Momentum Particle Maximum Likelihood

Jen Ning Lim [1] **Juan Kuntz** [2] **Samuel Power** [3] **Adam M. Johansen** [1]

Abstract

Maximum likelihood estimation (MLE) of latent variable models is often recast as the minimization of a free energy functional over an extended space of parameters and probability distributions. This perspective was recently combined with insights from optimal transport to obtain novel particle-based algorithms for fitting latent variable models to data. Drawing inspiration from prior works which interpret 'momentum-enriched' optimization algorithms as discretizations of ordinary differential equations, we propose an analogous dynamical-systems-inspired approach to minimizing the free energy functional. The result is a dynamical system that blends elements of Nesterov's Accelerated Gradient method, the underdamped Langevin diffusion, and particle methods. Under suitable assumptions, we prove that the continuous-time system minimizes the functional. By discretizing the system, we obtain a practical algorithm for MLE in latent variable models. The algorithm outperforms existing particle methods in numerical experiments and compares favourably with other MLE algorithms.

1. Introduction

In this work, we study parameter estimation for (probabilistic) latent variable models $p_\theta(y, x)$ with parameters $\theta \in \mathbb{R}^{d_\theta}$, unobserved (or latent) variables $x \in \mathbb{R}^{d_x}$, and observed variables $y \in \mathbb{R}^{d_y}$ (which we treat as fixed throughout). The type II maximum likelihood approach (Good, 1983) estimates θ by maximizing the *marginal likelihood* $p_\theta(y) := \int p_\theta(y, x)\, \mathrm{d}x$. However, for most models of practical interest, the integral has no known closed-form expressions and we are unable to optimize the marginal likelihood directly.

This issue is often overcome (e.g., Neal and Hinton (1998)) by constructing an objective defined over an extended space, whose optima are in one-to-one correspondence with those of the MLE problem. To this end, we define the 'free energy' functional:

$$\mathcal{E}(\theta, q) := \int \log\left(\frac{q(x)}{p_\theta(y, x)}\right) q(x)\, \mathrm{d}x. \tag{1}$$

The infimum of $\mathcal{E}$ over the *extended space* $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$ (where $\mathcal{P}(\mathbb{R}^{d_x})$ denotes the space of probability distributions over $\mathbb{R}^{d_x}$) coincides with negative of the log marginal likelihood's supremum:

$$\mathcal{E}^* := \inf_{(\theta,q)} \mathcal{E}(\theta, q) = -\sup_{\theta \in \mathbb{R}^{d_\theta}} \log p_\theta(y). \tag{2}$$

To see this, note that

$$\mathcal{E}(\theta, q) = -\log p_\theta(y) + \mathsf{KL}(q, p_\theta(\cdot \mid y)), \tag{3}$$

where KL denotes the Kullback–Leibler divergence and $p_\theta(\cdot|y) := p_\theta(y, \cdot)/p_\theta(y)$ the posterior distribution. So, if (θ^*, q^*) minimizes the free energy, then θ^* maximizes the marginal likelihood and $q^*(\cdot) = p_{\theta^*}(\cdot|y)$. In other words, by minimizing the free energy, we solve our MLE problem.

This perspective motivates the search for practical procedures that minimize the free energy $\mathcal{E}$. One such example is the classical Expectation Maximization (EM) algorithm applicable to models for which the posterior distributions $p_\theta(\cdot|y)$ are available in closed form. As shown in Neal and Hinton (1998), its iterates coincide with those of coordinate descent applied to the free energy $\mathcal{E}$. Recently, Kuntz et al. (2023) sought analogues of gradient descent (GD) applicable to the free energy $\mathcal{E}$. In particular, building on ideas popular in optimal transport (e.g., see Ambrosio et al. (2005)), they identified an appropriate notion for the free energy's gradient $\nabla\mathcal{E}$, discretized the corresponding gradient flow $(\dot{\theta}_t, \dot{q}_t) = -\nabla\mathcal{E}(\theta_t, q_t)$, and obtained a practical algorithm they called particle gradient descent (PGD).

In optimization, GD is well-known to be suboptimal: other practical first-order methods achieve better worst-case convergence guarantees and practical performance (Nemirovskij and Yudin, 1983; Nesterov, 1983). A common feature among algorithms that achieve the optimal 'accelerated' convergence rates is the presence of 'momentum' effects in the dynamics of the algorithm. Roughly speaking,

[1]University of Warwick [2]Polygeist [3]University of Bristol. Correspondence to: Jen Ning Lim <Jen-Ning.Lim@warwick.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

if GD for a function f is analogous to solving the ordinary differential equation (ODE)

$$\dot{\theta}_t = -\nabla_\theta f(\theta_t) \tag{4}$$

with $\nabla_\theta f$ denoting f's Euclidean gradient, then a 'momentum' method is akin to solving a second-order ODE like

$$\dot{\theta}_t = \eta m_t, \quad \dot{m}_t = -\gamma\eta m_t - \nabla_\theta f(\theta_t), \tag{5}$$

where γ, η denote positive hyperparameters (Su et al., 2014; Wibisono et al., 2016). In the context of machine learning, Sutskever et al. (2013) demonstrated empirically that incorporating momentum effects into stochastic gradient descent (SGD) often has substantial benefits such as eliminating the performance gap between standard SGD and competing 'Hessian-free' deterministic methods based on higher-order differential information (Martens, 2010).

Given the successes of momentum methods, it is natural to ask whether PGD can be similarly accelerated. In this work, we affirmatively answer this question. Our contributions are: (1) we construct a continuous-time flow that incorporates momentum effects into the gradient flow studied in Kuntz et al. (2023); (2) we derive a discretization that outperforms PGD and compares competitively against other methods. As theoretical contributions: (3) we prove in Proposition 3.1 the existence and uniqueness of the flow's solutions; (4) under suitable conditions, we show in Theorem 4.1 that the flow converges to $\mathcal{E}$'s minima; and, (5) we asymptotically justify in Proposition 5.1 the particle approximations we use when discretizing the flow. The proofs of Proposition 3.1 and Proposition 5.1 require generalizations of proofs pertaining to McKean–Vlasov SDEs, and the proof of Theorem 4.1 is a generalization of the Ma et al. (2021)'s convergence argument. These may be of independent interest for future works that operate in the extended space $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$.

The structure of this manuscript is as follows: in Section 2, we review gradient flows on the Euclidean space $\mathbb{R}^{d_\theta}$, probability space $\mathcal{P}(\mathbb{R}^{d_x})$, and extended space $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$. We also review momentum methods on $\mathbb{R}^{d_\theta}$ and $\mathcal{P}(\mathbb{R}^{d_x})$. In Section 3, we describe how momentum can be incorporated into the gradient flow on $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$, resulting in a dynamical system we refer to as the 'momentum-enriched flow'. In Section 4, we establish the flow's convergence. In Section 5, we discretize the flow and obtain an MLE algorithm we refer to as Momentum Particle Descent (MPD). In Section 6, we demonstrate empirically the efficacy of our method and study the effects of various design choices. As a large-scale experiment, we benchmark our proposed method for training Variational Autoencoders (Kingma and Welling, 2014) against current methods for training latent variable models. We conclude in Section 7 with a discussion of our results, their limitations, and future work.

2. Background

Our goal is to incorporate momentum into the gradient flow on the extended space $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$ considered in Kuntz et al. (2023) and obtain an algorithm for type II MLE that outperforms PGD. We begin by surveying the standard gradient flows on $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$'s component spaces and how momentum is incorporated in those settings.

2.1. Gradient Flows on $\mathbb{R}^d$ and their Acceleration

Given a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, f's Euclidean gradient flow is (4). Discretizing (4) in time using a standard forward Euler integrator yields GD. As observed in Su et al. (2014); Shi et al. (2021), algorithms such as Nesterov (1983)'s accelerated gradient (NAG) that achieve the optimal convergence rates are instead obtained by discretising the second-order ODE (5); cf. Appendix I.2. This ODE models the evolution of a kinetic particle with both position and momentum. In particular, for convex quadratic f, (5) reproduces the dynamics of a damped harmonic oscillator (McCall, 2010). Hence, the ODE's hyperparameters have intuitive interpretations: η^{-1} represents the particle's mass and γ determines the damping's strength.

We can rewrite (5) concisely as

$$(\dot{\theta}_t, \dot{m}_t) = -\mathsf{D}_\gamma \nabla_{(\theta,m)} f_\eta(\theta_t, m_t), \tag{6}$$

where f_η denotes the 'Hamiltonian function' and D_γ the 'damping matrix':

$$f_\eta(\theta, m) = f(\theta) + \frac{\eta}{2}\|m\|^2, \;\; \mathsf{D}_\gamma := \begin{pmatrix} 0_d & -I_d \\ I_d & \gamma I_d \end{pmatrix}. \tag{7}$$

In other words, (5) is a 'damped' or 'conformal' Hamiltonian flow (McLachlan and Perlmutter, 2001; Maddison et al., 2018). In particular, we can interpret NAG as a discretisation of F_η's 'γ-damped Hamiltonian flow' (Wibisono et al., 2016; Wilson et al., 2021).

2.2. Gradient Flows on $\mathcal{P}(\mathbb{R}^d)$ and their Acceleration

We can follow a similar program to solve optimization problems over $\mathcal{P}(\mathbb{R}^d)$. To do so, we require an analogue of the Euclidean gradient $\nabla_\theta f$ for (differentiable) functions f on $\mathbb{R}^d$ applicable to (sufficiently-regular) functionals F on $\mathcal{P}(\mathbb{R}^d)$. One way of defining $\nabla_\theta f(\theta)$ at a given point θ is as the unique vector satisfying

$$f(\theta') = f(\theta) + \langle \nabla_\theta f(\theta), \theta' - \theta \rangle + o(\mathsf{d}_E(\theta, \theta')), \tag{8}$$

where $\langle \cdot, \cdot \rangle$ and d_E denote the Euclidean inner product and metric related by the formula

$$\mathsf{d}_E(\theta, \theta')^2 = \inf_\lambda \int_0^1 \langle \lambda_t, \lambda_t \rangle \, \mathrm{d}t, \tag{9}$$

where the infimum is taken over all (sufficiently-regular) curves $\lambda : [0,1] \to \mathbb{R}^d$ connecting θ and θ'. Roughly speaking, we can reverse this process to define F's gradient for a given metric on $\mathcal{P}(\mathbb{R}^d)$. Here, we focus on the Wasserstein-2 metric W_2 for which (9)'s analogue is the Benamou-Brenier formula (Peyré and Cuturi, 2019, Eq.7.2):

$$\mathsf{W}_2(q,\tilde{q})^2 = \inf_{\rho} \int_0^1 g_{\rho_t}(\dot{\rho}_t, \dot{\rho}_t)\,\mathrm{d}t$$

where the infimum is taken over all (sufficiently-regular) curves $\rho : [0,1] \to \mathbb{R}^d$ connecting q and $\tilde{q}$ and

$$g_q(m, m') := \int \langle v(x), v'(x)\rangle\, q(x)\mathrm{d}x,$$

where v and v' solve the respective continuity equation, i.e., $m = -\nabla_x \cdot (qv), \quad m' = -\nabla_x \cdot (qv')$. Then, replacing $(\theta, \theta', f, \nabla_\theta, \langle\cdot,\cdot\rangle, \mathsf{d}_E)$ with $(q, q', F, \nabla_q, g_q(\cdot,\cdot), \mathsf{W}_2)$ in (8) and solving for $\nabla_q F$, we obtain

$$\nabla_q F(q) = -\nabla_x \cdot (q\nabla_x \delta_q F[q]),$$

where $\delta_q F[q] : \mathbb{R}^d \to \mathbb{R}$ denotes the functional's *first variation*: the function satisfying

$$F(q + \varepsilon\chi) = F(q) + \varepsilon \int \delta_q F[q](x)\ \chi(x)\,\mathrm{d}x + o(\varepsilon)$$

for all (sufficiently-regular) $\chi : \mathbb{R}^d \to \mathbb{R}$ s.t. $\int \chi(x)\mathrm{d}x = 0$. This is, of course, not the only choice; other metrics induce other 'geometries' and lead to other algorithms (e.g., see Duncan et al. (2023); Liu (2017); Sharrock and Nemeth (2023) for the Stein case). For a more rigorous presentation of these concepts, see Appendix C.1 and references therein.

We can now define F's Wasserstein gradient flow just as in the Euclidean case (except we have a PDE rather than an ODE): $\dot{q}_t = -\nabla_q F(q_t)$. A well-known example of such a flow is that for which $F(\cdot) := \mathsf{KL}(\cdot, p)$ for a fixed p in $\mathcal{P}(\mathbb{R}^d)$. In this case, $\delta_q F[q] = \log(q/p)$ and the flow reads $\dot{q}_t = \nabla_x \cdot (q_t \nabla_x \log(q_t/p))$. As noted in Jordan et al. (1998), this is the Fokker-Planck equation satisfied by the law of the overdamped Langevin SDE with invariant measure p: $\mathrm{d}X_t = \nabla_x \ell(X_t)\ \mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t$, where $\ell(x) := \log p(x)$ and $(W_t)_{t\geq 0}$ denotes the standard Wiener process. For this reason, papers such as Wibisono (2018); Ma et al. (2021) view the overdamped Langevin algorithm obtained by discretizing this SDE using the Euler–Maruyama scheme as an analogue of GD for F.

Ma et al. (2021) go a step further and search for 'accelerated' methods for minimizing F. To do so, they consider distributions over the momentum-enriched space $\mathbb{R}^{2d}$ rather than $\mathbb{R}^d$ and they replace F on $\mathcal{P}(\mathbb{R}^d)$ with the following 'Hamiltonian' on $\mathcal{P}(\mathbb{R}^d \times \mathbb{R}^d)$:

$$F_\eta(\cdot) := \mathsf{KL}(\cdot, p(\mathrm{d}x) \otimes r_\eta(\mathrm{d}u)), \tag{10}$$

where $\eta > 0$ denotes a fixed hyperparameter, $r_\eta := \mathcal{N}(0, \eta^{-1} I_d)$ a zero-mean η^{-1}-variance isotropic Gaussian distribution over the momentum variables u, and $p \otimes r_\eta$ the product measure. They then define the following $\mathcal{P}(\mathbb{R}^{2d})$-valued analogue of the γ-damped Hamiltonian flow (6):

$$\dot{q}_t = -\mathsf{D}_\gamma \nabla_q F_\eta(q_t) = \nabla_{(x,u)} \cdot \left(q_t \mathsf{D}_\gamma \nabla_{(x,u)} \delta_q F_\eta[q_t]\right). \tag{11}$$

This is the Fokker–Planck equation satisfied by the law of the underdamped Langevin SDE:

$$\mathrm{d}X_t = \eta U_t \mathrm{d}t,\quad \mathrm{d}U_t = [\nabla_x \ell(X_t) - \gamma\eta U_t]\mathrm{d}t + \sqrt{2\gamma}\mathrm{d}W_t. \tag{12}$$

Discretizing the SDE, Ma et al. (2021) recover the underdamped Langevin algorithm (e.g., see Cheng et al. (2018)) and show that, under suitable conditions, it achieves faster convergence rates than the overdamped Langevin algorithm.

2.3. Gradient Flows on $\mathbb{R}^{d_\theta} \times \mathcal{P}\left(\mathbb{R}^{d_x}\right)$

To minimize a functional $\mathcal{E}$ on $\mathbb{R}^{d_\theta} \times \mathcal{P}\left(\mathbb{R}^{d_x}\right)$, we can follow steps analogous to those in the previous two sections. First, given a metric d on this space, we can define a gradient $\nabla\mathcal{E}$ w.r.t. d similarly as in the beginning of Section 2.2; cf. Kuntz et al. (2023, Appendix A). Setting $\mathsf{d}((\theta, q), (\theta', q')) := \sqrt{\mathsf{d}_E(\theta,\theta')^2 + \mathsf{W}_2(q,q')^2}$, one finds that $\nabla\mathcal{E} = (\nabla_\theta \mathcal{E}, \nabla_q \mathcal{E})$, and the corresponding flow reads

$$\dot{\theta}_t = -\nabla_\theta \mathcal{E}(\theta_t, q_t), \quad \dot{q}_t = -\nabla_q \mathcal{E}(\theta_t, q_t).$$

Kuntz et al. (2023) were interested in MLE of latent variable problems and set $\mathcal{E}$ to be the free energy in (1). In this case, the above reduces to

$$\dot{\theta}_t = \int \nabla_\theta \ell(\theta_t, x)\, q_t(\mathrm{d}x), \quad \dot{q}_t = \nabla_x \cdot \left(q_t \nabla_x \log \frac{q_t}{\rho_{\theta_t}}\right),$$

where $\rho_\theta(\cdot) := p_\theta(y, \cdot)$. The above is the Fokker–Planck equation of the following McKean–Vlasov SDE:

$$d\theta_t = \int \nabla_\theta \ell(\theta_t, x)\, q_t(\mathrm{d}x),$$
$$\mathrm{d}X_t = \nabla_x \ell(\theta_t, X_t)\ \mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t,$$

where $q_t := \mathrm{Law}(X_t)$ and $\ell(\theta, x) := \log \rho_\theta(x)$. Because θ_t's drift depends on the X_t's law, this is a 'McKean–Vlasov' process (Kac, 1956; McKean Jr, 1966). Approximating the SDE as described in Appendix I.1, we obtain PGD.

3. Momentum-Enriched Flow for MLE

Following an approach analogous to those previously taken to obtain accelerated algorithms for minimizing functionals on $\mathbb{R}^{d_\theta}$ and $\mathcal{P}(\mathbb{R}^{d_x})$ (cf. Sections 2.1 and 2.2, respectively), we now derive a momentum-enriched flow that minimizes the free energy functional $\mathcal{E}$ over $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$ defined in (1). Approximating this flow, we will obtain an algorithm that experimentally outperforms PGD (Section 2.3).

We begin by enriching both components of the space with momentum variables: we replace θ and x in $\mathbb{R}^{d_\theta}$ and $\mathbb{R}^{d_x}$ with (θ, m) and (x, u) in $\mathbb{R}^{2d_\theta}$ and $\mathbb{R}^{2d_x}$; and we correspondingly substitute $\mathcal{P}(\mathbb{R}^{d_x})$ with $\mathcal{P}(\mathbb{R}^{2d_x})$. Next, we define a 'Hamiltonian' on the momentum-enriched space $\mathbb{R}^{2d_\theta} \times \mathcal{P}(\mathbb{R}^{2d_x})$ analogous to f_η and F_η in (7,10):

$$\mathcal{F}(\theta, m, q) := \int \log\left(\frac{q(x,u)}{\rho_{\theta,\eta_x}(x,u)}\right) q(x,u)\,\mathrm{d}x\,\mathrm{d}u + \frac{\eta_\theta}{2}\|m\|^2. \tag{13}$$

where (η_θ, η_x) denote positive hyperparameters and $\rho_{\theta,\eta_x}(x,u) := p_\theta(y,x)\, r_{\eta_x}(u)$ with $r_{\eta_x} := \mathcal{N}(0, \eta_x^{-1} I_{d_x})$. Consider now the following $(\gamma_\theta, \gamma_x)$-damped Hamiltonian flow of $\mathcal{F}$ analogous to (6,11):

$$(\dot{\theta}_t, \dot{m}_t) = -\mathsf{D}_{\gamma_\theta} \nabla_{(\theta,m)} \mathcal{F}(\theta_t, m_t, q_t), \tag{14a}$$
$$\dot{q}_t = \nabla_{(x,u)} \cdot \left(q_t \mathsf{D}_{\gamma_x} \nabla_{(x,u)} \delta_q \mathcal{F}[\theta_t, m_t, q_t]\right). \tag{14b}$$

This is the Fokker–Planck equation for the following McKean–Vlasov SDE:

$$\mathrm{d}\theta_t = \eta_\theta m_t\,\mathrm{d}t, \tag{15a}$$
$$\mathrm{d}m_t = \left[\int \nabla_\theta \ell(\theta_t, x)\, q_{t,X}(\mathrm{d}x) - \gamma_\theta \eta_\theta m_t\right]\mathrm{d}t, \tag{15b}$$
$$\mathrm{d}X_t = \eta_x U_t\,\mathrm{d}t, \tag{15c}$$
$$\mathrm{d}U_t = [\nabla_x \ell(\theta_t, X_t) - \gamma_x \eta_x U_t]\,\mathrm{d}t + \sqrt{2\gamma_x}\,\mathrm{d}W_t, \tag{15d}$$

where $q_{t,X} := \mathrm{Law}(X_t)$. Assuming that ℓ has a Lipschitz gradient, the SDE has a unique strong solution:

Assumption 1 (K-Lipschitz gradient)**.** *We assume that the potential ℓ has a K-Lipschitz gradient, i.e., there is some $K > 0$ such that the inequality holds*

$$\left\|\nabla_{(\theta,x)}\ell(\theta,x) - \nabla_{(\theta,x)}\ell(\theta',x')\right\| \le K\|(\theta,x) - (\theta',x')\|,$$

for all (θ, x) and (θ', x') in $\mathbb{R}^{d_\theta} \times \mathbb{R}^{d_x}$.

Proposition 3.1 (Existence and uniqueness of strong solutions to (15))**.** *Under Assumption 1, there exists a unique strong solution to (15) for any initial condition (θ_0, m_0, q_0) in $\mathbb{R}^{2d_\theta} \times \mathcal{P}(\mathbb{R}^{2d_x})$.*

Proof. The proof is a generalization of Carmona (2016, Theorem 1.7)'s proof; see Appendix G. □

4. Convergence of Momentum-Enriched Flow

To evidence that our approach is theoretically well-founded, we wish to show the momentum-enriched flow (14) solves the maximum likelihood problem. We do this by showing that the log marginal likelihood evaluated along $(\theta_t)_{t\ge0}$ (of the momentum-enriched flow (14)) converges exponentially fast to its supremum. Since directly working with the marginal likelihood is difficult, we exploit the following inequality: for all $m \in \mathbb{R}^{d_\theta}$, $q \in \mathcal{P}(\mathbb{R}^{2d_x})$,

$$-\log p_\theta(y) \le \mathcal{E}(\theta, q_X) \le \mathcal{F}(\theta, m, q), \tag{16}$$

where q_X denotes q's x-marginal. From (16) and the fact that both $\mathcal{E}$ and $\mathcal{F}$ share the same infimum $\mathcal{E}^*$, it holds that if $\mathcal{F}$ in (13) decays exponentially fast to its infimum along the solutions $(\theta_t, m_t, q_t)_{t\ge0}$ of (14), then we have the desired result of convergence in log marginal likelihood (as summarized in (16)). Hence, we dedicate the remainder of this section to establishing that $\mathcal{F}$ convergence exponentially along the momentum-enriched flow.

Because the momentum-enriched flow (14) is not a gradient flow, we are unable to prove the convergence using the techniques commonly applied to study gradient flows; e.g., see Otto and Villani (2000). Instead, we need to resort to an involved argument featuring a Lyapunov function $\mathcal{L}$ along the lines of those in Wilson et al. (2021); Wibisono et al. (2016); Ma et al. (2021).

Two immediate choices for Lyapunov functions are available: $\mathcal{E}$ and $\mathcal{F}$. However, unlike the gradient case, $\mathcal{E}$ along the momentum-enriched flow (15) is not monotonic and can fluctuate. As for $\mathcal{F}$, we show in Proposition E.1 that $\mathcal{F}$ along the flow denoted by $\mathcal{F}_t := \mathcal{F}(\theta_t, m_t, q_t)$ satisfies

$$\dot{\mathcal{F}}_t = -\gamma_\theta \|\nabla_m \mathcal{F}_t\|^2 - \gamma_x \|\nabla_u \delta_q \mathcal{F}_t\|^2 \le 0; \tag{17}$$

implying that $\mathcal{F}_t$ is non-increasing. This shows that the flow is in some sense stable, but it is not enough to prove the convergence: the two gradients could vanish without the flow necessarily having reached a minimizer. To preclude this happening, we need to force (θ, m)-gradients to feature in (17) similarly to the (x, u)-gradients. To achieve this, we follow Wilson et al. (2021); Ma et al. (2021) and introduce the following 'twisted' gradient terms:

$$\left\|\nabla_{(\theta,m)}\mathcal{F}\right\|^2_{T_{(\theta,m)}} := \left\langle \nabla_{(\theta,m)}\mathcal{F}, T_{(\theta,m)}\nabla_{(\theta,m)}\mathcal{F}\right\rangle,$$
$$\left\|\nabla_{(x,u)}\delta_q\mathcal{F}\right\|^2_{T_{(x,u)}} := \left\langle \nabla_{(x,u)}\delta_q\mathcal{F}, T_{(x,u)}\nabla_{(x,u)}\delta_q\mathcal{F}\right\rangle,$$

where

$$T_{(\theta,m)} = K^{-1}\begin{pmatrix} \tau_\theta I_{d_\theta} & \frac{\tau_{\theta m}}{2} I_{d_\theta} \\ \frac{\tau_{\theta m}}{2} I_{d_\theta} & \tau_m I_{d_\theta} \end{pmatrix},$$
$$T_{(x,u)} = K^{-1}\begin{pmatrix} \tau_x I_{d_x} & \frac{\tau_{xu}}{2} I_{d_x} \\ \frac{\tau_{xu}}{2} I_{d_x} & \tau_u I_{d_x} \end{pmatrix},$$

with K denoting the Lipschitz constant in Assumption 1 and $(\tau_\theta, \tau_{\theta m}, \tau_m, \tau_x, \tau_{xu}, \tau_x)$ non-negative constants such that the above matrices are positive-semidefinite. We then add the above terms to $\mathcal{F} - \mathcal{E}^*$ to obtain our Lyapunov function:

$$\mathcal{L} := \mathcal{F} - \mathcal{E}^* + \left\|\nabla_{(\theta,m)}\mathcal{F}\right\|^2_{T_{(\theta,m)}} + \left\|\nabla_{(x,u)}\delta_q\mathcal{F}\right\|^2_{T_{(x,u)}}. \tag{18}$$

We prove the convergence for models satisfying the a "log Sobolev inequality" (see Caprio et al. (2024) for a detailed discussion). It extends the log Sobolev inequality popular in optimal transport and the Polyak–Łojasiewicz inequality often used to argue linear convergence of GD algorithms.

Assumption 2 (Log Sobolev inequality)**.** *There exists a constant $C_{\mathcal{E}} > 0$, s.t. for all (θ, q) in $\mathbb{R}^{d_\theta} \times \in \mathcal{P}(\mathbb{R}^{d_x})$,*

$$\mathcal{E}(\theta, q) - \mathcal{E}^* \leq \frac{1}{2C_{\mathcal{E}}} \left|\left|\nabla\mathcal{E}(\theta, q)\right|\right|^2_q,$$

where $||\nabla\mathcal{E}(\theta, q)||_q := ||\nabla_\theta\mathcal{E}(\theta, q)||^2 + ||\nabla_x \log(q/\rho_\theta)||^2_q$ with $||\cdot||_q$ denoting the L^2_q norm.

Theorem 4.1 (Exponential convergence of the momentum-enriched flow)**.** *Suppose that Assumptions 1, 2 hold and there exists $(\varphi, \tau_\theta, \tau_{\theta m}, \tau_m, \tau_x, \tau_{xu}, \tau_u)$ satisfying the inequalities (51) in the supplement. If $\mathcal{L}_t := \mathcal{L}(\theta_t, m_t, q_t)$,*

$$\dot{\mathcal{L}}_t \leq -\varphi C \mathcal{L}_t,$$

where $C = \min\{C_{\mathcal{E}}, \eta_\theta, \eta_x\}$. Moreover,

$$\begin{aligned}&\sup_{\theta\in\mathbb{R}^{d_\theta}} \log p_\theta(y) - \log p_{\theta_t}(y)\\ &\leq \mathcal{E}_t - \mathcal{E}^* \leq \mathcal{F}_t - \mathcal{E}^* \leq \mathcal{L}_t \leq \mathcal{L}_0 \exp\left(-\varphi C t\right).\end{aligned} \tag{19}$$

Proof. The proof extends Ma et al. (2021, Proposition 1)'s proof. It requires generalizing from $\mathcal{P}(\mathbb{R}^{d_x})$ to the product space and controlling the terms that arise from the interaction between the two spaces; cf. Appendix F. □

It is not difficult to find hyperparameter choices for which the inequalities in (51) have solutions; cf. Appendix F.3 for some concrete examples.

5. Momentum Particle Descent

In order to realize an actionable algorithm, we need to discretize the dynamics of the momentum-enriched flow (15) both in space and in time. In this section, we derive the algorithm called Momentum Particle Descent (MPD) via discretization.

For the space discretization, following the approach of Kuntz et al. (2023), we approximate $q_{t,X} \approx \frac{1}{M}\sum_{i\in[M]} \delta_{X_t^{i,M}} =: q_{t,X}^M$ using a collection of particles. This yields the following system: for $i \in [M]$,

$$\mathrm{d}\theta_t^M = \eta_\theta m_t^M \,\mathrm{d}t \tag{20a}$$

$$\mathrm{d}m_t^M = -\gamma_\theta\eta_\theta m_t^M \,\mathrm{d}t - \nabla_\theta\mathcal{E}\left(\theta_t^M, q_{t,X}^M\right) \mathrm{d}t \tag{20b}$$

$$\mathrm{d}X_t^{i,M} = \eta_x U_t^{i,M} \,\mathrm{d}t \tag{20c}$$

$$\begin{aligned}\mathrm{d}U_t^{i,M} =& -\gamma_x\eta_x U_t^{i,M} \,\mathrm{d}t + \nabla_x\ell\left(\theta_t^M, X_t^{i,M}\right) \mathrm{d}t\\ &+ \sqrt{2\gamma_x}\,\mathrm{d}W_t^i.\end{aligned} \tag{20d}$$

where $\{W_t^i\}_{i=1}^M$ comprises M independent standard Wiener processes. Proposition 5.1 establishes asymptotic pointwise propagation of chaos which provides asymptotic justification for using a particle approximation for $q_{t,X}$ to approximate the flow in (15). The proof can be found in the Appendix H.

Proposition 5.1. *Under Assumption 1, we have*

$$\lim_{M\to\infty} \mathbb{E} \sup_{t\in[0,T]} \{\|\vartheta_t - \vartheta_t^M\|^2 + \mathrm{W}_2^2(q_t^{\otimes M}, q_t^M)\} = 0,$$

where $\vartheta_t := (\theta_t, m_t)$, $q_t^{\otimes M} := \Pi_{i=1}^M q_t$ with $q_t = \mathrm{Law}(X_t, U_t)$ are defined by the SDE in (15); and $\vartheta_t^M := (\theta_t^M, m_t^M)$, and $q_t^M = \mathrm{Law}(\{(X_t^{i,M}, U_t^{i,M})\}_{i=1}^M)$ in (20).

As for the time discretization, as noted by Ma et al. (2021), naive application of an Euler–Maruyama scheme to momentum-enriched dynamics may be insufficient for attaining an accelerated rate of convergence. We appeal to the literature on the discretization of underdamped Langevin dynamics to design an appropriate integrator. One integrator that can achieve accelerated rates is the (Euler) Exponential Integrator (Cheng et al., 2018; Sanz-Serna and Zygalakis, 2021; Hochbruck and Ostermann, 2010). The main strategy is to 'freeze' the nonlinear components of the dynamics in a suitable way, and then solve the resulting linear SDE. We also incorporate a partial update inspired by Sutskever et al. (2013)'s interpretation of NAG. More precisely, given $\left(\theta_0, m_0, \{\left(X_0^i, U_0^i\right)\}_{i=1}^M\right)$, define an approximating SDE on the time interval $t \in [0, h]$ as follows: for $i \in [M]$,

$$\mathrm{d}\tilde{\theta}_t = \eta_\theta\tilde{m}_t \,\mathrm{d}t \tag{21a}$$

$$\mathrm{d}\tilde{m}_t = -\gamma_\theta\eta_\theta\tilde{m}_t \,\mathrm{d}t - \nabla_\theta\mathcal{E}\left(\bar{\theta}_0, \tilde{q}_{0,X}^M\right) \mathrm{d}t \tag{21b}$$

$$\mathrm{d}\tilde{X}_t^i = \eta_x\tilde{U}_t^i \,\mathrm{d}t \tag{21c}$$

$$\mathrm{d}\tilde{U}_t^i = -\gamma_x\eta_x\tilde{U}_t^i \,\mathrm{d}t + \nabla_x\ell\left(\bar{\bar{\theta}}_0, X_0^i\right) \mathrm{d}t + \sqrt{2\gamma_x}\,\mathrm{d}W_t^i, \tag{21d}$$

where $\tilde{q}_{t,X}^M := \frac{1}{M}\sum_{i=1}^M \delta_{\tilde{X}_t^i}$, and the fixed variables $\bar{\theta}$ and $\bar{\bar{\theta}}_0$ (for reasons which will become clear when faced with the solved SDE) can be thought of as a partial update to $\tilde{\theta}_t$. This is a *linear* SDE, which can then be solved analytically, yielding the following iteration (see Appendix I.4 for details): for $i \in [M]$, we have

$$\begin{aligned}
\tilde{\theta}_k =& \tilde{\theta}_{k-1} + \frac{\iota_\theta(h)}{\gamma_\theta}\tilde{m}_{k-1} \\
&- \frac{1}{\gamma_\theta}\left(h - \frac{\iota_\theta(h)}{\gamma_\theta\eta_\theta}\right)\nabla_\theta\mathcal{E}\left(\bar{\theta}_{k-1}, \tilde{q}^M_{k-1,X}\right) \\
\tilde{m}_k =& (1-\iota_\theta(h))\tilde{m}_{k-1} - \frac{\iota_\theta(h)}{\gamma_\theta\eta_\theta}\nabla_\theta\mathcal{E}\left(\bar{\theta}_{k-1}, \tilde{q}^M_{k-1,X}\right) \\
\tilde{X}^i_k =& \tilde{X}^i_{k-1} + \frac{\iota_x(h)}{\gamma_x}\tilde{U}^i_{k-1} \\
&+ \frac{1}{\gamma_x}\left(h - \frac{\iota_x(h)}{\gamma_x\eta_x}\right)\nabla_x\ell\left(\bar{\bar{\theta}}_{k-1}, \tilde{X}^i_{k-1}\right) + L^{XX}_\Sigma\xi^i_k \\
\tilde{U}^i_k =& (1-\iota_x(h))\tilde{U}^i_{k-1} + \frac{\iota_x(h)}{\gamma_x\eta_x}\nabla_x\ell\left(\bar{\bar{\theta}}_{k-1}, \tilde{X}^i_{k-1}\right) \\
&+ L^{XU}_\Sigma\xi^i_k + L^{UU}_\Sigma\xi'^{,i}_k,
\end{aligned}$$

where $\iota_x(h) := 1 - \exp(-\eta_x\gamma_x h)$, $\left\{\xi^i_k, \xi'^{,i}_k\right\}_{i\in[M]} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0_{d_x}, I_{d_x})$, and $\left(L^{XX}_\Sigma, L^{XU}_\Sigma, L^{UU}_\Sigma\right)$ are suitable scalars that depend on (η_x, γ_x) that can be found in the appendix as Equation (83). We set

$$\bar{\theta}_{k-1} := \tilde{\theta}_{k-1} + \frac{\iota_\theta(h)}{\gamma_\theta}\tilde{m}_{k-1}, \quad \bar{\bar{\theta}}_{k-1} = \tilde{\theta}_k. \tag{22}$$

The astute reader will notice that our approximating ODE-SDE (i.e., Equation (21)) deviates from Cheng et al. (2018)'s style of discretization. Specifically, the difference lies in where we compute the gradient in Equations (21b) and (21d) which we refer to as *gradient correction*. The choices of $\bar{\theta}_{k-1}$ and $\bar{\bar{\theta}}_{k-1}$ in Equation (22) are chosen to reflect partial and full updates to $\tilde{\theta}_k$ respectively. This is reminiscent of NAG's usage of a partial parameter update to compute the gradient, unlike Polyak's Heavy Ball, which does not (Sutskever et al., 2013, Section 2.1). In our case, we empirically found that this choice is critical for a more stable discretization and enables the algorithm to be "faster" by taking larger step sizes (see Section 6.1).

The choice of the momentum parameters $(\eta_\theta, \gamma_\theta, \eta_x, \gamma_x)$ is crucial to the practical performance of MPD. We observe that (as is also the case for NAG, the underdamped Langevin SDE, and other similar systems) there are three different qualitative regimes for the dynamics: i) the underdamped regime, in which the parameter values oscillate, ii) the overdamped regime, in which one recovers PGD-type behaviour, and iii) the critically-damped regime, in which oscillations are largely suppressed, but the momentum effects are still able to accelerate the convergence behaviour relative to PGD. While rigorous approaches are limited to simple targets (e.g., Dockhorn et al. (2022)), one can utilize cross-validation to obtain these parameters in practice.

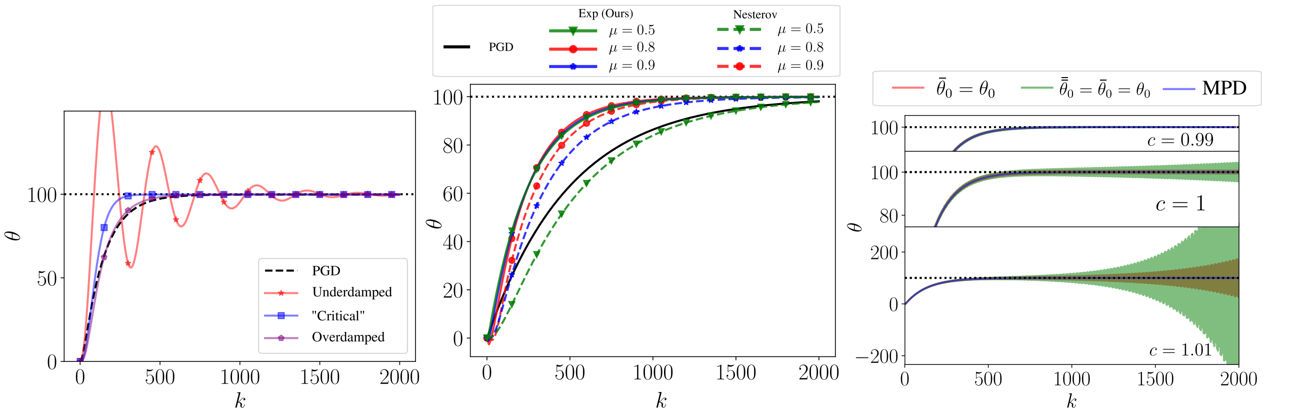
6. Experiments

In this section, we study various design choices of MPD and demonstrate the efficacy of our proposal through empirical results. The structure of this section is as follows: in Section 6.1, we demonstrate on the Toy Hierarchical model the effects of various design choices; then, in Section 6.2, we compare MPD with algorithms that incorporate momentum in either θ or X only; finally, in Section 6.3, we compare our proposal on training VAEs with other methods. The code for reproducibility is available online: `https://github.com/jenninglim/mpgd`.

6.1. Toy Hierarchical Model

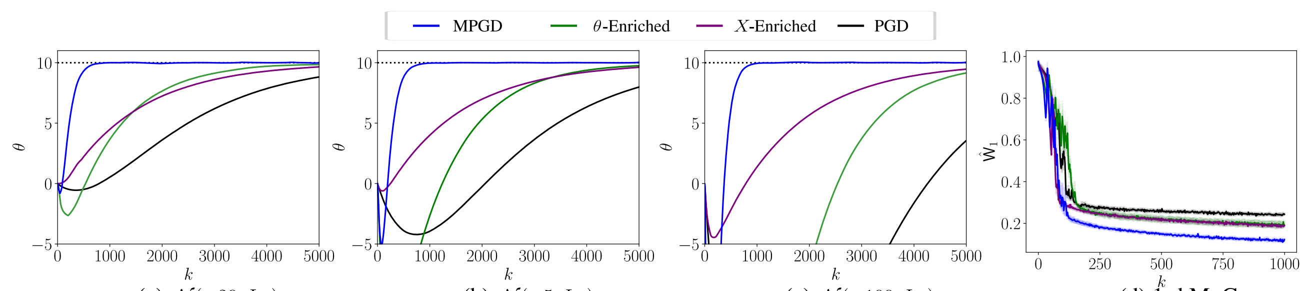
As a toy example, we consider the hierarchical model proposed in Kuntz et al. (2023). The model $\mathrm{ToyHM}(\theta, \sigma)$ is given by $p_\theta(y, x) := \prod_{i=1}^N \mathcal{N}(y_i|x_i, 1)\mathcal{N}(x_i|\theta, \sigma^2)$, where $N = 100$ is the number of data points. The dataset y is sampled from a model with $\theta = 100, \sigma = 1$. In this experiment, we wish to understand the behaviour of MPD compared against PGD. We are particularly interested in (1) how the momentum parameters $(\eta_\theta, \gamma_\theta, \eta_x, \gamma_x)$ affect the optimization process; (2) comparing our proposed discretization to another that follows in the style of NAG (detailed in Appendix I.3); and (3) the role of the gradient correction term described in Section 5. The results are shown in Figure 1. Further experiment details can be found in Appendix J.4.1.

In Figure 1a, we show that different regimes can arise from different choices of hyper-parameters (as discussed in Section 5). In Figure 1b, we compare the performance of MPD using (our) Exponential integrator with a NAG-like discretization for (θ_t, m_t)-component. We vary the "momentum coefficient", i.e., we have $\mu \in \{0.5, 0.8, 0.9\}$ with $\mu_\theta = \mu_x = \mu$. It can be seen that MPD with our exponential integrator for θ_t performs better than NAG-like integrator (see Appendix J.3 for more discussion). In Figure 1c, we show the effect of the gradient correction term for three different step sizes in (θ, m)-components and (X, U)-components. The different lines in the figure are generated from multiplying the stepsize of each component with a constant $c \in \{0.99, 1, 1.01\}$. It can be seen that our proposed gradient correction results in a more effective discretization. In red, we show MPD with the absence of gradient correction in Equation (21b) (i.e., when we use $\bar{\theta}_0 = \theta_0$ in Equation (21b)), and, in green, we show the MPD when the gradient correction is absent in both Equation (21b) and Equation (21d) (i.e. when we use $\bar{\bar{\theta}}_0 = \bar{\theta}_0 = \theta_0$ in Equation (21b) and Equation (21d)). It can be seen the gradient correction term results in a more stable algorithm.



(a) Different regimes in MPD. (b) θ_t integrator comparison. (c) Role of gradient correction.

Figure 1: **Toy Hierarchical Model**. (a) Different regimes that arise from varying the momentum parameters; (b) comparison between our Exponential (Exp) integrator and a Nesterov-like integrator for different momentum parameters; (c) we compare the performance of the MPD with and without gradient correction.



(a) $\mathcal{N}(-20, I_{d_x})$ (b) $\mathcal{N}(-5, I_{d_x})$ (c) $\mathcal{N}(-100, I_{d_x})$ (d) 1-d MoG

Figure 2: **Comparison of MPD with algorithms that only enrich one component.** (a), (b), (c) shows the performance on the $\mathrm{ToyHM}(10, 12)$ with different initialization of the particle cloud $\{X_0^i\}_{i=1}^M$. In (d) shows the $\hat{W}_1$ vs iterations of a density estimation problem on a Mixture of Gaussian (MoG) dataset. MPD is shown in blue, θ-only-enriched shown in green, X-only-enriched in purple, and PGD in black. The results are averaged over 10 independent trials.

6.2. Why Accelerate Both Components?

One may question the importance of momentum-enriching both components of PGD. We show in two experiments that enriching only one component results in a suboptimal algorithm. To show this, we consider the $\mathrm{ToyHM}(10, 12)$, and, as a more realistic example, a 1-d density estimation problem using a VAE on a Mixture of Gaussians with two equally weighted components which are $\mathcal{N}(2, 0.5)$ and $\mathcal{N}(-2, 0.5)$. To measure the performance of each algorithm, for the first problem, we examine the parameter of the model which should converge to the true value of 10; and, for the second, we compute the empirical Wasserstein-1 $\hat{W}_1$ using the empirical CDF.

Figure 2 shows the results of both experiments. Overall, it can be seen that the presence of momentum results in a better performance. Furthermore, we observe that the algorithms which only enrich one component can exhibit problems which are not present for MPD. It can be seen that the poor initialization of the cloud drastically lengthens the transient phase of the θ-only enriched algorithm, whereas MPD can recover more rapidly. This is a typical setting in high-dimensional settings (see Figure 2 (d)). Conversely, while the X-only enriched algorithm does not suffer as much from poor initialization, it can be observed to suffer from slow convergence in the θ-component (intuitively, one pays a larger price for poor conditioning of the ideal MLE objective). Overall, it can be observed that MPD noticeably outperforms PGD and all other intermediate methods in all settings. For details/discussion, see Appendix K.

6.3. Image Generation

For this task, we consider two datasets: MNIST (LeCun et al., 1998) and CIFAR-10 (Krizhevsky and Hinton, 2009). For the model, we use a variational autoencoder (Kingma and Welling, 2014) with a VampPrior (Tomczak and Welling, 2018). As baselines, we compare our proposed method MPD against PGD, Alternating Backpropagation

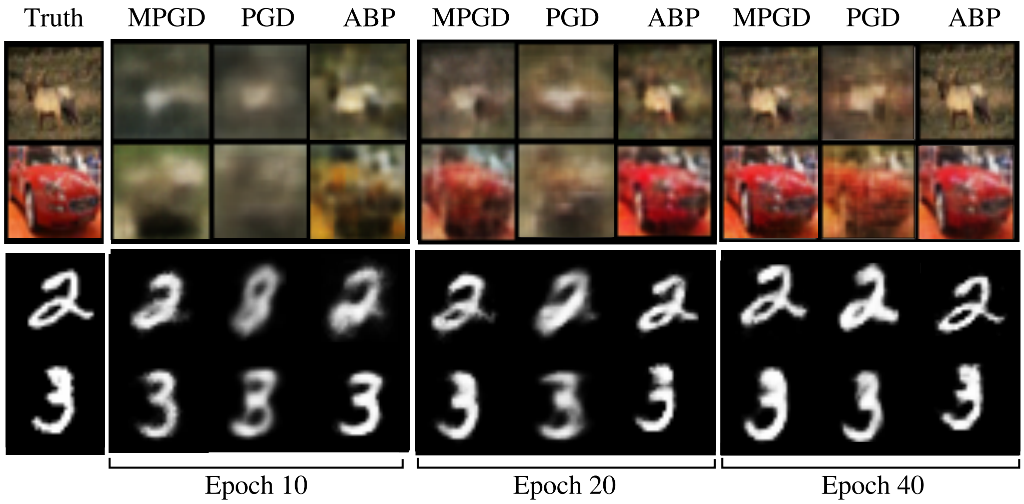


Figure 3: **Posterior Cloud vs Epochs**. We show the evolution of the reconstruction of a particle for *persistent* methods. The particle is taken at epoch $\{10, 20, 40\}$ on MNIST and CIFAR-10.

Dataset	MPD	PGD	ABP	SR	VI
MNIST	$\mathbf{47.9} \pm 0.6$	104.1 ± 1.2	$\mathbf{45.9} \pm 0.8$	109.97 ± 12.5	72.6 ± 1.3
CIFAR	$\mathbf{93.2} \pm 1.5$	106.2 ± 4.8	97.0 ± 8.0	126.25 ± 13.0	$\mathbf{94.2} \pm 6.7$

Table 1: **FID scores** on MNIST and CIFAR-10 after the final epoch. In **bold**, we indicate the lowest two scores on average. We write $(\mu \pm \sigma)$ to indicate the mean μ and standard deviation σ of the FID score calculated over three independent trials.

(ABP) (Han et al., 2017), Short Run (SR) (Nijkamp et al., 2020), and amortized variational inference (VI) (Kingma and Welling, 2014). ABP is the most similar to PGD. They differ in that ABP takes multiple steps of the (unadjusted and overdamped) Langevin algorithm (ULA) instead of PGD's single step. They (MPD, PGD, ABP) are also *persistent*, meaning that the ULA chain starts at the previous particle location, whereas SR restarts the chain at a random location sampled from the prior but like ABP runs the chain for several steps. Further experiment details can be found in Appendix J.4.2.

We are interested in the generative performance of the resulting models. For a qualitative measure, we show the samples produced by the model in Figure 6 for CIFAR (for MNIST samples, see Figure 5 in the Appendix). As a quantitative measure, we report the Fréchet inception distance (FID) (Heusel et al., 2017) as shown in Table 1. It can be seen that our proposed method does well compared against other baselines. The closest competitor is ABP. As noted by Kuntz et al. (2023), ABP can reap the benefits of taking multiple ULA steps to locate the posterior mode quickly and reduce the transient phase. This hypothesis is further confirmed in Figure 3, where we visualize the evolution of a single particle across various epochs. It can be seen that ABP's and MPD's distinct methods of reducing the transient phase (by taking multiple steps or utilizing momentum) are effective. As a competitor SR does perform badly, which we attribute to its non-persistent property: by restarting the particles at the prior and running short chains, it is unable to overcome the bias of short chains and locate the posterior mode. Variational inference performs well for CIFAR but surprisingly falls slightly short in MNIST.

7. Conclusion, Limitations, and Future work

We presented Momentum Particle Descent MPD, a method of incorporating momentum into particle gradient descent. We establish several theoretical results such as the convergence in continuous time; existence and uniqueness; as well as some guarantees for the usage of the particle approximation. Through experiments, we showed that, for suitably chosen momentum parameters, the resulting algorithm achieves better performance than PGD.

The main limitation that may impede the widespread adoption of MPD over PGD is the requirement for tuning the momentum parameters. This issue is inherited from the Underdamped Langevin dynamics where it is generally understood to be somewhat more challenging than its overdamped counterpart, as witnessed by the dearth of papers proposing practical tuning strategies for this class of algorithms. Although we found that the momentum coefficient heuristic (see Appendix J.3) worked well in our experiments, future work includes a systematic method of tuning momentum parameters $(\eta_\theta, \gamma_\theta, \eta_x, \gamma_x)$ following Riou-Durand et al. (2023) (or some justification for using the momentum coefficient heuristic). Another future work is a theoretical characterization of the difference between our discretization scheme compared with other potential schemes akin to Sanz-Serna and Zygalakis (2021).

Acknowledgements

We thank our anonymous reviewers for their helpful comments which improved the quality of the paper. JL acknowledge the support from the Feuer International Scholarship. JK and AMJ acknowledge support from the Engineering and Physical Sciences Research Council (EPSRC; grant EP/T004134/1). AMJ acknowledges further support from the EPSRC (grants EP/R034710/1 and EP/Y014650/1). SP acknowledges support from the Engineering and Physical Sciences Research Council (EPSRC; grant EP/R018561/1).

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Ambrosio, L., Gigli, N., and Savaré, G. (2005). *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive MCMC. *Statistics and computing*, 18:343–373.

Bakry, D. and Émery, M. (2006). Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84: Proceedings*, pages 177–206. Springer.

Bernton, E. (2018). Langevin Monte Carlo and JKO splitting. In *Conference on learning theory*, pages 1777–1798. PMLR.

Caprio, R., Kuntz, J., Power, S., and Johansen, A. M. (2024). Error bounds for particle gradient descent, and extensions of the log-Sobolev and Talagrand inequalities. *arXiv preprint arXiv:2403.02004*.

Carmona, R. (2016). *Lectures on BSDEs, stochastic control, and stochastic differential games with financial applications*. SIAM.

Chen, F., Lin, Y., Ren, Z., and Wang, S. (2024). Uniform-in-time propagation of chaos for kinetic mean field Langevin dynamics. *Electronic Journal of Probability*, 29:1–43.

Cheng, X., Chatterji, N. S., Bartlett, P. L., and Jordan, M. I. (2018). Underdamped Langevin MCMC: A non-asymptotic analysis. In *Conference on Learning Theory*, pages 300–323. PMLR.

Chizat, L. (2022). Mean-field langevin dynamics : Exponential convergence and annealing. *Transactions on Machine Learning Research*.

Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31.

De Bortoli, V., Durmus, A., Pereyra, M., and Vidal, A. F. (2021). Efficient stochastic optimisation by unadjusted Langevin Monte Carlo: Application to maximum marginal likelihood and empirical Bayesian estimation. *Statistics and Computing*, 31:1–18.

Delyon, B., Lavielle, M., and Moulines, E. (1999). Convergence of a stochastic approximation version of the EM algorithm. *Annals of statistics*, pages 94–128.

Diao, M. Z., Balasubramanian, K., Chewi, S., and Salim, A. (2023). Forward-backward Gaussian variational inference via JKO in the Bures-Wasserstein space. In *International Conference on Machine Learning*, pages 7960–7991. PMLR.

Dockhorn, T., Vahdat, A., and Kreis, K. (2022). Score-based generative modeling with critically-damped Langevin diffusion. In *International Conference on Learning Representations (ICLR)*.

Duncan, A., Nüsken, N., and Szpruch, L. (2023). On the geometry of Stein variational gradient descent. *Journal of Machine Learning Research*, 24(56):1–39.

Garbuno-Inigo, A., Nüsken, N., and Reich, S. (2020). Affine invariant interacting Langevin dynamics for Bayesian inference. *SIAM Journal on Applied Dynamical Systems*, 19(3):1633–1658.

Gelfand, I. M. and Silverman, R. A. (2000). *Calculus of Variations*. Courier Corporation.

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. University of Minnesota Press.

Han, T., Lu, Y., Zhu, S.-C., and Wu, Y. N. (2017). Alternating back-propagation for generator network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., and Hochreiter, S. (2017). GANs trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in Neural Information Processing Systems*, 30.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Hochbruck, M. and Ostermann, A. (2010). Exponential integrators. *Acta Numerica*, 19:209–286.

Hu, K., Ren, Z., Šiška, D., and Szpruch, Ł. (2021). Mean-field Langevin dynamics and energy landscape of neural networks. In *Annales de l'Institut Henri Poincare (B) Probabilites et statistiques*, volume 57, pages 2043–2065. Institut Henri Poincaré.

Jordan, R., Kinderlehrer, D., and Otto, F. (1998). The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17.

Kac, M. (1956). Foundations of kinetic theory. In *Proceedings of The third Berkeley symposium on mathematical statistics and probability*, volume 3, pages 171–197.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. In Bengio, Y. and LeCun, Y., editors, *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Krizhevsky, A. and Hinton, G. (2009). Learning multiple layers of features from tiny images.

Kruger, A. Y. (2003). On Fréchet subdifferentials. *Journal of Mathematical Sciences*, 116(3):3325–3358.

Kuntz, J., Lim, J. N., and Johansen, A. M. (2023). Particle algorithms for maximum likelihood training of latent variable models. In *Proceedings on 26th International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 206 of *Proceedings of Machine Learning Research*, pages 5134–5180.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. (2022). Variational inference via Wasserstein gradient flows. *Advances in Neural Information Processing Systems*, 35:14434–14447.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Liu, Q. (2017). Stein variational gradient descent as gradient flow. *Advances in Neural Information Processing Systems*, 30.

Loève, M. (1977). *Probability Concepts*, pages 151–176. Springer New York, New York, NY.

Ma, Y.-A., Chatterji, N. S., Cheng, X., Flammarion, N., Bartlett, P. L., and Jordan, M. I. (2021). Is there an analog of Nesterov acceleration for gradient-based MCMC? *Bernoulli*, 27(3):1942 – 1992.

Maddison, C. J., Paulin, D., Teh, Y. W., O'Donoghue, B., and Doucet, A. (2018). Hamiltonian descent methods. *arXiv preprint arXiv:1809.05042*.

Martens, J. (2010). Deep learning via Hessian-free optimization. In *International Conference on Machine Learning*, volume 27, pages 735–742.

McCall, M. W. (2010). *Classical Mechanics: From Newton to Einstein: A Modern Introduction*. John Wiley & Sons.

McKean Jr, H. P. (1966). A class of Markov processes associated with nonlinear parabolic equations. *Proceedings of the National Academy of Sciences*, 56(6):1907–1911.

McLachlan, R. and Perlmutter, M. (2001). Conformal Hamiltonian systems. *Journal of Geometry and Physics*, 39(4):276–300.

Mei, S., Montanari, A., and Nguyen, P.-M. (2018). A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671.

Neal, R. M. and Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. *Learning in graphical models*, pages 355–368.

Nemirovskij, A. S. and Yudin, D. B. (1983). *Problem complexity and method efficiency in optimization*. Wiley-Interscience.

Nesterov, Y. (2003). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Science & Business Media.

Nesterov, Y. E. (1983). A method of solving a convex programming problem with convergence rate $\mathcal{O}\left(\frac{1}{k^2}\right)$. In *Doklady Akademii Nauk*, volume 269, pages 543–547. Russian Academy of Sciences.

Nijkamp, E., Pang, B., Han, T., Zhou, L., Zhu, S.-C., and Wu, Y. N. (2020). Learning multi-layer latent variable model via variational optimization of short run MCMC for approximate inference. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 361–378. Springer.

Nitanda, A. and Suzuki, T. (2017). Stochastic particle gradient descent for infinite ensembles. *arXiv preprint arXiv:1712.05438*.

Nitanda, A., Wu, D., and Suzuki, T. (2022). Convex analysis of the mean field Langevin dynamics. In *International Conference on Artificial Intelligence and Statistics*, pages 9741–9757. PMLR.

Otto, F. and Villani, C. (2000). Generalization of an Inequality by Talagrand and Links with the Logarithmic Sobolev Inequality. *Journal of Functional Analysis*, 173(2):361–400.

O'Donoghue, B. and Candes, E. (2015). Adaptive restart for accelerated gradient schemes. *Foundations of Computational Mathematics*, 15:715–732.

Pang, B., Han, T., Nijkamp, E., Zhu, S.-C., and Wu, Y. N. (2020). Learning latent space energy-based prior model. *Advances in Neural Information Processing Systems*, 33:21994–22008.

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Foundations and Trends in Machine Learning*, 11(5-6):355–607.

Platen, E. and Bruti-Liberati, N. (2010). *Numerical Solution of Stochastic Differential Equations with Jumps in Finance*. Springer Science & Business Media.

Riou-Durand, L., Sountsov, P., Vogrinc, J., Margossian, C., and Power, S. (2023). Adaptive tuning for Metropolis adjusted Langevin trajectories. In *International Conference on Artificial Intelligence and Statistics*, pages 8102–8116. PMLR.

Roberts, G. O. and Rosenthal, J. S. (2009). Examples of adaptive MCMC. *Journal of computational and graphical statistics*, 18(2):349–367.

Santambrogio, F. (2017). {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7:87–154.

Sanz-Serna, J. M. and Zygalakis, K. C. (2021). Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations. *The Journal of Machine Learning Research*, 22(1):11006–11042.

Sharrock, L. and Nemeth, C. (2023). Coin sampling: Gradient-based Bayesian inference without learning rates. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J., editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 30850–30882. PMLR.

Shi, B., Du, S. S., Jordan, M. I., and Su, W. J. (2021). Understanding the acceleration phenomenon via high-resolution differential equations. *Mathematical Programming*, 195:79–148.

Silvester, J. R. (2000). Determinants of block matrices. *The Mathematical Gazette*, 84(501):460–467.

Staib, M., Reddi, S., Kale, S., Kumar, S., and Sra, S. (2019). Escaping saddle points with adaptive gradient methods. In *International Conference on Machine Learning*, pages 5956–5965. PMLR.

Su, W., Boyd, S., and Candes, E. (2014). A differential equation for modeling Nesterov's accelerated gradient method: theory and insights. *Advances in Neural Information Processing Systems*, 27.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International Conference on Machine Learning*, pages 1139–1147. PMLR.

Suzuki, T., Wu, D., and Nitanda, A. (2023). Mean-field langevin dynamics: Time-space discretization, stochastic gradient, and variance reduction. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Tieleman, T. and Hinton, G. (2012). Lecture 6.5-rmsprop, coursera: Neural networks for machine learning. *University of Toronto, Technical Report*, 6.

Tomczak, J. and Welling, M. (2018). VAE with a VampPrior. In *International Conference on Artificial Intelligence and Statistics*, pages 1214–1223. PMLR.

Van Handel, R. (2014). Probability in high dimension. *Lecture Notes (Princeton University)*.

Villani, C. (2009). *Optimal transport: old and new*, volume 338. Springer.

Wibisono, A. (2018). Sampling as optimization in the space of measures: The Langevin dynamics as a composite optimization problem. In *Conference on Learning Theory*, pages 2093–3027. PMLR.

Wibisono, A., Wilson, A. C., and Jordan, M. I. (2016). A variational perspective on accelerated methods in optimization. *Proceedings of the National Academy of Sciences*, 113(47):E7351–E7358.

Wilson, A. C., Recht, B., and Jordan, M. I. (2021). A Lyapunov analysis of accelerated methods in optimization. *Journal of Machine Learning Research*, 22(113):1–34.

Yao, R. and Yang, Y. (2022). Mean field variational inference via Wasserstein gradient flow. *arXiv preprint arXiv:2207.08074*.

Appendices

A. Notation

The following table summarizes some key notation used throughout.

$\mathcal{E}$	Free energy defined in Equation (1).
$\mathcal{F}$	Momentum-enriched free energy defined in Equation (13).
z_t	The tuple $(\theta_t, m_t, q_t) \in \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x} \times \mathbb{R}^{d_x})$.
∇f	Euclidean gradient of f. If $f : \mathbb{R}^n \to \mathbb{R}^m$ is vector-valued, we have $\nabla f \in \mathbb{R}^{n\times m}$.
$\nabla_a f$	Euclidean gradient of f w.r.t. a.
$\nabla_{(a,b)} f$	$[\nabla_a f, \nabla_b f]^\top$.
ℓ	$\ell(\theta, x) := \log p_\theta(y, x)$.
ρ_θ	$\rho_\theta(x) := p_\theta(y, x)$.
$\nabla \cdot f$	If $f : \mathbb{R}^n \to \mathbb{R}^n$, we have $\nabla \cdot f := \partial_i f_i$. If $f : \mathbb{R}^n \to \mathbb{R}^n \times \mathbb{R}^m$, we have $(\nabla \cdot f)_i := \partial_j f_{ji}$.
$\nabla_a \cdot$	Divergence operator w.r.t. a.
∇_a^*	Adjoint operator of ∇_a (see Appendix F.4.2).
$[n]$	$[n] := \{1, ..., n\}$.
Δ	Laplacian operator, $\Delta = \nabla \cdot \nabla$.
$\mathcal{P}(\mathbb{R}^d)$	The space of probability measures that are absolutely continuous w.r.t. Lebesgue measure (have densities) and possess finite second moments.
$\langle \cdot, \cdot \rangle$ (and $\Vert \cdot \Vert$)	Euclidean inner product or Frobenius inner product (and its inner norm).
$\langle \cdot, \cdot \rangle_\rho$ (and $\Vert \cdot \Vert_\rho$)	$L^2(\rho)$ inner product (and its norm).
$o(\epsilon)$	Bachmann–Landau little-o notation.
$\partial F(\mu)$	Fréchet subdifferential (see Definition C.1).

B. Related Work

The present work sits at the juncture of i) deterministic gradient flows for optimising objectives over "parameter" spaces, typically expressible through the discretization of ODEs, and ii) stochastic gradient flows for optimising objectives over the space of probability measures, typically expressible through discretisation of mean-field SDEs. The former class of problems is too vast to be properly surveyed here (for an overview, see Santambrogio (2017); Ambrosio et al. (2005)), effectively including a large proportion of modern continuous optimisation problems. The latter class has seen substantial growth over the past few years in particular, with various problems related to sampling (Liu, 2017; Bernton, 2018; Garbuno-Inigo et al., 2020; Duncan et al., 2023), variational inference (Yao and Yang, 2022; Lambert et al., 2022; Diao et al., 2023), and the training of shallow neural networks (Mei et al., 2018; Chizat and Bach, 2018; Nitanda and Suzuki, 2017; Hu et al., 2021; Nitanda et al., 2022; Chizat, 2022; Chen et al., 2024; Suzuki et al., 2023) being studied in this framework. While there exist earlier works which combine optimisation with Markovian sampling (e.g. stochastic approximation approaches to the EM algorithm (Delyon et al., 1999; De Bortoli et al., 2021), training of energy-based models (Hinton, 2002), and hyperparameter tuning in MCMC (Andrieu and Thoms, 2008; Roberts and Rosenthal, 2009)), the connection to gradient flows remains somewhat under-developed at present. We hope that the present work can encourage further exploration of these connections.

C. Gradient Flow on $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$

In this section, we describe gradient flows on the extended space $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$. We begin with an exposition of gradient flows in Wasserstein space, i.e. the space of distributions $\mathcal{P}(\mathbb{R}^{d_x})$ endowed with the Wasserstein-2 metric. We aim to give an intuitive introduction as opposed to a rigorous one; those readers with an interest in the latter are directed to Ambrosio et al. (2005). Using the ideas in Ambrosio et al. (2005), we show how the notion of gradients can be generalized to the product space $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$.

C.1. Gradient Flow on $\mathcal{P}(\mathbb{R}^{d_x})$

We begin by describing the gradient flow on $\mathcal{P}(\mathbb{R}^{d_x})$ endowed with the Wasserstein-2 metric. The Wasserstein distance is defined as

$$\mathsf{W}_2^2(p,q) = \inf_{\pi\in\Pi(p,q)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|x-y\|^2\, \pi(\mathrm{d}x\times \mathrm{d}y),$$

where $\Pi(p,q)$ is the set of all couplings between p and q.

In Ambrosio et al. (2005, Chapter 11), the authors discuss various approaches for adapting gradient flows on well-studied spaces (such as Euclidean and Riemannian spaces) to the Wasserstein space. One of these approaches proceeds by first defining suitable notions of tangent space and subdifferential, following which the simple definition of gradient flow modelled on Riemannian manifolds can then be reproduced. In this case, the Fréchet subdifferential (Ambrosio et al., 2005, Definition 10.1.1) is defined as follows:

Definition C.1 (Fréchet differential on Wasserstein Space)**.** *Let $F:\mathcal{P}(\mathbb{R}^{d_x})\to\mathbb{R}$ be a sufficiently regular function. We say that $\xi\in L^2(q)$ belongs to the Fréchet subdifferential $\partial F(q)$ if for all $q'\in\mathcal{P}(\mathbb{R}^{d_x})$, we have*

$$F(q') - F(q) \geq \left\langle \xi, t_q^{q'} - i\right\rangle_q + o(\mathsf{W}_2(q,q')),$$

where t_p^q is the optimal map between p and q (Ambrosio et al., 2005, see (7.1.4)) and i is the identity map. Furthermore, if $\xi\in\partial F(q)$ also satisfies

$$F(t_{\#}q) - F(q) \geq \langle \xi, t-i\rangle_q + o(\|t-i\|_q),$$

*for all $t\in L^2(q)$, then we say that ξ is a **strong** subdifferential.*

See Ambrosio et al. (2005, Definition 10.1.1) for more details. The strong subdifferential can be thought of as the (Wasserstein) "gradient" of F. Equipped with this notion of gradient, we can define the gradient (descent) flow of $F:\mathcal{P}(\mathbb{R}^{d_x})\to\mathbb{R}$ as follows:

Definition C.2 (Gradient Flow)**.** *We say that a curve $p_t:[0,1]\to\mathcal{P}(\mathbb{R}^{d_x})$ is a gradient flow of $F:\mathcal{P}(\mathbb{R}^{d_x})\to\mathbb{R}$ if for all $t>0$, it satisfies the continuity equation $\partial_t p_t + \nabla_x\cdot(v_t p_t) = 0$, where the tangent vector v_t satisfies $-v_t\in\partial F(p_t)$ for all t.*

Thus, for our application, we are interested in computing the strong subdifferential of F. If F is an integral of the type

$$F(p) = \int f(x, p(x), \nabla_x p(x))\mathrm{d}x, \tag{23}$$

where $f:\mathbb{R}^{d_x}\times\mathbb{R}\times\mathbb{R}^{d_x}\to\mathbb{R}$ is sufficiently regular, then we will see that its strong subdifferential admits an analytic solution.

Functionals of this form of great interest in the calculus of variations (e.g., see Gelfand and Silverman (2000)). A vital quantity which is used to study these functionals is the first variation. Writing $\delta_p F[p]:\mathbb{R}^{d_x}\to\mathbb{R}$ for the first variation of F, it is a function that satisfies

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}F(q_\epsilon)\bigg|_{\epsilon=0} = -\int \delta_q F[q](x)\, \nabla_x\cdot(q(x)v(x))\mathrm{d}x,$$

for all v such that $q_\epsilon := (i+\epsilon v)_{\#}q\in\mathcal{P}(\mathbb{R}^{d_x})$ for sufficiently small ϵ. One can readily establish that for (23)-typed F, it is given by

$$\delta_p F[p](x) := \nabla_{(2)} f(x,p(x),\nabla_x p(x)) - \nabla_x\cdot(\nabla_{(3)} f(x,p(x),\nabla_x p(x)),$$

where $\nabla_{(i)}$ denotes the partial derivative w.r.t. the i-th argument (Ambrosio et al., 2005, Eq. (10.4.2)). It can be shown that for any $\xi \in \partial F(p)$ which is a strong subdifferential, it holds that $\xi(x) \overset{p-a.e.}{=} \nabla_x \delta_p F[p](x)$ (Ambrosio et al., 2005, Lemma 10.4.1).

C.2. Gradient flow on $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$

We provided here a generalization of the Fréchet differential (for a broad survey of Fréchet differentials, see Kruger (2003)) to the extended space $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$:

Definition C.3 (Fréchet differential on $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$)**.** *Let $\mathcal{G} : \mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x}) \to \mathbb{R}$ be a sufficiently regular function. We say that $(\xi_\theta, \xi_q) \in \mathbb{R}^{d_\theta} \times L_2(q)$ belongs to the Fréchet subdifferential $\partial\mathcal{G}(\theta, q)$ if for all $(\theta', q') \in \mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$, it holds that*

$$\mathcal{G}(\theta', q') - \mathcal{G}(\theta, q) \geq \langle \xi_\theta, \theta' - \theta \rangle + \left\langle \xi_q, t_q^{q'} - i \right\rangle_q + o(\mathsf{W}_2(q', q) + \|\theta' - \theta\|).$$

Furthermore, if $(\xi_\theta, \xi_q) \in \partial\mathcal{G}(\theta, q)$ also satisfies

$$\mathcal{G}(\theta + \tau, t_{\#}q) - \mathcal{G}(\theta, q) \geq \langle \xi_\theta, \tau \rangle + \langle \xi_q, t - i \rangle_q + o(\|t - i\|_q + \|\tau\|),$$

*for all $(\tau, t) \in \mathbb{R}^{d_\theta} \times L_2(q)$, then we say that (ξ_θ, ξ_q) is a **strong** subdifferential.*

For the remainder of the section, we shall assume that for any perturbation (σ, v), there exists some $\delta_q F[\theta, q] : \mathbb{R}^{d_x} \to \mathbb{R}$ (called the first variation) such that the following expansion holds:

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon} \mathcal{G}(\theta_\epsilon, q_\epsilon)\big|_{\epsilon=0} = \langle \nabla_\theta \mathcal{G}(\theta, q), \sigma \rangle - \int \delta_q \mathcal{G}[\theta, q](x) \, \nabla_x \cdot (v(x) q(x)) \, \mathrm{d}x, \tag{24}$$

where $q_\epsilon := (i + \epsilon v)_{\#} q$ and $\theta_\epsilon := \theta + \epsilon\sigma$. For all $\mathcal{G}$ of interest in this paper, this assumption holds; for instance, see Proposition F.4.

We follow in the argument of Ambrosio et al. (2005, Lemma 10.4.1) to show that if $(\xi_\theta, \xi_q) \in \partial\mathcal{G}(\theta, q)$ is a strong subdifferential, then we have that $\xi_\theta = \nabla_\theta \mathcal{G}(\theta, q)$ and $\xi_q(x) \overset{q-a.e.}{=} \nabla_x \delta_q \mathcal{G}[\theta, q](x)$. Let $(\xi_\theta, \xi_p) \in \partial\mathcal{G}(\theta, p)$ be a strong subdifferential. By using the strong subdifferential property for some perturbation $(\epsilon\sigma, \epsilon v)$ and taking left and right limits of (C.3), we obtain that

$$\limsup_{\epsilon \uparrow 0} \frac{\mathcal{G}(\theta_\epsilon, q_\epsilon) - \mathcal{G}(\theta, q)}{\epsilon} \leq \langle \xi_\theta, \sigma \rangle + \langle \xi_q, v \rangle_q \leq \liminf_{\epsilon \downarrow 0} \frac{\mathcal{G}(\theta_\epsilon, q_\epsilon) - \mathcal{G}(\theta, q)}{\epsilon}.$$

On the other hand, from our assumption (24), we have that

$$\begin{aligned}
\lim_{\epsilon \to 0} \frac{\mathcal{G}(\theta_\epsilon, q_\epsilon) - \mathcal{G}(\theta, q)}{\epsilon} &= \langle \nabla_\theta \mathcal{G}(\theta, q), \sigma \rangle - \int \delta_q \mathcal{G}[\theta, q](x) \, \nabla_x \cdot (v(x) q(x)) \, \mathrm{d}x \\
&= \langle \nabla_\theta \mathcal{G}(\theta, q), \sigma \rangle + \langle \nabla_x \delta_q \mathcal{G}[\theta, q], v \rangle_q ,
\end{aligned}$$

where the last equality follows from integration by parts and the divergence theorem. Hence, we have that $\xi_\theta = \nabla_\theta \mathcal{G}(\theta, q)$, and $\xi_q \overset{q-a.e.}{=} \nabla_x \delta_q \mathcal{G}[\theta, q]$.

C.3. First Variation

In this section, we derive the first variation for integral expressions taking a certain form. In particular, we are interested in computing the strong subdifferential of $\mathcal{G}$ of the following types:

$$\mathcal{G}(\theta, q) := \int f(\theta, x, q(x), \nabla_x q(x)) \, \mathrm{d}x, \text{ where } f : \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_x} \times \mathbb{R} \times \mathbb{R}^{d_x}; \tag{VI-I}$$

$$\mathcal{G}(\theta, q) := \int \int f(\theta, x, x', q(x), q(x')) \, \mathrm{d}x \mathrm{d}x', \text{ where } f : \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \times \mathbb{R} \times \mathbb{R}. \tag{VI-II}$$

An exampled of (VI-I)-typed $\mathcal{G}$ is when $\mathcal{G}$ is the free energy $\mathcal{E}$. Following standard techniques from the calculus of variations (Gelfand and Silverman, 2000), consider a perturbation $(\sigma, v) \in \mathbb{R}^{d_\theta} \times L_2(q)$ and define a mapping Φ by

$$\Phi(\epsilon) := \mathcal{G}(\theta_\epsilon, q_\epsilon),$$

where $q_\epsilon := (i + \epsilon v)_{\#} q$ and $\theta_\epsilon := \theta + \epsilon\sigma$.

(VI-I)**-typed** $\mathcal{G}$. By application of the change-of-variables formula (for instance, see Ambrosio et al. (2005, Lemma 5.5.3)), we can compute the density q_ϵ and its derivative $\frac{\partial}{\partial \epsilon} q_\epsilon(y)$ as follows:

$$q_\epsilon(y) = \frac{q}{\det(I + \epsilon \nabla_x v)} \circ (i + \epsilon v)^{-1}(y), \quad \frac{\partial}{\partial \epsilon} q_\epsilon(y)\big|_{\epsilon=0} = -\nabla_x \cdot [qv](y).$$

Thus, we can compute the derivative of Φ, provided that the interchange of derivative and integral can be justified, as

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\epsilon}\Phi(\epsilon)\bigg|_{\epsilon=0} &= \int \frac{\mathrm{d}}{\mathrm{d}\epsilon} f(\theta_\epsilon, x, q_\epsilon(x), \nabla_x q_\epsilon(x))\,\mathrm{d}x \\
&= \int \langle \nabla_\theta f(\theta, x, q(x), \nabla_x q(x)), \sigma \rangle \,\mathrm{d}x \\
&\quad - \int \nabla_{(3)} f(\theta, x, q(x), \nabla_x q(x))\, \nabla_x \cdot [vq](x)\,\mathrm{d}x \\
&\quad - \int \big\langle \nabla_{(4)} f(\theta, x, q(x), \nabla_x q(x)), \nabla_x[\nabla_x \cdot [vq]](x) \big\rangle \,\mathrm{d}x.
\end{aligned}$$

Applying integration by parts and the divergence theorem, we obtain that

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\epsilon}\Phi(\epsilon)\bigg|_{\epsilon=0} &= \langle \nabla_\theta \mathcal{G}(\theta, q), \sigma \rangle \\
&\quad - \int \nabla_{(3)} f(\theta, x, q(x), \nabla_x q(x))\, [\nabla_x \cdot (vq)](x)\,\mathrm{d}x \\
&\quad + \int \nabla_x \cdot [\nabla_{(4)} f](\theta, x, q(x), \nabla_x q(x))[\nabla_x \cdot (vq)](x)\,\mathrm{d}x.
\end{aligned}$$

We can hence write the first variation of (VI-I)-typed $\mathcal{G}$ as

$$\delta_q \mathcal{G}[\theta, q](x) = \nabla_{(3)} f(\theta, x, q(x), \nabla_x q(x)) - \nabla_x \cdot (\nabla_{(4)} f(\theta, x, q(x), \nabla_x q(x)). \tag{25}$$

(VI-II)**-typed** $\mathcal{G}$. Similarly to above, we can define $\Phi(\epsilon) := \mathcal{G}(\theta_\epsilon, q_\epsilon)$, whose derivative is given by

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\epsilon}\Phi(\epsilon)\bigg|_{\epsilon=0} &= \langle \nabla_\theta \mathcal{G}(\theta, q), \sigma \rangle \\
&\quad - \int\int \nabla_{(4)} f(\theta, x, y, q(x), q(y))\, \nabla_x \cdot [qv](x)\,\mathrm{d}x\,\mathrm{d}y \\
&\quad - \int\int \nabla_{(5)} f(\theta, x, y, q(x), q(y))\, \nabla_y \cdot [qv](y)\,\mathrm{d}x\,\mathrm{d}y.
\end{aligned}$$

Hence, the first variation is given by

$$\delta_q \mathcal{G}[\theta, q](x) = \int \left[\nabla_{(4)} f(\theta, x, y, q(x), q(y)) + \nabla_{(5)} f(\theta, y, x, q(y), q(x))\right] \mathrm{d}y. \tag{26}$$

Example 1 (First Variation of $\mathcal{E}$)**.** *It can be seen that the free energy is* (VI-I)*-typed with* $f(\theta, x, a, g) = a \cdot \log \frac{a}{p_\theta(y,x)}$. *Hence, by combining the formula* (25) *with the fact that*

$$\nabla_3 f(\theta, x, a, b) = \log \frac{a}{p_\theta} + 1,$$

$$\nabla_4(f(\theta, x, a, g) = 0,$$

we obtain the following expression for the first variation:

$$\delta_q \mathcal{E}[\theta, q](x) = \log \frac{q(x)}{p_\theta(y, x)} + 1.$$

This provides an alternative derivation of this expression to that given in Kuntz et al. (2023, Lemma 1).

D. A log Sobolev inequality for $\mathcal{E}$ can be transferred to $\mathcal{F}$

In this section, we show that nice properties of the functional $\mathcal{E}$ transfer to the functional $\mathcal{F}$, i.e. if $\mathcal{E}$ satisfies a log Sobolev inequality (Assumption 2), then so does $\mathcal{F}$ (with a modified constant):

Proposition D.1. *If Assumption 2 holds, then upon defining $C := \min\{C_\mathcal{E}, \eta_\theta, \eta_x\}$, it holds that*

$$\mathcal{F}(\theta, m, q) - \mathcal{E}^* \leq \frac{1}{2C} ||\nabla \mathcal{F}(\theta, m, q)||_q^2, \quad \forall \theta, m \in \mathbb{R}^{d_\theta}, \;\; q \in \mathcal{P}(\mathbb{R}^{2d_x}),$$

where

$$||\nabla \mathcal{F}(\theta, m, q)||_q^2 := \|\nabla_{(\theta,m)} \mathcal{F}(\theta, m, q)\|^2 + \|\nabla_{(x,u)} \log q - \nabla_{(x,u)} \log \rho_{\theta,\eta_x}\|_q^2. \tag{27}$$

Proof. As we show at the end of the proof, we have the fact that

$$\left|\left|\nabla_{(x,u)} \log \left(\frac{q}{\rho_\theta \otimes r_{\eta_x}}\right)\right|\right|_q^2 \geq \left|\left|\nabla_x \log \left(\frac{q_X}{\rho_\theta}\right)\right|\right|_{q_X}^2 + \left|\left|\nabla_u \log \frac{q_{U|X}}{r_{\eta_x}}\right|\right|_q^2. \tag{28}$$

From the definition, we have

$$||\nabla \mathcal{F}(\theta, m, q)||_q^2 = ||\nabla_\theta \mathcal{F}(\theta, m, q)||^2 + ||\eta_\theta m||^2 + \left|\left|\nabla_{(x,u)} \log \frac{q}{\rho_{\theta,\eta_x}}\right|\right|_q^2.$$

Using (28) and the fact that

$$\nabla_\theta \mathcal{F}(\theta, m, q) = -\int \nabla_\theta \log \rho_{\theta,\eta_x}(x, u) q(\mathrm{d}x, \mathrm{d}u) = -\int \nabla_\theta \log \rho_\theta(x) q_X(\mathrm{d}x) = \nabla_\theta \mathcal{E}(\theta, q_X),$$

then it follows that

$$||\nabla \mathcal{F}(\theta, m, q)||_q^2 \geq ||\nabla_\theta \mathcal{E}(\theta, q_X)||_q^2 + \eta_\theta^2 ||m||^2 + \left|\left|\nabla_u \log \frac{q_{U|X}}{r_\eta}\right|\right|_q^2.$$

Because $C = \min\{C_\mathcal{E}, \eta_\theta, \eta_x\}$, the above implies that

$$\frac{1}{2C} ||\nabla \mathcal{F}(\theta, m, q)||_q^2 \geq \frac{1}{2C_\mathcal{E}} ||\nabla_\theta \mathcal{E}(\theta, q)||_q^2 + \frac{\eta_\theta}{2} ||m||^2 + \frac{1}{2\eta_x} \left|\left|\nabla_u \log \frac{q_{U|X}}{r_{\eta_x}}\right|\right|_q^2. \tag{29}$$

Since $r_{\eta_x} = \mathcal{N}(0, \eta_x^{-1} I_{d_x})$ is η_x-strongly log-concave, by the Bakry–Émery criterion of Bakry and Émery (2006), it holds that

$$\frac{1}{2\eta_x} \left|\left|\nabla_u \log \left(\frac{q(\cdot|x)}{r_{\eta_x}}\right)\right|\right|_{q(\mathrm{d}u|x)}^2 \geq \mathsf{KL}(q_{U|x}|r_{\eta_x}) \quad \forall x \in \mathcal{X},$$

and so we obtain that

$$\begin{aligned}\frac{1}{2\eta_x} \left|\left|\nabla_u \log \frac{q_{U|X}}{r_{\eta_x}}\right|\right|_q^2 &= \frac{1}{2\eta_x} \mathbb{E}_{q_X} \left[\left|\left|\nabla_u \log \frac{q(\cdot|X)}{r_{\eta_x}}\right|\right|_{q(\mathrm{d}u|X)}^2\right] \\ &\geq \mathbb{E}_{q_X} \left[\mathsf{KL}(q_{U|X}|r_{\eta_x})\right].\end{aligned}$$

Plugging the above into (29) and using the log Sobolev inequality for $\mathcal{E}$, we obtain that for $\mathcal{F}$:

$$\begin{aligned}\frac{1}{2C}\left|\left|\nabla \mathcal{F}(\theta, m, q)\right|\right|_q^2 &\geq \mathcal{E}(\theta, q_X) - \mathcal{E}^* + \frac{\eta_\theta}{2}\|m\|_2^2 + \mathbb{E}_{q_X}\left[\mathsf{KL}(q_{U|X}|r_{\eta_x})\right] \\ &\geq \mathcal{F}(\theta, m, q) - \mathcal{E}^*.\end{aligned}$$

We have one loose end to tie up: proving (28). To do so, note that

$$\begin{aligned}\left|\left|\nabla_{(x,u)} \log\left(\frac{q}{\rho_\theta \otimes r_{\eta_x}}\right)\right|\right|_q^2 &= \left|\left|\nabla_{(x,u)} \log\left(\frac{q_X}{\rho_\theta}\right) + \nabla_{(x,u)} \log\left(\frac{q_{U|X}}{r_{\eta_x}}\right)\right|\right|_q^2 \\ &= \left|\left|\nabla_x \log\left(\frac{q_X}{\rho_\theta}\right) + \nabla_x \log q_{U|X}\right|\right|_q^2 \qquad (30) \\ &\quad + \left|\left|\nabla_u \log\left(\frac{q_{U|X}}{r_{\eta_x}}\right)\right|\right|_q^2. \qquad (31)\end{aligned}$$

Taking the first term and expanding the square, we have

$$\begin{aligned}\left|\left|\nabla_x \log\left(\frac{q_X}{\rho_\theta}\right) + \nabla_x \log q_{U|X}\right|\right|_q^2 &= \left|\left|\nabla_x \log\left(\frac{q_X}{\rho_\theta}\right)\right|\right|_q^2 \\ &\quad + 2\left\langle \nabla_x \log\left(\frac{q_X}{\rho_\theta}\right), \nabla_x \log q_{U|X}\right\rangle_q \\ &\quad + \left|\left|\nabla_x \log q_{U|X}\right|\right|_q^2.\end{aligned}$$

Applying Jensen's inequality to the final term, we obtain

$$\begin{aligned}\left|\left|\nabla_x \log\left(\frac{q_X}{\rho_\theta}\right) + \nabla_x \log q_{U|X}\right|\right|_q^2 &\geq \left|\left|\nabla_x \log\left(\frac{q_X}{\rho_\theta}\right)\right|\right|_{q_X}^2 \\ &\quad + 2\left\langle \nabla_x \log\left(\frac{q_X}{\rho_\theta}\right), \mathbb{E}_{q(\mathrm{d}u|\cdot)}\left[\nabla_x \log q(U|\cdot)\right]\right\rangle_{q_X} \\ &\quad + \left|\left|\mathbb{E}_{q(\mathrm{d}u|\cdot)}\left[\nabla_x \log q(U|\cdot)\right]\right|\right|_{q_X}^2.\end{aligned}$$

One can compute explicitly that for all x, it holds that $\mathbb{E}_{q(\mathrm{d}u|x)}\left[\nabla_x \log q(U|x)\right] = 0$; we thus obtain that

$$\left|\left|\nabla_x \log \frac{q_X}{\rho_\theta} + \nabla_x \log q_{U|X}\right|\right|_q^2 \geq \left|\left|\nabla_x \log \frac{q_X}{\rho_\theta}\right|\right|_{q_X}^2,$$

and so (28) follows from (31). □

E. $\mathcal{F}(\theta_t, m_t, q_t)$ is non-increasing

In the following proposition, we show that $\mathcal{F}_t := \mathcal{F}(\theta_t, m_t, q_t)$ is non-increasing in time.

Proposition E.1. *For any $\gamma_x \geq 0$ and $\gamma_\theta \geq 0$, it holds that*

$$\dot{\mathcal{F}}_t = -\gamma_\theta \|\nabla_m \mathcal{F}_t\|^2 - \gamma_x \|\nabla_u \delta_q \mathcal{F}_t\|_{q_t}^2 \leq 0.$$

Proof. We begin by computing the time derivative

$$\dot{\mathcal{F}}_t = -\left\langle \nabla_{(\theta,m)}\mathcal{F}_t, \begin{pmatrix} 0 & -I_{d_\theta} \\ I_{d_\theta} & \gamma_\theta I_{d_\theta}\end{pmatrix} \nabla_{(\theta,m)}\mathcal{F}_t \right\rangle \qquad (32)$$

$$-\left\langle \nabla_{(x,u)}\delta_q\mathcal{F}_t, \begin{pmatrix} 0 & -I_{d_x} \\ I_{d_x} & \gamma_x I_{d_x} \end{pmatrix} \nabla_{(x,u)}\delta_q\mathcal{F}_t \right\rangle_{q_t}. \tag{33}$$

Decomposing the matrix $\begin{pmatrix} 0 & -I_{d_\theta} \\ I_{d_\theta} & \gamma_\theta I_{d_\theta} \end{pmatrix}$ into symmetric and skew-symmetric components (write D, Q respectively), i.e.

$$\begin{pmatrix} 0 & -I_{d_\theta} \\ I_{d_\theta} & \gamma_\theta I_{d_\theta} \end{pmatrix} = \underbrace{\begin{pmatrix} 0 & -I_{d_\theta} \\ I_{d_\theta} & 0 \end{pmatrix}}_{=:Q} + \underbrace{\begin{pmatrix} 0 & 0 \\ 0 & \gamma_\theta I_{d_\theta} \end{pmatrix}}_{=:D},$$

we can simplify the RHS of (32) to

$$\text{RHS (32)} = -\left\langle \nabla_{(\theta,m)}\mathcal{F}_t, D\nabla_{(\theta,m)}\mathcal{F}_t \right\rangle = -\gamma_\theta \|\nabla_m \mathcal{F}_t\|^2.$$

Similarly, we can show that

$$(33) = -\gamma_x \|\nabla_u \delta_q \mathcal{F}_t\|^2_{q_t}.$$

As such, for $\gamma_x \geq 0$, $\gamma_\theta \geq 0$, the claim follows. □

F. Proof of Theorem 4.1

The outline of the proof of Theorem 4.1 is:

- **Step 1** (Appendix F.1): Explicitly computing an upper bound of the time derivative of $\mathcal{F}$ as a quadratic form.
- **Step 2** (Appendix F.2): Under the conditions specified in (51), we show that this time derivative is bounded above by another quadratic form that allows us to apply the log Sobolev inequality of Proposition D.1.
- **Step 3**: Using log Sobolev inequality, Proposition D.1, and Grönwall's inequality, we obtain the desired result.

The remainder of the sections are dedicated to supporting the proof of Steps 1 and 2. This is done by developing the technical tools and carrying out explicit computations. The particular roles of the following sections are:

1. Appendix F.4.1: Computing the time derivative of $\mathcal{L}$ for Step 1.
2. Appendices F.4.2 to F.4.4: Introducing the adjoint, commutator and another operator, as well as their explicit forms.
3. Appendix F.4.5: Using the operators and explicit forms in Appendices F.4.2 to F.4.4, we can upper bound terms introduced in Appendix F.4.1 for Step 1. We utilize this in Step 1.
4. Appendix F.4.6. Bounding the cross terms (or interaction terms) between θ and x that arises in the time derivative.
5. Appendix F.4.7. Establishing sufficient conditions for a matrix with a particular form to be positive semi-definite given in Proposition F.18. This is used in Step 2.

Recall that the Lyapunov function is given by:

$$\mathcal{L} := \mathcal{F} - \mathcal{E}^* + \left\|\nabla_{(\theta,m)}\mathcal{F}\right\|^2_{T_{(\theta,m)}} + \left\|\nabla_{(x,u)}\delta_q\mathcal{F}\right\|^2_{T_{(x,u)}}$$

where

$$T_{(\theta,m)} := \frac{1}{K}\begin{pmatrix} \tau_\theta I_{d_\theta} & \frac{\tau_{\theta m}}{2} I_{d_\theta} \\ \frac{\tau_{\theta m}}{2} I_{d_\theta} & \tau_m I_{d_\theta} \end{pmatrix}, \quad T_{(x,u)} := \frac{1}{K}\begin{pmatrix} \tau_x I_{d_x} & \frac{\tau_{xu}}{2} I_{d_x} \\ \frac{\tau_{xu}}{2} I_{d_x} & \tau_u I_{d_x} \end{pmatrix} \tag{34}$$

Proof of Theorem 4.1. The proof is completed in the following three steps:

Step 1. In Section F.1, we show that the time derivative of $\mathcal{L}_t := \mathcal{L}(z_t)$ satisfies the following upper bound:

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_t \leq -\left\langle \nabla_{(\theta,m)}\mathcal{F}_t, S_{(\theta,m)}\nabla_{(\theta,m)}\mathcal{F}_t\right\rangle - \left\langle \nabla_{(x,u)}\delta_q\mathcal{F}_t, S_{(x,u)}\nabla_{(x,u)}\delta_q\mathcal{F}_t\right\rangle_{q_t}.$$

where $S_{(x,u)}$ and $S_{(\theta,m)}$ are suitable matrices defined in (47) and (48), respectively.

Step 2. Then, in Section F.2, when Equation (51) holds for some rate $\varphi > 0$, we have that

$$S_{(\theta,m)} \succeq \varphi C\left(T_{(\theta,m)} + \frac{1}{2C}I_{d_\theta}\right),$$
$$S_{(x,u)} \succeq \varphi C\left(T_{(x,u)} + \frac{1}{2C}I_{d_x}\right),$$

where $C := \min\{C_{\mathcal{E}}, \eta_\theta, \eta_x\}$.

Step 3. By Step 1 and Step 2, we obtain

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_t \leq & -\varphi C\left(\frac{1}{2C}\left|\left|\nabla_{(\theta,m)}\mathcal{F}_t\right|\right|^2 + \left\|\nabla_{(\theta,m)}\mathcal{F}_t\right\|^2_{T_{(\theta,m)}}\right) \\ & -\varphi C\left(\frac{1}{2C}\left|\left|\nabla_{(x,u)}\delta_q\mathcal{F}_t\right|\right|^2_{q_t} + \left\|\nabla_{(x,u)}\delta_q\mathcal{F}_t\right\|^2_{T_{(x,u)}}\right).\end{aligned}$$

Comparing the above to (27), applying Proposition D.1 and the Equation (18),

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_t \leq -\varphi C\mathcal{L}_t,$$

and application of Gronwall's inequality yields the final result

$$\mathcal{F}_t - \mathcal{E}^* \leq \mathcal{L}_0 \exp\left(-\varphi C t\right).$$

□

F.1. Proof of Step 1

As shown in Proposition F.2, the time derivative of the Lyapunov function $\mathcal{L}$ is given by

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_t = -\left\langle \nabla_{(\theta,m)}\mathcal{L}_t, \begin{pmatrix} 0 & -I_{d_\theta} \\ I_{d_\theta} & \gamma_\theta I_{d_\theta}\end{pmatrix}\nabla_{(\theta,m)}\mathcal{F}_t\right\rangle \tag{35}$$

$$-\left\langle \nabla_{(x,u)}\delta_q\mathcal{L}_t, \begin{pmatrix} 0 & -I_{d_x} \\ I_{d_x} & \gamma_\theta I_{d_x}\end{pmatrix}\nabla_{(x,u)}\delta_q\mathcal{F}_t\right\rangle_{q_t}, \tag{36}$$

where we abbreviate $\mathcal{L}_t := \mathcal{L}(\theta_t, m_t, q_t)$ and $\mathcal{F}_t := \mathcal{F}(\theta_t, m_t, q_t)$.

We deal with the inner products in (35,36) one-by-one, starting with (35): we first compute the gradient of $\mathcal{L}$ w.r.t. to (θ, m), finding that

$$\nabla_{(\theta,m)}\mathcal{L} = \nabla_{(\theta,m)}\mathcal{F} + 2[\nabla^2_{(\theta,m)}\mathcal{F}]\,T_{(\theta,m)}\nabla_{(\theta,m)}\mathcal{F} + 2\mathbb{E}_q\left[[\nabla_{(\theta,m)}\nabla_{(x,u)}\delta_q\mathcal{F}]\,T_{(x,u)}\nabla_{(x,u)}\delta_q\mathcal{F}\right],$$

where, as before, assume that the derivative-integral exchange can justified (see Appendix C.3). It then follows that

$$(35) = -\left\langle \nabla_{(\theta,m)}\mathcal{F}_t, \tilde{\Gamma}_\theta\nabla_{(\theta,m)}\mathcal{F}_t\right\rangle \tag{37}$$

$$-2\left\langle \nabla_{(\theta,m)}\nabla_{(x,u)}[\delta_q\mathcal{F}_t]\,T_{(x,u)}\nabla_{(x,u)}\delta_q\mathcal{F}_t, \begin{pmatrix} 0 & -I_{d_\theta} \\ I_{d_\theta} & \gamma_\theta I_{d_\theta}\end{pmatrix}\nabla_{(\theta,m)}\mathcal{F}_t\right\rangle_{q_t}. \tag{38}$$

where

$$\tilde{\Gamma}_\theta := \begin{pmatrix} 0 & -I_{d_\theta} \\ I_{d_\theta} & \gamma_\theta I_{d_\theta} \end{pmatrix} + 2T_{(\theta,m)}\nabla^2_{(\theta,m)}\mathcal{F}_t \begin{pmatrix} 0 & -I_{d_\theta} \\ I_{d_\theta} & \gamma_\theta I_{d_\theta} \end{pmatrix}.$$

By $\mathcal{F}$'s definition,

$$\nabla^2_{(\theta,m)}\mathcal{F} = \begin{pmatrix} \nabla^2_\theta \mathcal{F} & 0_{d_\theta} \\ 0_{d_\theta} & \eta_\theta I_{d_\theta} \end{pmatrix};$$

whence we see that

$$\begin{aligned}
\nabla^2_{(\theta,m)}\mathcal{F} \begin{pmatrix} 0_{d_\theta} & -I_{d_\theta} \\ I_{d_\theta} & \gamma_\theta I_{d_\theta} \end{pmatrix} &= \begin{pmatrix} 0_{d_\theta} & -\nabla^2_\theta \mathcal{F} \\ \eta_\theta I_{d_\theta} & \gamma_\theta \eta_\theta I_{d_\theta} \end{pmatrix} \\
\Rightarrow T_{(\theta,m)}\nabla^2_{(\theta,m)}\mathcal{F} \begin{pmatrix} 0_{d_\theta} & -I_{d_\theta} \\ I_{d_\theta} & \gamma_\theta I_{d_\theta} \end{pmatrix} &= \begin{pmatrix} \frac{\tau_{\theta m}\eta_\theta}{2} I_{d_\theta} & \frac{\tau_{\theta m}\gamma_\theta\eta_\theta}{2} I_{d_\theta} - \tau_\theta \nabla^2_\theta \mathcal{F} \\ \tau_m \eta_\theta I_{d_\theta} & \tau_m \gamma_\theta \eta_\theta I_{d_\theta} - \frac{\tau_{\theta m}}{2}\nabla^2_\theta \mathcal{F} \end{pmatrix}, \\
\Rightarrow \tilde{\Gamma}_\theta &= \begin{pmatrix} \tau_{\theta m}\eta_\theta I_{d_\theta} & [\tau_{\theta m}\gamma_\theta\eta_\theta - 1] I_{d_\theta} - 2\tau_\theta \nabla^2_\theta \mathcal{F}_t \\ [2\tau_m\eta_\theta + 1] I_{d_\theta} & [2\tau_m \eta_\theta + 1]\gamma_\theta I_{d_\theta} - \tau_{\theta m}\nabla^2_\theta \mathcal{F}_t \end{pmatrix}.
\end{aligned}$$

Given that, for any matrix A, we have $\langle \cdot, A\cdot \rangle = \langle \cdot, A^{sym}\cdot \rangle$ where $A^{sym} = \left(A + A^\top\right)/2$, then with $\Gamma_\theta := \frac{\tilde{\Gamma}_\theta + \tilde{\Gamma}_\theta^\top}{2}$, we obtain that

$$(37) = -\left\langle \nabla_{(\theta,m)}\mathcal{F}_t, \Gamma_\theta \nabla_{(\theta,m)}\mathcal{F}_t \right\rangle, \tag{39}$$

and hence that

$$\Gamma_\theta = \begin{pmatrix} \tau_{\theta m}\eta_\theta I_{d_\theta} & \left[\tau_m + \frac{\tau_{\theta m}\gamma_\theta}{2}\right] \eta_\theta I_{d_\theta} - \tau_\theta \nabla^2_\theta \mathcal{F}_t \\ \left[\tau_m + \frac{\tau_{\theta m}\gamma_\theta}{2}\right] \eta_\theta I_{d_\theta} - \tau_\theta \nabla^2_\theta \mathcal{F}_t & [2\tau_m\eta_\theta + 1]\gamma_\theta I_{d_\theta} - \tau_{\theta m}\nabla^2_\theta \mathcal{F}_t \end{pmatrix}. \tag{40}$$

We now turn our attention to (36). Defining $\mathcal{H}_t(x,u) := \sqrt{\frac{q_t(x,u)}{\rho_{\theta_t,\eta_x}(x,u)}}$ and Proposition F.2, we see that

$$\begin{aligned}
(36) = &- 4\left\langle \nabla_{(x,u)} \log \mathcal{H}_t, \begin{pmatrix} 0_{d_x} & -I_{d_x} \\ I_{d_x} & \gamma_x I_{d_x} \end{pmatrix} \nabla_{(x,u)} \log \mathcal{H}_t \right\rangle_{q_t} && (41) \\
&- 8\left\langle \nabla_{(x,u)} \frac{\nabla^*_{(x,u)} T_{(x,u)} \nabla_{(x,u)}\mathcal{H}_t}{\mathcal{H}_t}, \begin{pmatrix} 0_{d_x} & -I_{d_x} \\ I_{d_x} & \gamma_x I_{d_x} \end{pmatrix} \nabla_{(x,u)} \log \mathcal{H}_t \right\rangle_{q_t} && (42) \\
&+ 4\left\langle \nabla_{(x,u)}\nabla_{(\theta,m)} \left[\log \rho_{\theta_t}\right] T_{(\theta,m)}\nabla_{(\theta,m)}\mathcal{F}_t, \begin{pmatrix} 0_{d_x} & -I_{d_x} \\ I_{d_x} & \gamma_x I_{d_x} \end{pmatrix} \nabla_{(x,u)} \log \mathcal{H}_t \right\rangle_{q_t}. && (43)
\end{aligned}$$

Furthermore, we have

$$(41) = -4\gamma_x \left|\left|\nabla_u \log \mathcal{H}_t\right|\right|^2_{q_t} = -\gamma_x \left|\left|\nabla_u \log \frac{q_t}{\rho_{\theta_t,\eta_x}}\right|\right|^2_{q_t}.$$

In Proposition F.11, we show that (42) can be further simplified as follows:

$$(42) = -2\gamma_x \left\langle \nabla_{(x,u)}\nabla_u \left[\log \frac{q_t}{\rho_{\theta_t,\eta_x}}\right], T_{(x,u)}\nabla_{(x,u)}\nabla_u \left[\log \frac{q_t}{\rho_{\theta_t,\eta_x}}\right] \right\rangle_{q_t}$$

$$-\left\langle \nabla_{(x,u)} \log \frac{q_t}{\rho_{\theta_t,\eta_x}}, \tilde{\Gamma}_x \nabla_{(x,u)} \log \frac{q_t}{\rho_{\theta_t,\eta_x}} \right\rangle_{q_t},$$

where

$$\tilde{\Gamma}_x = \begin{pmatrix} \tau_{xu}\eta_x I_{d_x} & \frac{\tau_u+\tau_{xu}\gamma_x}{2}\eta_x I_{d_x} + \tau_x \nabla_x^2 \ell \\ \frac{\tau_u+\tau_{xu}\gamma_x}{2}\eta_x I + \tau_x \nabla_x^2 \ell & 2\gamma_x \tau_u \eta_x I_{d_x} + \tau_{xu}\nabla_x^2 \ell \end{pmatrix}.$$

We can then combine (41) and (42) to obtain

$$\begin{aligned}(41) + (42) = &-2\gamma_x \left\langle \nabla_{(x,u)} \nabla_u \left[\log \frac{q_t}{\rho_{\theta_t,\eta_x}}\right], T_{(x,u)} \nabla_{(x,u)} \nabla_u \left[\log \frac{q_t}{\rho_{\theta_t,\eta_x}}\right] \right\rangle_{q_t} \\ &- \left\langle \nabla_{(x,u)} \log \frac{q_t}{\rho_{\theta_t,\eta_x}}, \Gamma_x \nabla_{(x,u)} \log \frac{q_t}{\rho_{\theta_t,\eta_x}} \right\rangle_{q_t}.\end{aligned}$$

where $\Gamma_x = \tilde{\Gamma}_x + \begin{pmatrix} 0_{d_x} & 0_{d_x} \\ 0_{d_x} & \gamma_x I_{d_x} \end{pmatrix}$ is again positive semi-definite.

Given any p.s.d. matrix $M \in \mathbb{R}^{d\times d}$ and a general matrix $A \in \mathbb{R}^{d\times d}$, it holds generically that

$$\langle A, MA \rangle = \langle A, LL^\top A \rangle = \langle L^\top A, L^\top A \rangle = \|L^\top A\|^2 \geq 0,$$

where $M = LL^\top$ is Cholesky decomposition of M. Thus, we have the following upper bound:

$$\begin{aligned}(41) + (42) &\leq -\left\langle \nabla_{(x,u)} \log \frac{q_t}{\rho_{\theta_t,\eta_x}}, \Gamma_x \nabla_{(x,u)} \log \frac{q_t}{\rho_{\theta_t,\eta_x}} \right\rangle_{q_t} && (44)\\ &= -\left\langle \nabla_{(x,u)} \delta_q \mathcal{F}_t, \Gamma_x \nabla_{(x,u)} \delta_q \mathcal{F}_t \right\rangle_{q_t}. && (45)\end{aligned}$$

Cross Terms (38) and (43). We will now deal with the cross terms of (43) and (38). In Proposition F.17, we show the following upper-bound: for all $\varepsilon > 0$, it holds that

$$(43) + (38) \leq \varepsilon \left\langle \nabla_{(\theta,m)} \mathcal{F}_t, \Gamma_\times \nabla_{(\theta,m)} \mathcal{F}_t \right\rangle + \frac{1}{\varepsilon} \left\langle \nabla_{(x,u)} \delta_q \mathcal{F}_t, \nabla_{(x,u)} \delta_q \mathcal{F}_t \right\rangle_{q_t}, \tag{46}$$

where $\Gamma_\times = K^2 \begin{pmatrix} \tau_\theta^2 I_{d_\theta} & \tau_\theta \left(\frac{\tau_{xu}+\tau_{\theta m}}{2}\right) I_{d_\theta} \\ \tau_\theta \left(\frac{\tau_{xu}+\tau_{\theta m}}{2}\right) I_{d_\theta} & \left(\left(\frac{\tau_{xu}+\tau_{\theta m}}{2}\right)^2 + \tau_x^2\right) I_{d_\theta} \end{pmatrix}$.

Conclusion of Step 1. Combining the results in (39, 45, 46), we upper bound the time derivative of the Lyapunov function as

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_t \leq -\left\langle \nabla_{(\theta,m)} \mathcal{F}_t, S_{(\theta,m)} \nabla_{(\theta,m)} \mathcal{F}_t \right\rangle - \left\langle \nabla_{(x,u)} \delta_q \mathcal{F}_t, S_{(x,u)} \nabla_{(x,u)} \delta_q \mathcal{F}_t \right\rangle_{q_t},$$

where

$$S_{(x,u)} = \Gamma_x - \frac{1}{\varepsilon} I_{2d_x} = \begin{pmatrix} \left(\tau_{xu}\eta_x - \frac{1}{\varepsilon}\right) I_{d_x} & \frac{\tau_u+\tau_{xu}\gamma_x}{2}\eta_x I_{d_x} + \tau_x \nabla_x^2 \ell \\ \frac{\tau_u+\tau_{xu}\gamma_x}{2}\eta_x I_{d_x} + \tau_x \nabla_x^2 \ell & \left(\gamma_x(2\tau_u\eta_x + 1) - \frac{1}{\varepsilon}\right) I_{d_x} + \tau_{xu}\nabla_x^2 \ell \end{pmatrix}, \tag{47}$$

$$S_{(\theta,m)} = \Gamma_\theta - \varepsilon \Gamma_\times = \begin{pmatrix} s_\theta I_{d_\theta} & s_{\theta m} I_{d_\theta} - \tau_\theta \nabla_\theta^2 \mathcal{F}_t \\ s_{\theta m} I_{d_\theta} - \tau_\theta \nabla_\theta^2 \mathcal{F}_t & s_m I_{d_\theta} - \tau_{\theta m} \nabla_\theta^2 \mathcal{F}_t \end{pmatrix}, \tag{48}$$

with constants defined as

$$\begin{aligned} s_\theta &:= \tau_{\theta m}\eta_\theta - \varepsilon K^2 \tau_\theta^2, \quad s_{\theta m} := \tau_m \eta_\theta + \frac{\tau_{\theta m}\gamma_\theta}{2}\eta_\theta - \varepsilon K^2 \tau_\theta \frac{\tau_{xu}+\tau_{\theta m}}{2}, \\ s_m &:= (2\tau_m\eta_\theta + 1)\gamma_\theta - \varepsilon K^2 \left(\left(\frac{\tau_{xu}+\tau_{\theta m}}{2}\right)^2 + \tau_x^2\right). \end{aligned}$$

F.2. Proof of Step 2

By the definitions of $T_{(\theta,m)}$ and $T_{(x,u)}$ in (34), and those of $S_{(\theta,m)}$ and $S_{(x,u)}$ in (47, 48),

$$S_{(\theta,m)} - \varphi C\left(T_{(\theta,m)} + \frac{1}{2C}I_{d_\theta}\right) = \begin{pmatrix} \alpha_\theta I_{d_\theta} & \beta_\theta I_{d_\theta} - \tau_\theta \nabla_\theta^2 \mathcal{F} \\ \beta_\theta I_{d_\theta} - \tau_\theta \nabla_\theta^2 \mathcal{F} & \kappa_\theta I_{d_\theta} - \tau_{\theta m}\nabla_\theta^2 \mathcal{F}\end{pmatrix}, \tag{49}$$

$$S_{(x,u)} - \varphi C\left(T_{(x,u)} + \frac{1}{2C}I_{d_x}\right) = \begin{pmatrix} \alpha_x I_{d_x} & \beta_x I_{d_x} + \tau_x \nabla_x^2 \ell \\ \beta_x I_{d_x} + \tau_x \nabla_x^2 \ell & \kappa_x I_{d_x} + \tau_{xu}\nabla_x^2 \ell\end{pmatrix}. \tag{50}$$

where

$$\begin{aligned}
\alpha_\theta &:= \tau_{\theta m}\eta_\theta - \varepsilon K^2 \tau_\theta^2 - \varphi C\left(\tau_\theta + \frac{1}{2C}\right),\\
\beta_\theta &:= \left(\tau_m + \frac{\tau_{\theta m}\gamma_\theta}{2}\right)\eta_\theta - \varepsilon K^2 \tau_\theta\left(\frac{\tau_{xu}+\tau_{\theta m}}{2}\right) - \varphi C\frac{\tau_{\theta m}}{2},\\
\kappa_\theta &:= (2\tau_m\eta_\theta + 1)\gamma_\theta - \varepsilon K^2\left(\left(\frac{\tau_{xu}+\tau_{\theta m}}{2}\right)^2 + \tau_x^2\right) - \varphi C\left(\tau_m + \frac{1}{2C}\right),\\
\alpha_x &= \tau_{xu}\eta_x - \frac{1}{\varepsilon} - \varphi C\left(\tau_x + \frac{1}{2C}\right),\\
\beta_x &= \frac{(\tau_u + \tau_{xu}\gamma_x)\eta_x}{2} - \varphi C\frac{\tau_{xu}}{2},\\
\kappa_x &= \gamma_x(2\tau_u\eta_x + 1) - \frac{1}{\varepsilon} - \varphi C\left(\tau_u + \frac{1}{2C}\right).
\end{aligned}$$

Proposition F.1. *Defining $(\alpha_\theta, \beta_\theta, \kappa_\theta, \alpha_x, \beta_x, \kappa_x)$ as above, assume that $K \geq 2C_{\mathcal{E}}$ if*

$$\tau_{\theta m}K - \alpha_\theta - \kappa_\theta \leq 0, \tag{51a}$$
$$\tau_\theta^2 K^2 + (\tau_{\theta m}\alpha_\theta - 2\beta_\theta\tau_\theta)K - \alpha_\theta\kappa_\theta + \beta_\theta^2 \leq 0, \tag{51b}$$
$$\tau_\theta^2 K^2 - (\tau_{\theta m}\alpha_\theta - 2\beta_\theta\tau_\theta)K - \alpha_\theta\kappa_\theta + \beta_\theta^2 \leq 0, \tag{51c}$$
$$\tau_{xu}K - \alpha_x - \kappa_x \leq 0, \tag{51d}$$
$$\tau_x^2 K^2 - (\tau_{xu}\alpha_x - 2\beta_x\tau_x)K - \alpha_x\kappa_x + \beta_x^2 \leq 0, \tag{51e}$$
$$\tau_x^2 K^2 + (\tau_{xu}\alpha_x - 2\beta_x\tau_x)K - \alpha_x\kappa_x + \beta_x^2 \leq 0, \tag{51f}$$

hold for some rate $\varphi > 0$. The following lower bounds for $S_{(x,u)}$ and $S_{(\theta,m)}$ then hold:

$$S_{(x,u)} \succeq \varphi C\left(T_{(x,u)} + \frac{1}{2C}I_{d_x}\right), \quad S_{(\theta,m)} \succeq \varphi C\left(T_{(\theta,m)} + \frac{1}{2C}I_{d_\theta}\right),$$

where $C := \min\{C_{\mathcal{E}}, \eta_\theta, \eta_x\}$.

Proof. Note the matrices in (49, 50) have the same form as that in (66). As we show below,

$$-KI_{d_x} \preceq \nabla_x^2 \ell \preceq KI_{d_x}, \quad -KI_{d_\theta} \preceq \nabla_\theta^2 \mathcal{F} \preceq KI_{d_\theta}. \tag{52}$$

Then, applying Proposition F.18 tells us that (49, 50) are positive semidefinite if the conditions (51) are satisfied.

Now, to prove the inequalities in (52). Assumption 1 and (Nesterov, 2003, Lemma 1.2.2) imply the first two inequalities and

$$-KI_{d_\theta} \preceq \nabla_\theta^2 \ell \preceq KI_{d_\theta}.$$

The other two inequalities in (52) follow from the above. To see this, for $\nabla_\theta^2 \mathcal{F} \preceq KI_{d_\theta}$, we have

$$\left\langle v, (KI_{d_\theta} - \nabla_\theta^2 \mathcal{F})v\right\rangle = \int \left\langle v, (KI_{d_\theta} + \nabla_\theta^2 \ell)v\right\rangle q(\mathrm{d}x \times \mathrm{d}u),$$

and since $\langle v, \nabla_\theta^2[\ell]v\rangle \geq -K\|v\|^2$, we have shown that $\left\langle v, (KI_{d_\theta} - \nabla_\theta^2 \mathcal{F})v\right\rangle \geq 0$ implying the desired result of $\nabla_\theta^2 \mathcal{F} \preceq KI_{d_\theta}$. A similar argument can be made for $-KI_{d_\theta} \preceq \nabla_\theta^2 \mathcal{F}$. □

F.3. Examples of (51) holding

We verified the above with a symbolic calculator (see https://github.com/jenninglim/mpgd) written in Mathematica for the choices of $\eta_x, \eta_\theta, \gamma_x, \gamma_\theta$ hold for the following choices:

1. Rate $\varphi = \frac{1}{10}$, momentum parameters $\gamma_x = \frac{3}{2}$, $\gamma_\theta = 3$, $\eta_x = 2K$, $\eta_\theta = 2K$, and elements of the Lyapunov function $\tau_\theta = \frac{3}{8}, \tau_{\theta m} = \frac{5}{4}, \tau_m = 2, \tau_x = \frac{1}{4}, \tau_{xu} = \frac{5}{4}, \tau_u = \frac{7}{4}$, and $\varepsilon = 15$.
2. Rate $\varphi = \frac{1}{30}$, momentum parameters $\gamma_x = \frac{6}{5}$, $\gamma_\theta = \frac{5}{2}$, $\eta_x = 2K$, $\eta_\theta = 2K$, and elements of the Lyapunov function $\tau_\theta = \frac{3}{8}, \tau_{\theta m} = \frac{5}{4}, \tau_m = 2, \tau_x = \frac{1}{4}, \tau_{xu} = 1, \tau_u = \frac{3}{2}$, and $\varepsilon = 15$.

F.4. Supporting Proofs and Derivations

F.4.1. The derivative of $\mathcal{L}_t$ along the flow

Here, we prove (35,36).

Proposition F.2. *The derivative of the Lyapunov function $\mathcal{L}$ is given by*

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathcal{L}_t = -\left\langle \nabla_{(\theta,m)}\mathcal{L}_t, \begin{pmatrix} 0 & -I_{d_\theta} \\ I_{d_\theta} & \gamma_\theta I_{d_\theta} \end{pmatrix} \nabla_{(\theta,m)}\mathcal{F}_t \right\rangle - \left\langle \nabla_{(x,u)}\delta_q\mathcal{L}_t, \begin{pmatrix} 0 & -I_{d_x} \\ I_{d_x} & \gamma_\theta I_{d_x} \end{pmatrix} \nabla_{(x,u)}\delta_q\mathcal{F}_t \right\rangle_{q_t},$$

where the first variation is given by

$$\begin{aligned}\delta_q\mathcal{L}_t =& \delta_q\mathcal{F}_t + \delta_q\mathcal{L}_t^1 + \delta_q\mathcal{L}_t^2 \\ =& \log\frac{q_t}{\rho_{\theta_t,\eta_x}} + 1 - 2\left\langle \nabla_{(\theta,m)}\log\rho_{\theta_t}, T_{(\theta,m)}\nabla_{(\theta,m)}\mathcal{F}_t\right\rangle \\ &+ \frac{4}{\mathcal{H}_t}\nabla^*_{(x,u)}T_{(x,u)}\nabla_{(x,u)}\mathcal{H}_t,\end{aligned}$$

with ∇^* and $\mathcal{H}_t := \mathcal{H}_{\theta_t,q_t}$ defined in Proposition F.5.

Proof. From the linearity of $\mathcal{L}$, we have

$$\dot{\mathcal{L}}_t = \dot{\mathcal{F}}_t + \dot{\mathcal{L}}_t^1 + \dot{\mathcal{L}}_t^2,$$

where

$$\begin{aligned}\mathcal{L}_t^1 &:= \left\langle \nabla_{(\theta,m)}\mathcal{F}_t, T_{(\theta,m)}\nabla_{(\theta,m)}\mathcal{F}_t\right\rangle, \\ \mathcal{L}_t^2 &:= \left\langle \nabla_{(x,u)}\delta_q\mathcal{F}_t, T_{(x,u)}\nabla_{(x,u)}\delta_q\mathcal{F}_t\right\rangle_{q_t}.\end{aligned}$$

From Appendix C.2, it can be seen that the time derivatives of $\mathcal{F}_t$, $\mathcal{L}_t^1$, and $\mathcal{L}_t^2$, are given by

$$\begin{aligned}\dot{\mathcal{F}}_t &= \left\langle \nabla_{(\theta,m)}\mathcal{F}_t, (\dot\theta_t, \dot m_t)\right\rangle + \left\langle \nabla_{(x,u)}\delta_q\mathcal{F}_t, (\dot x_t, \dot u_t)\right\rangle_{q_t}, \\ \dot{\mathcal{L}}_t^1 &= \left\langle \nabla_{(\theta,m)}\mathcal{L}_t^1, (\dot\theta_t, \dot m_t)\right\rangle + \left\langle \nabla_{(x,u)}\delta_q\mathcal{L}_t^1, (\dot x_t, \dot u_t)\right\rangle_{q_t}, \\ \dot{\mathcal{L}}_t^2 &= \left\langle \nabla_{(\theta,m)}\mathcal{L}_t^2, (\dot\theta_t, \dot m_t)\right\rangle + \left\langle \nabla_{(x,u)}\delta_q\mathcal{L}_t^2, (\dot x_t, \dot u_t)\right\rangle_{q_t},\end{aligned}$$

where $(\dot\theta_t, \dot m_t) = \begin{pmatrix} 0 & -I_{d_\theta} \\ I_{d_\theta} & \gamma_\theta I_{d_\theta} \end{pmatrix}\nabla_{(\theta,m)}\mathcal{F}_t$ and $(\dot x_t, \dot u_t) = \begin{pmatrix} 0 & -I_{d_x} \\ I_{d_x} & \gamma_x I_{d_x} \end{pmatrix}\nabla_{(x,u)}\delta_q\mathcal{F}_t$. From Proposition F.3, Proposition F.4, and Proposition F.5, and linearity of the inner product, we obtain as desired. □

Proposition F.3. *The first variation of $\mathcal{F}$ is given by*

$$\delta_q\mathcal{F}[z](x,u) = \log\frac{q(x,u)}{\rho_{\theta,\eta_x}(x,u)} + 1,$$

where $z = (\theta, m, q)$.

Proof. It can be seen that $\mathcal{F}$ is (VI-I)-typed equation with $f((\theta, m), (x, u), q, g) = q \log \frac{q}{\rho_{\theta,\eta_x}(x)}$. Thus, using the formula (25), we obtain as its first variation

$$\delta_q \mathcal{F}[z](x, u) = \log \frac{q(x, u)}{\rho_{\theta,\eta_x}(x, u)} + 1,$$

Hence, we have as desired. □

For the first variation of $\mathcal{L}^1$, we have the following result:

Proposition F.4. *The first variation of $\mathcal{L}^1$ is given by*

$$\delta_q \mathcal{L}^1[z](x, u) = -2 \left\langle \nabla_{(\theta,m)} \log \rho_\theta(x), T_{(\theta,m)} \nabla_{(\theta,m)} \mathcal{F}[z] \right\rangle,$$

where $z = (\theta, m, q)$.

Proof. It can be seen that $\mathcal{L}^1$ is a variational integral of type (VI-II) with

$$\begin{aligned} f((\theta, m), (x, u), (x', u'), q, q') =& qq' \langle \nabla_\theta \log \rho_\theta(x), T_{\theta\theta} \nabla_\theta \log \rho_\theta(x') \rangle \\ &- q \langle \nabla_\theta \log \rho_\theta(x), T_{\theta m} m \rangle \\ &- q' \langle m, T_{m\theta} \nabla_\theta \log \rho_\theta(x') \rangle \\ &+ \langle m, T_{mm} m \rangle , \end{aligned}$$

where we write $T_{\theta\theta}$ for the upper left block of $T_{(\theta,m)}$, and similarly for $T_{\theta m}$, and T_{mm}.

Using the formula (26) and the fact that

$$\nabla_{(4)} f((\theta, m), (x, u), (x', u'), q, q') =q' \langle \nabla_\theta \log \rho_\theta(x), T_{\theta\theta} \nabla_\theta \log \rho_\theta(x') \rangle - \langle \nabla_\theta \log \rho_\theta(x), T_{\theta m} m \rangle ,$$

For the first term of (26), we have

$$\begin{aligned} &\int \nabla_{(4)} f((\theta, m), (x, u), (x', u'), q(x, u), q(x', u')) \mathrm{d}x' \mathrm{d}u' \\ &\qquad = - \left\langle \nabla_{(\theta,m)} \log \rho_\theta(x), T_{(\theta,m)} \nabla_{(\theta,m)} \mathcal{F}[z] \right\rangle , \end{aligned}$$

and by symmetry, it follows similarly for the last term of (26). We have obtained as desired. □

Proposition F.5. *The first variation of $\mathcal{L}^2$ is given by*

$$\delta_q \mathcal{L}^2[z] = \frac{4}{\mathcal{H}_{\theta,q}} \nabla^*_{(x,u)} T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_{\theta,q},$$

where $\mathcal{H}_{\theta,q}(x, u) := \sqrt{\frac{q(x,u)}{\rho_{\theta,\eta_x}(x,u)}}$ and the adjoint operator $\nabla^*_{(x,u)}$ (see Appendix F.4.2) is defined as

$$\nabla^*_{(x,u)}(f)(x, u) := - \left\langle \nabla_{(x,u)} \log \rho_{\theta,\eta_x}(x, u), f(x, u) \right\rangle - \nabla_{(x,u)} \cdot [f](x, u),$$

where $\nabla_x \cdot$ is the divergence operator w.r.t. x.

Proof. First note that $\mathcal{L}^2$ can be written equivalently as

$$\left\langle \frac{\nabla_{(x,u)} q}{q} - \nabla_{(x,u)} \log \rho_{\theta,\eta_x}, T_{(x,u)} \left[\frac{\nabla_{(x,u)} q}{q} - \nabla_{(x,u)} \log \rho_{\theta,\eta_x} \right] \right\rangle_q .$$

It can be seen that $\mathcal{L}^2$ is a variational integral of type (VI-I) with

$$f((\theta, m), (x, u), q, g) = q \left\langle \frac{g}{q} - \nabla_{(x,u)} \log \rho_{\theta,\eta_x}(x, u), T_{(x,u)} \left[\frac{g}{q} - \nabla_{(x,u)} \log \rho_{\theta,\eta_x}(x, u) \right] \right\rangle .$$

As

$$\begin{aligned}\nabla_{(3)} f((\theta, m),(x, u), q, g) =& \left\langle \frac{g}{q} - \nabla_{(x,u)} \log \rho_{\theta,\eta_x}(x,u), T_{(x,u)} \left[\frac{g}{q} - \nabla_{(x,u)} \log \rho_{\theta,\eta_x}(x,u)\right]\right\rangle \\ &- 2q \left\langle \frac{g}{q^2}, T_{(x,u)} \left[\frac{g}{q} - \nabla_{(x,u)} \log \rho_{\theta,\eta_x}(x,u)\right]\right\rangle \\ =& - \left\langle \frac{g}{q} + \nabla_{(x,u)} \log \rho_{\theta,\eta_x}(x,u), T_{(x,u)} \left[\frac{g}{q} - \nabla_{(x,u)} \log \rho_{\theta,\eta_x}(x,u)\right]\right\rangle,\end{aligned}$$

and

$$\nabla_{(4)} f((\nabla, m),(x, u), q, g) = 2T_{(x,u)} \left[\frac{g}{q} - \nabla_{(x,u)} \log \rho_{\theta,\eta_x}(x,u)\right],$$

using formula (25), we obtain

$$\delta_q \mathcal{L}^2[z](x,u) = - \left\langle \nabla_{(x,u)} \log q(x,u) + \nabla_{(x,u)} \log \rho_{\theta,\eta_x}(x,u), T_{(x,u)} \nabla_{(x,u)} \log \frac{q(x,u)}{\rho_{\theta,\eta_x}(x,u)}\right\rangle \tag{53a}$$

$$- \nabla_{(x,u)} \cdot \left(2T_{(x,u)} \nabla_{(x,u)} \log \frac{q(x,u)}{\rho_{\theta,\eta_x}(x,u)}\right). \tag{53b}$$

To obtain the desired result, note that we can rewrite

$$\begin{aligned}(53a) &= - \left\langle \nabla_{(x,u)} \log \sqrt{q} + \nabla_{(x,u)} \log \sqrt{\rho_{\theta,\eta_x}}, \frac{4}{\mathcal{H}_{\theta,q}} T_{(x,u)} \nabla_{(x,u)} \log \mathcal{H}_{\theta,q}\right\rangle \\ (53b) &= -\nabla_{(x,u)} \cdot \left(\frac{4}{\mathcal{H}_{\theta,q}} T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_{\theta,q}\right), \\ &= - \left\langle \nabla_{(x,u)} \frac{4}{\mathcal{H}_{\theta,q}}, T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_{\theta,q}\right\rangle - \frac{4}{\mathcal{H}_{\theta,q}} \nabla_{(x,u)} \cdot (T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_{\theta,q}) \\ &= \left\langle \nabla_{(x,u)} \log \mathcal{H}_{\theta,q}, \frac{4}{\mathcal{H}_{\theta,q}} T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_{\theta,q}\right\rangle - \frac{4}{\mathcal{H}_{\theta,q}} \nabla_{(x,u)} \cdot (T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_{\theta,q}).\end{aligned}$$

Hence, we have obtained the desired result. □

F.4.2. THE ADJOINT OF ∇

Throughout the proofs, we use ∇^* to denote the linear map that is an adjoint of ∇ (the Euclidean gradient operator).

In the view of ∇ as a linear map from $(L^2_\rho(\mathcal{X};\mathbb{R}), \langle\cdot,\cdot\rangle_\rho)$ to $(L^2_\rho(\mathcal{X};\mathcal{X}), \langle\cdot,\cdot\rangle_\rho)$ where $\mathcal{X} \in \{\mathbb{R}^{d_x}, \mathbb{R}^{d_x} \times \mathbb{R}^{d_x}\}$, $\rho \in \mathcal{P}(\mathcal{X})$, and

$$L^2_\rho(\mathcal{X};\mathcal{X}) := \{\text{sufficiently regular } v : \mathcal{X} \to \mathcal{X} \text{ such that } ||v||_\rho < \infty\}.$$

Then its adjoint ∇^* is given by

$$\nabla^* v := -\frac{1}{\rho} \nabla \cdot [\rho v] = - \langle \nabla \log \rho, v\rangle - \nabla \cdot v.$$

where $\nabla\cdot$ denotes the divergence operator. Whatever the gradient ∇ is taken with respect to, the adjoint ∇^* is taken with respect to the corresponding quantity.

We show the adjoint property for the case where $\mathcal{X} = \mathbb{R}^{d_x} \times \mathbb{R}^{d_x}$ as the other case follows similarly. To begin.

$$\begin{aligned}\langle \nabla f, v\rangle_\rho &= \int \langle \nabla f(x,u), v(x,u)\rangle \, \rho(\mathrm{d}x \times du) \\ &= \int \langle \nabla f(x,u), \rho(x,u) v(x,u)\rangle \, \mathrm{d}x \times du \\ &= - \int f(x,u) \nabla \cdot [\rho(x,u) v(x,u)] \, \mathrm{d}x \times du = \langle f, \nabla^* v\rangle_\rho,\end{aligned}$$

where we used integration by parts combined with the divergence theorem and the fact that ρ must vanish at the boundary to obtain the second equation.

Another case of interest is the view of ∇ as a linear map from $(L^2_\rho(\mathcal{X};\mathcal{X}'), \langle\cdot,\cdot\rangle_\rho)$ to $(L^2_\rho(\mathcal{X};\mathcal{X}\times\mathcal{X}'), \langle\cdot,\cdot\rangle_\rho)$, where the underlying inner product on the latter space is induced by the Frobenius inner product, and as before, $\mathcal{X},\mathcal{X}' \in \{\mathbb{R}^{d_x}, \mathbb{R}^{d_x}\times\mathbb{R}^{d_x}\}$, $\rho \in \mathcal{P}(\mathcal{X})$.

The adjoint of ∇ for some $M \in L^2(\mathcal{X}, \mathcal{X}\times\mathcal{X}')$ is defined as

$$(\nabla^* M) := -M^\top \nabla \log \rho - \nabla\cdot M \in \mathcal{X}',$$

where $\nabla \cdot M$ denotes the divergence operator for a matrix field defined as $(\nabla\cdot M)_i = \partial_j M_{ji}$ in Einstein's notation. Each element of $(\nabla^* M)_i$ is given by

$$(\nabla^* M)_i := -\left\langle \nabla \log\rho, M^\top e_i \right\rangle - \nabla\cdot(M^\top e_i),$$

where $\{e_i\}_i$ is the standard basis of $\mathcal{X}'$.

For the case $\mathcal{X} = \mathbb{R}^{d_x}\times\mathbb{R}^{d_x}$, the adjoint ∇^* is property can be shown as follows: let $M \in L^2(\mathcal{X},\mathcal{X}\times\mathcal{X}')$ and $v \in L^2(\mathcal{X},\mathcal{X}')$

$$\begin{aligned}
\langle \nabla v, M\rangle_\rho &= \int \sum_{i=1}^{d_\mathcal{X}} \sum_{j=1}^{d_{\mathcal{X}'}} (\nabla v)_{ij} M_{ij}\, \rho(\mathrm{d}x \times \mathrm{d}u) \\
&= \int \sum_{i=1}^{d_\mathcal{X}} \sum_{j=1}^{d_{\mathcal{X}'}} \partial_i[v_j]\, M_{ij}\, \rho(\mathrm{d}x\times\mathrm{d}u) = \int \sum_{j=1}^{d_{\mathcal{X}'}} \langle \nabla v_j, M e_j\rangle\, \rho(\mathrm{d}x\times\mathrm{d}u) \\
&= -\int \sum_{j=1}^{d_{\mathcal{X}'}} (v_j) \nabla\cdot(\rho M e_j)\, \mathrm{d}x\times\mathrm{d}u \\
&= -\int \sum_{j=1}^{d_{\mathcal{X}'}} (v_j)[\langle \nabla\log\rho, M e_j\rangle + \nabla\cdot(M e_j)]\, \rho(\mathrm{d}x\times\mathrm{d}u) \\
&= \langle \nabla^* M, v\rangle_\rho .
\end{aligned}$$

where, for the second line, we apply integration by parts combined with the divergence theorem and the fact that ρ vanishes at the boundary.

F.4.3. Commutator $[\cdot,\cdot]$

Another quantity widely used in the proof is the commutator denoted by $[\cdot,\cdot]$. It can be thought of as an indicator of whether two operators commute. It is defined as

$$[A,B]f = (AB)f - (BA)f.$$

If $[A,B]f = 0$, then A and B is said to commute.

In this section, we list propositions and their proofs that will be useful for the proof of Theorem 4.1.

Proposition F.6. *Let $f : \mathbb{R}^{d_x}\times\mathbb{R}^{d_x} \to \mathbb{R}$ be sufficiently regular, then we have*

$$[\nabla_u, \nabla_u^*]\nabla_u f = \eta_x \nabla_u f.$$

Proof. Expanding out the commutator, we obtain

$$\begin{aligned}
[\nabla_u, \nabla_u^*]\nabla_u f &= \nabla_u \nabla_u^* \nabla_u f - \nabla_u^* \nabla_u \nabla_u f \\
&\qquad - \nabla_u \left\langle \nabla_u \log \rho_{\theta,\eta_x}, \nabla_u f\right\rangle - \nabla_u[\nabla_u \cdot \nabla_u f] + \nabla_u^2 f\, \nabla_u \log\rho_{\theta,\eta_x} + \nabla_u\cdot(\nabla_u^2 f),
\end{aligned}$$

and since $(\nabla_u[\nabla_u\cdot\nabla_u f])_i = \partial_{u_i}\partial_{u_j}\partial_{u_j} f = \partial_{u_j}\partial_{u_j}\partial_{u_i} f = (\nabla_u\cdot(\nabla_u^2 f))_i$. We have that

$$[\nabla_u, \nabla_u^*]\nabla_u f = -\nabla_u^2 \log r_{\eta_x}\, \nabla_u f = \eta_x \nabla_u f.$$

□

Proposition F.7. *Let $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$ be sufficiently regularly, we have that ∇_u and $\nabla_x^* \nabla_x$ commutes. In other words, we have*

$$[\nabla_u, \nabla_x^* \nabla_x] f = 0.$$

Proof. We begin with the definition

$$[\nabla_u, \nabla_x^* \nabla_x] f = \nabla_u \nabla_x^* \nabla_x f - \nabla_x^* \nabla_x \nabla_u f,$$

For the first term, we have

$$\begin{aligned} \nabla_u \nabla_x^* \nabla_x f &= -\nabla_u [\langle \nabla_x \log \rho, \nabla_x f \rangle + \nabla_x \cdot \nabla_x f] \\ &= -\nabla_u \nabla_x f \, \nabla_x \log \rho - \nabla_u [\nabla_x \cdot \nabla_x f]. \end{aligned}$$

As for the second term, we have

$$\nabla_x^* \nabla_x \nabla_u f = -\nabla_u \nabla_x f \nabla_x \log \rho - \nabla_x \cdot (\nabla_x \nabla_u f).$$

Equality or $[\nabla_u, \nabla_x^* \nabla_x] f = 0$ follows from the following fact:

$$(\nabla_x \cdot (\nabla_x \nabla_u f))_i = \partial_{x_j} \partial_{x_j} \partial_{u_i} f = \partial_{u_i} \partial_{x_j} \partial_{x_j} f = (\nabla_u \nabla_x \cdot (\nabla_x f))_i.$$

□

Proposition F.8. *For $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$ be sufficiently regular, we have that ∇_x and $\nabla_u^* \nabla_u$ commutes, i.e., we have*

$$[\nabla_x, \nabla_u^* \nabla_u] f = 0.$$

Proof. Expanding out the commutator, we shall show that

$$[\nabla_x, \nabla_u^* \nabla_u] f = \nabla_x \nabla_u^* \nabla_u f - \nabla_u^* \nabla_u \nabla_x f = 0.$$

This follows since the first term can be written as

$$\begin{aligned} \nabla_x \nabla_u^* \nabla_u f &= \nabla_x \langle \nabla_u \log r_{\eta_x}, \nabla_u f \rangle + \nabla_x [\nabla_u \cdot (\nabla_u f)] \\ &= \nabla_x \nabla_u [f] \, \nabla_u \log r_{\eta_x} + \nabla_x [\nabla_u \cdot (\nabla_u f)], \end{aligned}$$

and, for the second term, we have

$$\nabla_u^* \nabla_u \nabla_x f = \nabla_u \nabla_x [f] \, \nabla_u \log r_{\eta_x} + \nabla_u \cdot (\nabla_u \nabla_x f).$$

Note that $\nabla_u \cdot (\nabla_u \nabla_x f) = \nabla_x [\nabla_u \cdot (\nabla_u f)]$, which can be seen from

$$(\nabla_x \nabla_u \cdot (\nabla_u f))_i = \partial_{x_i} \partial_{u_j} \partial_{u_j} f = \partial_{u_j} \partial_{u_j} \partial_{x_i} f = (\nabla_u \cdot (\nabla_u \nabla_x f))_i.$$

Hence, we have as desired. □

F.4.4. Another operator B

To simplify the notation, we drop the subscripts and write $\rho := \rho_{\theta, \eta_x}$. Let $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$, another operator (denoted by B) that we are interested in is defined as follow:

$$B[f] : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R} := \nabla_{(x,u)}^* \left[Q \nabla_{(x,u)} f \right]$$

where $Q := \begin{pmatrix} 0 & -I_{d_x} \\ I_{d_x} & 0 \end{pmatrix}$. By expanding out the adjoint, B can be written equivalently as

$$B[f] = -\left\langle \nabla_{(x,u)} \log \rho, Q \nabla_{(x,u)} f \right\rangle - \nabla_{(x,u)} \cdot \left(Q \nabla_{(x,u)} f \right)$$

$$= - \langle \nabla_x f, \nabla_u \log \rho \rangle + \langle \nabla_x \log \rho, \nabla_u f \rangle . \tag{54}$$

One can generalize the operator B to vector-valued functions $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}^n$ as follows (its equivalence when $n = 1$ is easy to see):

$$B[f] : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}^n := \nabla^*_{(x,u)} \left[Q \nabla_{(x,u)} f \right] .$$

This can be equivalently written as

$$\begin{aligned} B[f] &= - \nabla_{(x,u)} f^\top Q^\top \nabla_{(x,u)} \log \rho - \nabla_{(x,u)} \cdot \left(Q \nabla_{(x,u)} f \right) \\ &= - \nabla_{(x,u)} f^\top \begin{pmatrix} \nabla_u \log \rho \\ -\nabla_x \log \rho \end{pmatrix} , \end{aligned} \tag{55}$$

where we use the fact that

$$\left(\nabla_{(x,u)} \cdot \left(Q \nabla_{(x,u)} f \right) \right)_i = \partial_{(x,u)_j} \left[\begin{pmatrix} \nabla_u f \\ -\nabla_x f \end{pmatrix}_{ji} \right] = 0.$$

In the following proposition, we show that the operator is anti-symmetric.

Proposition F.9. *For all $f, g : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}^n$, we have $\langle B[f], g \rangle_\rho = - \langle f, B[g] \rangle_\rho$.*

Proof. From the definition, we have

$$\begin{aligned} \langle B[f], g \rangle_\rho &= \left\langle \nabla^*_{(x,u)} \left[Q \nabla_{(x,u)} f \right], g \right\rangle_\rho \\ &= \left\langle Q \nabla_{(x,u)} f, \nabla_{(x,u)} g \right\rangle_\rho , \end{aligned}$$

where we use the adjoint operator of ∇^* for the second equality. Continuing from the last line,

$$\langle B[f], g \rangle_\rho = \left\langle \nabla_{(x,u)} f, Q^\top \nabla_{(x,u)} g \right\rangle_\rho = - \left\langle f, \nabla^*_{(x,u)} Q \nabla_{(x,u)} g \right\rangle_\rho = - \langle f, B[g] \rangle_\rho .$$

□

Proposition F.10. *For $f : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$, we have*

$$[\nabla_u, B] f = \eta_x \nabla_x f, \quad \textit{and} \quad [\nabla_x, B] f = \nabla_x^2 \log \rho \, \nabla_u f$$

Proof. For the first equality, we begin by expanding out the commutator

$$[\nabla_u, B] f = \nabla_u B[f] - B \nabla_u [f].$$

From Equation (54), we have

$$\nabla_u B[f] = -\nabla_u \nabla_x f \, \nabla_u \log \rho - \nabla_u^2 \log \rho \, \nabla_x f + \nabla_u^2 f \, \nabla_x \log \rho,$$

and, from Equation (55), we have

$$B \nabla_u [f] = -\nabla_u \nabla_{(x,u)} f \begin{pmatrix} \nabla_u \log \rho \\ -\nabla_x \log \rho \end{pmatrix} = -\nabla_u \nabla_x f \, \nabla_u \log \rho + \nabla_u^2 f \, \nabla_x \log \rho,$$

then, we must have

$$[\nabla_u, B] f = -\nabla_u^2 \log \rho \, \nabla_x f = \eta_x \nabla_x f.$$

For the second inequality, we begin similarly

$$[\nabla_x, B] f = \nabla_x B[f] - B \nabla_x [f].$$

For the first term, from Equation (54), we have

$$\nabla_x B[f] = -\nabla_x^2 f \, \nabla_u \log \rho + \nabla_x^2 \log \rho \, \nabla_u f + \nabla_x \nabla_u f \, \nabla_x \log \rho.$$

For the second term, from Equation (55), we have

$$B\nabla_x[f] = -\nabla_x \nabla_{(x,u)} f \begin{pmatrix} \nabla_u \log \rho \\ -\nabla_x \log \rho \end{pmatrix} = -\nabla_x^2 f \, \nabla_u \log \rho + \nabla_x \nabla_u f \, \nabla_x \log \rho.$$

Hence, we have

$$[\nabla_x, B]f = [\nabla_x, B]f = \nabla_x^2 \log \rho \, \nabla_u f.$$

□

F.4.5. SIMPLIFYING (42)

Proposition F.11 (Simplifying (42))**.** *We have that*

$$\begin{aligned}(42) = &- 2\gamma_x \left\langle \nabla_{(x,u)} \nabla_u \left[\log \frac{q_t}{\rho_t}\right], T_{(x,u)} \nabla_{(x,u)} \nabla_u \left[\log \frac{q_t}{\rho_t}\right] \right\rangle_{q_t} \\ &- \left\langle \nabla_{(x,u)} \log \frac{q_t}{\rho_t}, \tilde{\Gamma}_x \nabla_{(x,u)} \log \frac{q_t}{\rho_t} \right\rangle_{q_t},\end{aligned}$$

where $\rho_t := \rho_{\theta_t, \eta_x}$ *and*

$$\tilde{\Gamma}_x = \frac{1}{2} \begin{pmatrix} 2\tau_{xu}\eta_x I_{d_x} & (\tau_u \eta_x + \tau_{xu}\gamma_x \eta_x) I_{d_x} + 2\tau_x \nabla_x^2[\ell] \\ (\tau_u \eta_x + \tau_{xu}\gamma_x \eta_x) I_{d_x} + 2\tau_x \nabla_x^2[\ell] & 4\gamma_x \tau_u \eta_x I_{d_x} + 2\tau_{xu} \nabla_x^2[\ell] \end{pmatrix}.$$

Proof. Recall that $\mathcal{H}_t := \sqrt{\frac{q_t}{\rho_{\theta_t}}}$. We begin by applying the quotient rule to obtain the fact that

$$\begin{aligned}&\nabla_{(x,u)} \left(\frac{\nabla^*_{(x,u)} T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_t}{\mathcal{H}_t} \right) \\ =&\frac{\nabla_{(x,u)} \nabla^*_{(x,u)} \left[T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_t\right]}{\mathcal{H}_t} - \frac{\nabla_{(x,u)} \mathcal{H}_t \nabla^*_{(x,u)} \left[T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_t\right]}{\mathcal{H}_t^2}.\end{aligned}$$

Then, we can write (42) as

$$(42) = - 8 \left\langle \frac{\nabla_{(x,u)} \nabla^*_{(x,u)} \left[T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_t\right]}{\mathcal{H}_t}, \begin{pmatrix} 0_{d_x} & -I_{d_x} \\ I_{d_x} & \gamma_x I_{d_x} \end{pmatrix} \nabla_{(x,u)} \log \mathcal{H}_t \right\rangle_{q_t} \tag{56}$$

$$+ 8 \left\langle \frac{\nabla_{(x,u)} \mathcal{H}_t \nabla^*_{(x,u)} \left[T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_t\right]}{\mathcal{H}_t^2}, \begin{pmatrix} 0_{d_x} & -I_{d_x} \\ I_{d_x} & \gamma_x I_{d_x} \end{pmatrix} \nabla_{(x,u)} \log \mathcal{H}_t \right\rangle_{q_t}. \tag{57}$$

We can simplify the terms individually given in Proposition F.13 and Proposition F.12 for (56) and (57) Hence, we can write (42) equivalently the following:

$$(42) = 16\gamma_x \left\langle \nabla_{(x,u)} \nabla_u \left[\mathcal{H}_t\right], T_{(x,u)} \frac{\nabla_{(x,u)} \mathcal{H}_t}{\mathcal{H}_t} \otimes \nabla_u \mathcal{H}_t \right\rangle_{\rho_t} \tag{58a}$$

$$- 8\gamma_x \left\langle \frac{\nabla_{(x,u)} \mathcal{H}_t}{\mathcal{H}_t} \otimes \nabla_u \mathcal{H}_t, T_{(x,u)} \frac{\nabla_{(x,u)} \mathcal{H}_t}{\mathcal{H}_t} \otimes \nabla_u \mathcal{H}_t \right\rangle_{\rho_t} \tag{58b}$$

$$- 8\gamma_x \left\langle \nabla_{(x,u)} \nabla_u \left[\mathcal{H}_t\right], T_{(x,u)} \nabla_{(x,u)} \nabla_u \left[\mathcal{H}_t\right] \right\rangle_{\rho_t} \tag{58c}$$

$$- 4\tau_{xu}\eta_x \|\nabla_x \mathcal{H}_t\|^2_{\rho_t} \tag{58d}$$
$$- 4 \left\langle \nabla_x \mathcal{H}_t, \left[\left(\tau_u \eta_x + \tau_{xu}\gamma_x\eta_x\right) I_{d_x} + 2\tau_x \nabla_x^2 \left[\ell\right]\right] \nabla_u \mathcal{H}_t \right\rangle_{\rho_t} \tag{58e}$$
$$- 8 \left\langle \nabla_u \mathcal{H}_t, \left[\tau_u\eta_x\gamma_x I_{d_x} + \frac{\tau_{xu}}{2}\nabla_x^2 \left[\ell\right]\right] \nabla_u \mathcal{H}_t \right\rangle_{\rho_t}. \tag{58f}$$

We can rewrite the sum of (58a), (58b), (58c) as

$$\begin{aligned}
&(58a) + (58b) + (58c)\\
&= -8\gamma_x \left\langle \nabla_{(x,u)}\nabla_u \left[\mathcal{H}_t\right] - \frac{\nabla_{(x,u)}\mathcal{H}_t}{\mathcal{H}_t} \otimes \nabla_u \mathcal{H}_t, T_{(x,u)} \left(\nabla_{(x,u)}\nabla_u \left[\mathcal{H}_t\right] - \frac{\nabla_{(x,u)}\mathcal{H}_t}{\mathcal{H}_t} \otimes \nabla_u \mathcal{H}_t \right) \right\rangle_{\rho_t}.
\end{aligned}$$

Since we have that $\partial_x\partial_u \mathcal{H}_t = \partial_x(\mathcal{H}_t \partial_u \log \mathcal{H}_t) = (\partial_x \mathcal{H}_t)(\partial_u \log \mathcal{H}_t) + \mathcal{H}_t(\partial_x\partial_u \log \mathcal{H}_t)$, we can write the above as

$$\begin{aligned}
(58a) + (58b) + (58c) &= -8\gamma_x \left\langle \nabla_{(x,u)}\nabla_u \left[\log \mathcal{H}_t\right], T_{(x,u)}\nabla_{(x,u)}\nabla_u \left[\log \mathcal{H}_t\right]\right\rangle_{q_t}\\
&= -2\gamma_x \left\langle \nabla_{(x,u)}\nabla_u \left[\log \frac{q_t}{\rho_t}\right], T_{(x,u)}\nabla_{(x,u)}\nabla_u \left[\log \frac{q_t}{\rho_t}\right]\right\rangle_{q_t}.
\end{aligned}$$

As for the other terms, we have that

$$\begin{aligned}
(58d) + (58e) + (58f) =& - 4 \left\langle \nabla_{(x,u)} \log \mathcal{H}_t, \tilde{\Gamma}_x \nabla_{(x,u)} \log \mathcal{H}_t \right\rangle_{q_t}\\
=& - \left\langle \nabla_{(x,u)} \log \frac{q_t}{\rho_t}, \tilde{\Gamma}_x \nabla_{(x,u)} \log \frac{q_t}{\rho_t} \right\rangle_{q_t},
\end{aligned}$$

where

$$\tilde{\Gamma}_x := \frac{1}{2}\begin{pmatrix} 2\tau_{xu}\eta_x I_{d_x} & \left(\tau_u\eta_x + \tau_{xu}\gamma_x\eta_x\right) I_{d_x} + 2\tau_x\nabla_x^2\ell \\ \left(\tau_u\eta_x + \tau_{xu}\gamma_x\eta_x\right) I_{d_x} + 2\tau_x\nabla_x^2\ell & 4\gamma_x\tau_u\eta_x I_{d_x} + 2\tau_{xu}\nabla_x^2\ell \end{pmatrix}.$$

We have as desired. □

Proposition F.12 (Simplifying (57))**.** *We have*

$$\begin{aligned}
(57) =& 16\gamma_x \left\langle \nabla_{(x,u)}\nabla_u \left[\mathcal{H}_t\right], T_{(x,u)} \frac{\nabla_{(x,u)}\mathcal{H}_t}{\mathcal{H}_t} \otimes \nabla_u \mathcal{H}_t \right\rangle_{\rho_t}\\
&- 8\gamma_x \left\langle \frac{\nabla_{(x,u)}\mathcal{H}_t}{\mathcal{H}_t} \otimes \nabla_u \mathcal{H}_t, T_{(x,u)} \frac{\nabla_{(x,u)}\mathcal{H}_t}{\mathcal{H}_t} \otimes \nabla_u \mathcal{H}_t \right\rangle_{\rho_t}.
\end{aligned}$$

Proof. Noting that $\frac{1}{\mathcal{H}_t}\nabla^*_{(x,u)}\left[T_{(x,u)}\nabla_{(x,u)}\mathcal{H}_t\right]$ is scalar-valued and using $\mathcal{H}_t^2$ to change the inner product from $\langle\cdot,\cdot\rangle_{q_t}$ to $\langle\cdot,\cdot\rangle_{\rho_t}$, we obtain

$$\begin{aligned}
(57) =& 8 \left\langle \frac{\nabla^*_{(x,u)}\left[T_{(x,u)}\nabla_{(x,u)}\mathcal{H}_t\right]}{\mathcal{H}_t}, \left\langle \nabla_{(x,u)}\mathcal{H}_t, \begin{pmatrix} 0_{d_x} & -I_{d_x} \\ I_{d_x} & \gamma_x I_{d_x}\end{pmatrix} \nabla_{(x,u)}\mathcal{H}_t \right\rangle \right\rangle_{\rho_t}\\
=& 8\gamma_x \left\langle \frac{\nabla^*_{(x,u)}\left[T_{(x,u)}\nabla_{(x,u)}\mathcal{H}_t\right]}{\mathcal{H}_t}, \|\nabla_u \mathcal{H}_t\|_2^2 \right\rangle_{\rho_t}.
\end{aligned}$$

Then, using the adjoint of operator $\nabla^*_{(x,u)}$ as was described in Appendix F.4.2, we obtain

$$(57) = 8\gamma_x \left\langle \nabla_{(x,u)} \frac{\|\nabla_u \mathcal{H}_t\|_2^2}{\mathcal{H}_t}, T_{(x,u)}\nabla_{(x,u)}\mathcal{H}_t \right\rangle_{\rho_t}.$$

Then, using the quotient rule, we have that

$$
\begin{aligned}
(57) =& 8\gamma_x \left\langle \frac{\nabla_{(x,u)} \|\nabla_u \mathcal{H}_t\|^2}{\mathcal{H}_t}, T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_t \right\rangle_{\rho_t} - 8\gamma_x \left\langle \frac{\|\nabla_u \mathcal{H}_t\|_2^2 \nabla_{(x,u)} \mathcal{H}_t}{\mathcal{H}_t^2}, T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_t \right\rangle_{\rho_t} \\
=& 16\gamma_x \left\langle \frac{\nabla_{(x,u)} \nabla_u \left[\mathcal{H}_t\right] \nabla_u \mathcal{H}_t}{\mathcal{H}_t}, T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_t \right\rangle_{\rho_t} - 8\gamma_x \left\langle \frac{\|\nabla_u \mathcal{H}_t\|^2 \nabla_{(x,u)} \mathcal{H}_t}{\mathcal{H}_t^2}, T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_t \right\rangle_{\rho_t}.
\end{aligned}
$$

Finally, we obtain

$$
\begin{aligned}
(57) =& 16\gamma_x \left\langle \nabla_{(x,u)} \nabla_u \left[\mathcal{H}_t\right], T_{(x,u)} \frac{\nabla_{(x,u)} \mathcal{H}_t}{\mathcal{H}_t} \otimes \nabla_u \mathcal{H}_t \right\rangle_{\rho_t} \\
& - 8\gamma_x \left\langle \frac{\nabla_{(x,u)} \mathcal{H}_t}{\mathcal{H}_t} \otimes \nabla_u \mathcal{H}_t, T_{(x,u)} \frac{\nabla_{(x,u)} \mathcal{H}_t}{\mathcal{H}_t} \otimes \nabla_u \mathcal{H}_t \right\rangle_{\rho_t},
\end{aligned}
$$

where $\otimes$ denotes the outer product, and we use the fact that $\langle Mv, u \rangle = \langle M, u \otimes v \rangle_F$ and

$$
\|\nabla_u \mathcal{H}_t\|^2 \nabla_{(x,u)} \mathcal{H}_t = \left(\nabla_{(x,u)} \mathcal{H}_t \otimes \nabla_u \mathcal{H}_t\right) \nabla_u \mathcal{H}_t.
$$

□

Proposition F.13 (Simplifying (56))**.** *We have*

$$
\begin{aligned}
(56) =& - 8\gamma_x \left\langle \nabla_{(x,u)} \nabla_u \left[\mathcal{H}_t\right], T_{(x,u)} \nabla_{(x,u)} \nabla_u \left[\mathcal{H}_t\right] \right\rangle_{\rho_t} \\
& - 4\tau_{xu} \eta_x \|\nabla_x \mathcal{H}_t\|_{\rho_t}^2 \\
& - 8 \left\langle \nabla_u \mathcal{H}_t, \left(\tau_x \eta_x \gamma_x I_{d_x} + \frac{\tau_{xu}}{2} \nabla_x^2 \log \rho_t\right) \nabla_u \mathcal{H}_t \right\rangle_{\rho_t} \\
& - 4 \left\langle \nabla_x \mathcal{H}_t, \left(\left[\tau_x \eta_x + \tau_{xu} \gamma_x \eta_x\right] I_{d_x} + 2\tau_x \nabla_x^2 \log \rho_t\right) \nabla_u \mathcal{H}_t \right\rangle_{\rho_t},
\end{aligned}
$$

where $\rho_t := \rho_{\theta_t, \eta_x}$.

Proof. We begin by changing the inner product from $\langle \cdot, \cdot \rangle_{q_t}$ to $\langle \cdot, \cdot \rangle_{\rho_t}$

$$
(56) = -8 \left\langle \nabla_{(x,u)} \nabla_{(x,u)}^* \left[T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_t\right], \begin{pmatrix} 0_{d_x} & -I_{d_x} \\ I_{d_x} & \gamma_x I_{d_x} \end{pmatrix} \nabla_{(x,u)} \mathcal{H}_t \right\rangle_{\rho_t}.
$$

Since we have

$$
\nabla_{(x,u)}^* \left[T_{(x,u)} \nabla_{(x,u)} \mathcal{H}_t\right] = \nabla_x^* \left[\tau_x \nabla_x \mathcal{H}_t + \frac{\tau_{xu}}{2} \nabla_u \mathcal{H}_t\right] + \nabla_u^* \left[\frac{\tau_{xu}}{2} \nabla_x \mathcal{H}_t + \tau_u \nabla_u \mathcal{H}_t\right],
$$

we obtain

$$
\begin{aligned}
(56) =& - 4\tau_{xu} \left\langle \nabla_{(x,u)} \left[\nabla_x^* \nabla_u \mathcal{H}_t + \nabla_u^* \nabla_x \mathcal{H}_t\right], \begin{pmatrix} 0_{d_x} & -I_{d_x} \\ I_{d_x} & \gamma_x I_{d_x} \end{pmatrix} \nabla_{(x,u)} \mathcal{H}_t \right\rangle_{\rho_t} && \text{(59a)} \\
& - 8\tau_x \left\langle \nabla_{(x,u)} \left[\nabla_x^* \left(\nabla_x \mathcal{H}_t\right)\right], \begin{pmatrix} 0_{d_x} & -I_{d_x} \\ I_{d_x} & \gamma_x I_{d_x} \end{pmatrix} \nabla_{(x,u)} \mathcal{H}_t \right\rangle_{\rho_t} && \text{(59b)} \\
& - 8\tau_u \left\langle \nabla_{(x,u)} \left[\nabla_u^* \left(\nabla_u \mathcal{H}_t\right)\right], \begin{pmatrix} 0_{d_x} & -I_{d_x} \\ I_{d_x} & \gamma_x I_{d_x} \end{pmatrix} \nabla_{(x,u)} \mathcal{H}_t \right\rangle_{\rho_t}. && \text{(59c)}
\end{aligned}
$$

In Proposition F.14, F.15 and F.16 we deal with these respective terms:

Proposition F.14 (Simplifying (59a))**.** *We have*

$$\begin{aligned}(59a) = & - 4\tau_{xu}\eta_x\|\nabla_x\mathcal{H}_t\|^2_{\rho_t} \\ & - 4\tau_{xu}\left\langle\nabla_u\mathcal{H}_t, \nabla_x^2\left[\ell\right]\,\nabla_u\mathcal{H}_t\right\rangle_{\rho_t} \\ & - 8\gamma_x\tau_{xu}\left\langle\nabla_u^2\left[\mathcal{H}_t\right], \nabla_u\nabla_x\left[\mathcal{H}_t\right]\right\rangle_{\rho_t} \\ & - 4\gamma_x\eta_x\tau_{xu}\left\langle\nabla_x\mathcal{H}_t, \nabla_u\mathcal{H}_t\right\rangle_{\rho_t}.\end{aligned}$$

Proposition F.15 (Simplifying (59b))**.** *We have*

$$(59b) = -8\tau_x\gamma_x\left\|\nabla_x\nabla_u\mathcal{H}_t\right\|^2_{\rho_t} - 8\tau_x\left\langle\nabla_x\mathcal{H}_t, \nabla_x^2\left[\ell\right]\nabla_u\mathcal{H}_t\right\rangle_{\rho_t},$$

Proposition F.16 (Simplifying (59c))**.**

$$(59c) = -8\tau_u\gamma_x\left\|\nabla_u^2\mathcal{H}_t\right\|^2_{\rho_t} - 8\tau_u\eta_x\gamma_x\left\|\nabla_u\mathcal{H}_t\right\|^2_{\rho_t} - 4\tau_u\eta_x\langle\nabla_x\mathcal{H}_t, \nabla_u\mathcal{H}_t\rangle_{\rho_t}.$$

Their proofs can be found below.

Thus, summing up the results of Proposition F.14, Proposition F.15 and Proposition F.16, we obtain

$$\begin{aligned}(56) =& (59a) + (59b) + (59c) \\ =& - 8\tau_{xu}\gamma_x\left\langle\nabla_u^2\mathcal{H}_t, \nabla_u\nabla_x\left[\mathcal{H}_t\right]\right\rangle_{\rho_t} && (60a) \\ & - 8\tau_u\gamma_x\left\|\nabla_u^2\mathcal{H}_t\right\|^2_{\rho_t} && (60b) \\ & - 8\tau_x\gamma_x\left\langle\nabla_x\nabla_u\mathcal{H}_t, \nabla_x\nabla_u\mathcal{H}_t\right\rangle_{\rho_t} && (60c) \\ & - 4\tau_{xu}\eta_x\|\nabla_x\mathcal{H}_t\|^2_{\rho_t} \\ & - 8\left\langle\nabla_u\mathcal{H}_t, \left(\tau_u\eta_x\gamma_x I_{d_x} + \frac{\tau_{xu}}{2}\nabla_x^2\left[\ell\right]\right)\nabla_u\mathcal{H}_t\right\rangle_{\rho_t} \\ & - 4\left\langle\nabla_x\mathcal{H}_t, \left[\left(\tau_u\eta_x + \tau_{xu}\gamma_x\eta_x\right)I_{d_x} + 2\tau_x\nabla_x^2\left[\ell\right]\right]\nabla_u\mathcal{H}_t\right\rangle_{\rho_t}.\end{aligned}$$

Since we can write the following sum equivalently as:

$$(60a) + (60b) + (60c) = -8\gamma_x\left\langle\nabla_{(x,u)}\nabla_u\left[\mathcal{H}_t\right], T_{(x,u)}\nabla_{(x,u)}\nabla_u\left[\mathcal{H}_t\right]\right\rangle_{\rho_t},$$

we obtain as desired. □

Proof of Proposition F.14. Beginning from the definition, we have

$$\begin{aligned}(59a) =& -4\tau_{xu}\left\langle\nabla_{(x,u)}\left[\nabla_x^*\nabla_u\mathcal{H}_t + \nabla_u^*\nabla_x\mathcal{H}_t\right], \begin{pmatrix}0_{d_x} & 0_{d_x} \\ 0_{d_x} & \gamma_x I_{d_x}\end{pmatrix}\nabla_{(x,u)}\mathcal{H}_t\right\rangle_{\rho_t} \\ & -4\tau_{xu}\left\langle\nabla_{(x,u)}\left[\nabla_x^*\nabla_u\mathcal{H}_t + \nabla_u^*\nabla_x\mathcal{H}_t\right], \begin{pmatrix}0_{d_x} & -I_{d_x} \\ I_{d_x} & 0_{d_x}\end{pmatrix}\nabla_{(x,u)}\mathcal{H}_t\right\rangle_{\rho_t} \\ =& -4\gamma_x\tau_{xu}\left\langle\nabla_u\left[\nabla_x^*\nabla_u\mathcal{H}_t + \nabla_u^*\nabla_x\mathcal{H}_t\right], \nabla_u\mathcal{H}_t\right\rangle_{\rho_t} && (61a) \\ & -4\tau_{xu}\left\langle\nabla_{(x,u)}\left[\nabla_x^*\nabla_u\mathcal{H}_t + \nabla_u^*\nabla_x\mathcal{H}_t\right], \begin{pmatrix}0_{d_x} & -I_{d_x} \\ I_{d_x} & 0_{d_x}\end{pmatrix}\nabla_{(x,u)}\mathcal{H}_t\right\rangle_{\rho_t}. && (61b)\end{aligned}$$

We will deal with (61a) and (61b) separately.

Using the adjoints of $\nabla_u^*, \nabla_u, \nabla_x$, we obtain

$$(61a) = -4\gamma_x\tau_{xu}\left\langle \nabla_u\mathcal{H}_t, \nabla_x\nabla_u^*\nabla_u\mathcal{H}_t\right\rangle_{\rho_t} - 4\gamma_x\tau_{xu}\left\langle \nabla_x\mathcal{H}_t, \nabla_u\nabla_u^*\nabla_u\mathcal{H}_t\right\rangle_{\rho_t},$$

From Proposition F.8, we have that ∇_x commutes with $\nabla_u^*\nabla_u$

$$(61a) = -4\gamma_x\tau_{xu}\left\langle \nabla_u\mathcal{H}_t, \nabla_u^*\nabla_u\nabla_x\left[\mathcal{H}_t\right]\right\rangle_{\rho_t} - 4\gamma_x\tau_{xu}\left\langle \nabla_x\mathcal{H}_t, \nabla_u\nabla_u^*\nabla_u\mathcal{H}_t\right\rangle_{\rho_t},$$

We can write this in terms of the commutator $[\cdot,\cdot]$ (defined in Appendix F.4.3) using the fact that $[\nabla_u, \nabla_u^*]\nabla_u f = \eta_x\nabla_u f$ (as shown in Proposition F.6). Thus, we have

$$\begin{aligned}(61a) &= -8\gamma_x\tau_{xu}\left\langle \nabla_u\mathcal{H}_t, \nabla_u^*\nabla_u\nabla_x\left[\mathcal{H}_t\right]\right\rangle_{\rho_t} - 4\gamma_x\tau_{xu}\left\langle \nabla_x\mathcal{H}_t, [\nabla_u, \nabla_u^*]\nabla_u\mathcal{H}_t\right\rangle_{\rho_t} \\ &= -8\gamma_x\tau_{xu}\left\langle \nabla_u^2\left[\mathcal{H}_t\right], \nabla_u\nabla_x\left[\mathcal{H}_t\right]\right\rangle_{\rho_t} - 4\gamma_x\eta_x\tau_{xu}\left\langle \nabla_x\mathcal{H}_t, \nabla_u\mathcal{H}_t\right\rangle_{\rho_t}.\end{aligned} \tag{62}$$

As for (61b), we start by writing it in terms of B (see Appendix F.4.4) $B[\mathcal{H}_t] := \nabla_{(x,u)}^* Q\nabla_{(x,u)}\mathcal{H}_t$, we can write

$$(61b) = -4\tau_{xu}\left\langle \nabla_{(x,u)}^*\nabla_{(u,x)}\mathcal{H}_t, B[\mathcal{H}_t]\right\rangle_{\rho_t}.$$

Using the fact that the adjoint of B denoted by B^* w.r.t. the inner product $\langle\cdot,\cdot\rangle_\rho$ is $B^* = -B$ (see Proposition F.9) we obtain

$$\begin{aligned}(61b) &= -4\tau_{xu}\left\langle \nabla_{(u,x)}\mathcal{H}_t, \nabla_{(x,u)}B[\mathcal{H}_t]\right\rangle_{\rho_t} \\ &= -4\tau_{xu}\left(\langle \nabla_x\mathcal{H}_t, \nabla_u B[\mathcal{H}_t]\rangle_{\rho_t} + \langle \nabla_u\mathcal{H}_t, \nabla_x B[\mathcal{H}_t]\rangle_{\rho_t}\right) \\ &= -4\tau_{xu}\left(\langle \nabla_x\mathcal{H}_t, \nabla_u B[\mathcal{H}_t]\rangle_{\rho_t} + \langle \nabla_u\mathcal{H}_t, B\nabla_x[\mathcal{H}_t]\rangle_{\rho_t} + \langle \nabla_u\mathcal{H}_t, [\nabla_x, B][\mathcal{H}_t]\rangle_{\rho_t}\right) \\ &= -4\tau_{xu}\left(\langle \nabla_x\mathcal{H}_t, \nabla_u B[\mathcal{H}_t]\rangle_{\rho_t} - \langle B\nabla_u\mathcal{H}_t, \nabla_x[\mathcal{H}_t]\rangle_{\rho_t} + \langle \nabla_u\mathcal{H}_t, [\nabla_x, B][\mathcal{H}_t]\rangle_{\rho_t}\right) \\ &= -4\tau_{xu}\left(\langle \nabla_x\mathcal{H}_t, [\nabla_u, B][\mathcal{H}_t]\rangle_{\rho_t} + \langle \nabla_u\mathcal{H}_t, [\nabla_x, B][\mathcal{H}_t]\rangle_{\rho_t}\right).\end{aligned}$$

From Proposition F.10, we have $[\nabla_u, B][\mathcal{H}_t] = \eta_x\nabla_x\mathcal{H}_t$ and $[\nabla_x, B][\mathcal{H}_t] = \nabla_x^2[\ell]\nabla_u\mathcal{H}_t$, and so

$$(61b) = -4\tau_{xu}\eta_x\|\nabla_x\mathcal{H}_t\|_{\rho_t}^2 - 4\tau_{xu}\left\langle \nabla_u\mathcal{H}_t, \nabla_x^2\left[\ell\right]\nabla_u\mathcal{H}_t\right\rangle_{\rho_t}. \tag{63}$$

Thus, summing up (62) and (63), we have as desired

$$\begin{aligned}(59a) = (62) + (63) = &-4\tau_{xu}\eta_x\|\nabla_x\mathcal{H}_t\|_{\rho_t}^2 \\ &-4\tau_{xu}\left\langle \nabla_u\mathcal{H}_t, \nabla_x^2\left[\ell\right]\nabla_u\mathcal{H}_t\right\rangle_{\rho_t} \\ &-8\gamma_x\tau_{xu}\left\langle \nabla_u^2\left[\mathcal{H}_t\right], \nabla_u\nabla_x\left[\mathcal{H}_t\right]\right\rangle_{\rho_t} \\ &-4\gamma_x\eta_x\tau_{xu}\left\langle \nabla_x\mathcal{H}_t, \nabla_u\mathcal{H}_t\right\rangle_{\rho_t}.\end{aligned}$$

□

Proof of Proposition F.15. Beginning from the definition and following in the same expansion as in Proposition F.14, and using the adjoint of $\nabla_{(x.u)}$ (i.e., $\nabla_{(x.u)}^*$) and adjoint of ∇_x^* (i.e., ∇_x^*) we obtain the following:

$$(59b) = -8\tau_x\gamma_x\left\langle \nabla_u\left[\nabla_x^*\left(\nabla_x\mathcal{H}_t\right)\right], \nabla_u\mathcal{H}_t\right\rangle_{\rho_t} - 8\tau_x\left\langle \nabla_x\mathcal{H}_t, \nabla_x B[\mathcal{H}_t]\right\rangle_{\rho_t}.$$

Writing this in terms of the commutator, we get

$$\begin{aligned}(59b) &= -8\tau_x\gamma_x\left\langle \nabla_u\left[\nabla_x^*\left(\nabla_x\mathcal{H}_t\right)\right], \nabla_u\mathcal{H}_t\right\rangle_{\rho_t} - 8\tau_x\left\langle \nabla_x\mathcal{H}_t, [\nabla_x, B]\mathcal{H}_t + B\nabla_x\mathcal{H}_t\right\rangle_{\rho_t} \\ &= -8\tau_x\gamma_x\left\langle \nabla_x\nabla_u\mathcal{H}_t, \nabla_x\nabla_u\mathcal{H}_t\right\rangle_{\rho_t} - 8\tau_x\left\langle \nabla_x\mathcal{H}_t, \nabla_x^2\left[\ell\right]\nabla_u\mathcal{H}_t\right\rangle_{\rho_t},\end{aligned}$$

where, for the first term, we use the fact that ∇_u commutes with $\nabla_x^*\nabla_x$ (see Proposition F.7); and, for the second term, we use Proposition F.10 and $\langle \nabla_x\mathcal{H}_t, B\nabla_x\mathcal{H}_t\rangle_{\rho_t} = 0$ from antisymmetry (see Proposition F.9). □

Proof of Proposition F.16. Beginning with the definition and following in the same expansion as in Proposition F.14, we obtain

$$\begin{aligned}(59c) =& -8\tau_u\gamma_x \left\langle \nabla_u\nabla_u^* \left[\nabla_u\mathcal{H}_t\right], \nabla_u\mathcal{H}_t\right\rangle_{\rho_t} \\ &- 8\tau_u \left\langle \nabla_u\mathcal{H}_t, \nabla_u\nabla_{(x,u)}^* \left[\begin{pmatrix} 0_{d_x} & -I_{d_x} \\ I_{d_x} & 0_{d_x}\end{pmatrix} \nabla_{(x,u)}\mathcal{H}_t\right]\right\rangle_{\rho_t} \\ =& -8\tau_u\gamma_x \left\langle \nabla_u\nabla_u^* \left[\nabla_u\mathcal{H}_t\right], \nabla_u\mathcal{H}_t\right\rangle_{\rho_t} - 4\tau_u \left\langle \nabla_u\mathcal{H}_t, \nabla_u B[\mathcal{H}_t]\right\rangle_{\rho_t}.\end{aligned}$$

Writing in terms of the commutator $[\cdot,\cdot]$ and B, we have equivalently

$$\begin{aligned}(59c) =& -8\tau_u\gamma_x \left\langle \nabla_u\nabla_u \left[\mathcal{H}_t\right], \nabla_u\nabla_u\left[\mathcal{H}_t\right]\right\rangle_{\rho_t} - 8\tau_u\gamma_x \left\langle \left[\nabla_u, \nabla_u^*\right]\nabla_u\mathcal{H}_t, \nabla_u\mathcal{H}_t\right\rangle_{\rho_t} \\ &- 4\tau_u \left\langle \nabla_u\mathcal{H}_t, B\nabla_u\mathcal{H}_t\right\rangle_{\rho_t} - 4\tau_u\eta_x\langle \nabla_x\mathcal{H}_t, \nabla_u\mathcal{H}_t\rangle_{\rho_t},\end{aligned}$$

where, for the last term, we use the fact of Proposition F.10 to obtain

$$\left\langle \nabla_u\mathcal{H}_t, [\nabla_u, B]\mathcal{H}_t\right\rangle_{\rho_t} = \eta_x \left\langle \nabla_x\mathcal{H}_t, \nabla_u\mathcal{H}_t\right\rangle_{\rho_t}$$

Using the antisymmetric property of B, we have $\left\langle \nabla_u\mathcal{H}_t, B\nabla_u\mathcal{H}_t\right\rangle_{\rho_t} = 0$; and, from Proposition F.6, the fact that $[\nabla_u, \nabla_u^*]\nabla_u f = \eta_x\nabla_u f$, we have

$$\begin{aligned}(59c) =& -8\tau_u\gamma_x \left\langle \nabla_u^2\left[\mathcal{H}_t\right], \nabla_u^2\left[\mathcal{H}_t\right]\right\rangle_{\rho_t} \\ &- 8\tau_u\eta_x\gamma_x \left\langle \nabla_u\mathcal{H}_t, \nabla_u\mathcal{H}_t\right\rangle_{\rho_t} \\ &- 4\tau_u\eta_x\langle \nabla_x\mathcal{H}_t, \nabla_u\mathcal{H}_t\rangle_{\rho_t}.\end{aligned}$$

□

F.4.6. Bounding the Cross terms (38) and (43).

Proposition F.17. *For all $\varepsilon > 0$, we have that*

$$(38) + (43) \leq \varepsilon \left\langle \nabla_{(\theta,m)}\mathcal{F}_t, \Gamma_\times \nabla_{(\theta,m)}\mathcal{F}_t\right\rangle + \frac{1}{\varepsilon}\|\nabla_{(x,u)}\delta_q\mathcal{F}_t\|_{q_t}^2,$$

where $\Gamma_\times = K^2 \begin{pmatrix} \tau_\theta^2 I_{d_\theta} & \tau_\theta\left(\frac{\tau_{xu}+\tau_{\theta m}}{2}\right) I_{d_\theta} \\ \tau_\theta\left(\frac{\tau_{xu}+\tau_{\theta m}}{2}\right) I_{d_\theta} & \left(\left(\frac{\tau_{xu}+\tau_{\theta m}}{2}\right)^2 + \tau_x^2\right) I_{d_\theta}\end{pmatrix}$.

Proof. Because $T_{(x,u)}$ is symmetric,

$$(38) = -2\left\langle A_t\nabla_{(\theta,m)}\mathcal{F}_t, \nabla_{(x,u)}\delta_q\mathcal{F}_t\right\rangle_{q_t}, \quad (43) = 2\left\langle A_t^{'}\nabla_{(\theta,m)}\mathcal{F}_t, \nabla_{(x,u)}\delta_q\mathcal{F}_t\right\rangle_{q_t},$$

where

$$\begin{aligned} A_t &:= T_{(x,u)}\nabla_{(x,u)}\nabla_{(\theta,m)}\left[\delta_q\mathcal{F}_t\right]\begin{pmatrix} 0_{d_\theta} & -I_{d_\theta} \\ I_{d_\theta} & \gamma_\theta I_{d_\theta}\end{pmatrix}, \\ A_t^{'} &:= \begin{pmatrix} 0_{d_x} & I_{d_x} \\ -I_{d_x} & \gamma_x I_{d_x}\end{pmatrix}\nabla_{(x,u)}\nabla_{(\theta,m)}\left[\log\rho_{\theta_t}\right] T_{(\theta,m)}.\end{aligned}$$

Given that

$$\nabla_{(x,u)}\nabla_{(\theta,m)}\left[\delta_q\mathcal{F}_t\right] = -\nabla_{(x,u)}\nabla_{(\theta,m)}\left[\log\rho_{\theta_t,\eta_x}\right] = -\begin{pmatrix} \nabla_x\nabla_\theta\left[\log\rho_{\theta_t}\right] & 0_{d_x\times d_\theta} \\ 0_{d_x\times d_\theta} & 0_{d_x\times d_\theta}\end{pmatrix},$$

we can re-write A_t and $A_t^{'}$ as

$$A_t = \begin{pmatrix} 0_{d_x \times d_\theta} & \tau_x \nabla_x \nabla_\theta \left[\log \rho_{\theta_t}\right] \\ 0_{d_x \times d_\theta} & \frac{\tau_{xu}}{2} \nabla_x \nabla_\theta \left[\log \rho_{\theta_t}\right] \end{pmatrix}, \quad A_t^{'} = - \begin{pmatrix} 0_{d_x \times d_\theta} & 0_{d_x \times d_\theta} \\ \tau_\theta \nabla_x \nabla_\theta \left[\log \rho_{\theta_t}\right] & \frac{\tau_{\theta m}}{2} \nabla_x \nabla_\theta \left[\log \rho_{\theta_t}\right] \end{pmatrix}.$$

Hence,

$$(38) + (43) = 2 \left\langle A_t \nabla_{(\theta,m)} \mathcal{F}_t, \nabla_{(x,u)} \delta_q \mathcal{F}_t \right\rangle_{q_t}$$

where

$$\tilde{A}_t := A_t^{'} - A_t = - \begin{pmatrix} 0 & \tau_x \nabla_x \nabla_\theta \left[\log \rho_{\theta_t}\right] \\ \tau_\theta \nabla_x \nabla_\theta \left[\log \rho_{\theta_t}\right] & \frac{(\tau_{xu} + \tau_{\theta m})}{2} \nabla_x \nabla_\theta \left[\log \rho_{\theta_t}\right] \end{pmatrix}.$$

Fix any $\varepsilon > 0$. Applying the Cauchy-Schwarz inequality and Young's inequality, we obtain that

$$\begin{aligned} (38) + (43) \leq & 2 \left|\left| \tilde{A}_t \nabla_{(\theta,m)} \mathcal{F}_t \right|\right|_{q_t} \left|\left| \nabla_{(x,u)} \delta_q \mathcal{F}_t \right|\right|_{q_t} \\ \leq & \varepsilon \left|\left| \tilde{A}_t \nabla_{(\theta,m)} \mathcal{F}_t \right|\right|_{q_t}^2 + \frac{1}{\varepsilon} \left|\left| \nabla_{(x,u)} \delta_q \mathcal{F}_t \right|\right|_{q_t}^2 . \end{aligned} \tag{64}$$

But,

$$\begin{aligned} \|\tilde{A}_t \nabla_{(\theta,m)} \mathcal{F}_t\|_{q_t}^2 = & \|\tau_x \nabla_x \nabla_\theta \left[\log \rho_{\theta_t}\right] \nabla_m \mathcal{F}_t\|_{q_t}^2 \\ & + \left\| \nabla_x \nabla_\theta \left[\log \rho_{\theta_t}\right] \left[\tau_\theta \nabla_\theta \mathcal{F}_t + \frac{\tau_{xu} + \tau_{\theta m}}{2} \nabla_m \mathcal{F}_t \right] \right\|_{q_t}^2 . \end{aligned}$$

Given that $\nabla_\theta \mathcal{F}(z)$ and $\nabla_m \mathcal{F}(z)$ are constant in x, the above reads

$$\begin{aligned} \|\tilde{A}_t \nabla_{(\theta,m)} \mathcal{F}_t\|_{q_t}^2 = & \|\nabla_x \left\langle \nabla_\theta \log \rho_{\theta_t}, \tau_x \nabla_m \mathcal{F}_t \right\rangle \|_{q_t}^2 \\ & + \left\| \nabla_x \left\langle \nabla_\theta \log \rho_{\theta_t}, \left(\tau_\theta \nabla_\theta \mathcal{F}_t + \frac{\tau_{xu} + \tau_{\theta m}}{2} \nabla_m \mathcal{F}_t \right) \right\rangle \right\|_{q_t}^2 . \end{aligned} \tag{65}$$

Fix any θ, v in $\mathbb{R}^{d_\theta}$. Because $\nabla_{(\theta,x)} \ell$ is K-Lipschitz (Assumption 1), the function

$$f_{v,\theta}(x) := \langle \nabla_\theta \log \rho_\theta(x), v \rangle = \langle \nabla_\theta \ell(\theta, x), v \rangle \quad \forall x \in \mathbb{R}^{d_x},$$

is $(K\|v\|)$-Lipschitz: by the Cauchy-Schwarz inequality,

$$\begin{aligned} f_{v,\theta}(x) - f_{v,\theta}(x') = & \langle \nabla_\theta \log \ell(\theta, x) - \nabla_\theta \ell(\theta, x'), v \rangle \leq \|\nabla_\theta \ell(\theta, x) - \nabla_\theta \ell(\theta, x')\| \|v\|, \\ \leq & \|\nabla_{(\theta,m)} \ell(\theta, x) - \nabla_{(\theta,m)} \ell(\theta, x')\| \|v\| \leq K \|v\| \|x - x'\| \quad \forall x, x' \in \mathbb{R}^{d_x}. \end{aligned}$$

Recalling that Lipschitz seminorms can be estimated by suprema of the norm of the gradient (e.g., see Van Handel (2014, Lemma 4.3)), we then see that

$$\|\nabla_x f_{v,\theta}\|_{q_t}^2 \leq \sup_{x \in \mathbb{R}^{d_x}} \|\nabla_x f_{v,\theta}(x)\|^2 \leq K^2 \|v\|^2.$$

Applying the above inequality to (65), we find that

$$\begin{aligned} \|\tilde{A}_t \nabla_{(\theta,m)} \mathcal{F}_t\|_{q_t}^2 \leq & K^2 \left[\|\tau_x \nabla_m \mathcal{F}_t\|^2 + \left\| \tau_\theta \nabla_\theta \mathcal{F}_t + \left(\frac{\tau_{xu} + \tau_{\theta m}}{2} \right) \nabla_m \mathcal{F}_t \right\|^2 \right] \\ = & \left\langle \nabla_{(\theta,m)} \mathcal{F}_t, \Gamma_\times \nabla_{(\theta,m)} \mathcal{F}_t \right\rangle . \end{aligned}$$

Plugging the above into (64) yields the desired inequality. □

F.4.7. POSITIVE SEMI-DEFINITENESS CONDITIONS

In this section, we establish sufficient conditions to ensure that a matrix with the form described in (66).

Proposition F.18. *Given $\alpha, \beta, \kappa, a, c \in \mathbb{R}$, let A be a symmetric matrix with the following form:*

$$A = \begin{pmatrix} \alpha I & \beta I + aH \\ \beta I + aH & \kappa I + cH \end{pmatrix}, \tag{66}$$

where H is a symmetric matrix that satisfies $-KI \preceq H \preceq KI$ for some $K > 0$. If A satisfies the following conditions:

$$\begin{cases} -cK - \alpha - \kappa \leq 0 & \textit{if } c \leq 0, \\ cK - \alpha - \kappa \leq 0 & \textit{otherwise}. \end{cases}$$
$$a^2K^2 - (c\alpha - 2\beta a)K - \alpha\kappa + \beta^2 \leq 0,$$
$$a^2K^2 + (c\alpha - 2\beta a)K - \alpha\kappa + \beta^2 \leq 0,$$

then, we have that A is positive semi-definite.

Proof. We prove this by showing that if the conditions are satisfied, then the eigenvalues of A are non-negative.

The eigenvalues l of A satisfy its characteristic equation

$$\det(A - lI) = 0.$$

We have that

$$\det(A - lI) = \det((\alpha - l)(\kappa - l)I + c(\alpha - l)H - (\beta I + aH)^2),$$

where we use the fact that $(\kappa I + cH)$ and $(\beta I + aH)$ are symmetric matrices and so their multiplication commutes (e.g., see Silvester (2000)).

Let $H = U\Lambda U^\top$ be the eigenvalue decomposition of H, then observe that

$$\det(A - lI) = \det(U(((\alpha - l)(\kappa - l) - \beta^2)I + (c(\alpha - l) - 2\beta a)\Lambda - a^2\Lambda^2)U^\top).$$

Hence, we obtain for each eigenvalue Λ_i of H the following constraints:

$$((\alpha - l)(\kappa - l) - \beta^2) + (c(\alpha - l) - 2\beta a)\Lambda_i - a^2\Lambda_i^2 = 0.$$

These constraints can be written equivalently as

$$l^2 + (-c\Lambda_i - \alpha - \kappa)\, l + \big(\alpha\kappa - \beta^2 - a^2\Lambda_i^2 + (c\alpha - 2\beta a)\Lambda_i\big) = 0.$$

Since the equality constraint is a quadratic function in l, we can utilize the quadratic formula to solve for l. It can be verified that the discriminant to the above quadratic equation is non-negative (hence has real roots). To ensure that l is positive, we require that

$$-c\Lambda_i - \alpha - \kappa \leq 0, \tag{67}$$
$$a^2\Lambda_i^2 - (c\alpha - 2\beta a)\Lambda_i - \alpha\kappa + \beta^2 \leq 0, \tag{68}$$

for all $\Lambda_i \in [-K, K]$. The constaints on Λ_i come from the fact that $-KI \preceq H \preceq KI$ and the min-max theorem.

Since (67) is a linear function and (68) is a quadratic function in Λ_i, we only require that the inequalities are satisfied at the end points. Namely, we obtain the following conditions

$$\begin{cases} -cK - \alpha - \kappa \leq 0 & \text{if } c \leq 0, \\ cK - \alpha - \kappa \leq 0 & \text{otherwise}. \end{cases}$$
$$a^2K^2 - (c\alpha - 2\beta a)K - \alpha\kappa + \beta^2 \leq 0,$$
$$a^2K^2 + (c\alpha - 2\beta a)K - \alpha\kappa + \beta^2 \leq 0.$$

This concludes the proof. □

G. Existence and Uniqueness of Strong solutions to (15)

In this section, we show the existence and uniqueness of the McKean–Vlasov SDE (15) under Lipschitz assumption 1. The structure is as follows:

1. We begin by showing that $\mathcal{F}$ has Lipschitz θ-gradients (Proposition G.1).
2. Then, we show that the drift, defined in Equation (69), is Lipschitz (Proposition G.2).
3. Finally, we prove the existence and uniqueness (Proposition 3.1).

In this section, we write the SDE (15) equivalently as

$$\mathrm{d}(\vartheta, \Upsilon)_t = b(\Upsilon_t, \vartheta_t, \mathrm{Law}(\Upsilon_t))\mathrm{d}t + \boldsymbol{\beta}\mathrm{d}W_t, \tag{69}$$

where $\vartheta_t = (\theta_t, m_t) \in \mathbb{R}^{2d_\theta}$, $\Upsilon_t = (X_t, U_t) \in \mathbb{R}^{2d_x}$, $\boldsymbol{\beta} = \sqrt{2}\mathrm{Diag}([0_{2d_\theta}, 0_{d_x}, 1_{d_x}])$ (where $\mathrm{Diag}(x)$ returns a Diagonal matrix whose elements are given by x), and $b : \mathbb{R}^{2d_x} \times \mathbb{R}^{2d_\theta} \times \mathcal{P}(\mathbb{R}^{2d_x}) \times \mathbb{R}^{2d_\theta} \to \mathbb{R}^{2d_\theta+2d_x}$ defined as

$$b((x,u),(\theta,m),q) = \begin{pmatrix} \eta_\theta m \\ -\gamma_\theta\eta_\theta m - \nabla_\theta \mathcal{F}(\theta, q) \\ \eta_x u \\ -\gamma_x\eta_x u + \nabla_x \ell(\theta, x) \end{pmatrix}.$$

We now prove that $\nabla_\theta \mathcal{F}$ is Lipschitz. Since the θ-gradients do not depend on the momentum parameter m, we will drop the dependence on m for brevity.

Proposition G.1 ($\mathcal{F}$ has Lipschitz θ-gradient). *Under Assumption 1, we have that $\mathcal{F}$ is Lipschitz,i.e., there exist a constant $K_{\mathcal{F}} > 0$ such that the following inequality holds:*

$$\|\nabla_\theta \mathcal{F}(\theta, q) - \nabla_\theta \mathcal{F}(\theta', q')\| \leq K_{\mathcal{F}}(\|\theta - \theta'\| + \mathsf{W}_1(q, q')),$$

for all $\theta, \theta' \in \mathbb{R}^{d_\theta}$ and $q, q' \in \mathcal{P}(\mathbb{R}^{2d_x})$.

Proof. From the definition, and adding and subtracting the same quantities and triangle inequality, we obtain

$$\|\nabla_\theta \mathcal{F}(\theta, q) - \nabla_\theta \mathcal{F}(\theta', q')\| \leq \|\nabla_\theta \mathcal{F}(\theta, q) - \nabla_\theta \mathcal{F}(\theta', q)\| + \|\nabla_\theta \mathcal{F}(\theta', q) - \nabla_\theta \mathcal{F}(\theta', q')\|.$$

We treat the terms on the RHS separately. For the first, from Jensen's inequality, we obtain

$$\begin{aligned} \|\nabla_\theta \mathcal{F}(\theta, q) - \nabla_\theta \mathcal{F}(\theta', q)\| &= \left\|\int \nabla_\theta \ell(\theta, x) - \nabla_\theta \ell(\theta', x)\, q_X(\mathrm{d}x)\right\| \\ &\leq \int \|\nabla_\theta \ell(\theta, x) - \nabla_\theta \ell(\theta', x)\|\; q_X(\mathrm{d}x) \\ &\leq K\|\theta - \theta'\|. \end{aligned}$$

As for the other term, we have

$$\begin{aligned} \|\nabla_\theta \mathcal{F}(\theta', q) - \nabla_\theta \mathcal{F}(\theta', q')\| &\leq \left\|\int \nabla_\theta \ell(\theta', x)(q - q')(x, u)\mathrm{d}x\mathrm{d}u\right\| \\ &\leq K \int \frac{1}{K}\|\nabla_\theta \ell(\theta', x)\|\, |q - q'|(x, u)\mathrm{d}x\mathrm{d}u \\ &\overset{(a)}{\leq} 2K\mathsf{W}_1(q, q'). \end{aligned}$$

For (a), we use the fact that $(x, u) \mapsto \|\nabla_\theta \ell(\theta', x)\|$ is K-Lipschitz (under assumption 1) (and so the map $x \mapsto \frac{1}{K}\|\nabla_\theta \ell(\theta', x)\|$ is 1-Lipschitz), from the dual representation of W_1 we have as desired.

To conclude, by combining the two bounds, we have shown that $\nabla_\theta \mathcal{F}$ is Lipschitz with constant $K_{\mathcal{F}} = 2K$. $\square$

We now prove that the drift of the SDE (69) is Lipschitz.

Proposition G.2 (Lipschitz Drift)**.** *Under Assumption 1, the drift b is Lipschitz, i.e., there is some constant $K_b > 0$ such that the following inequality holds:*

$$\|b(\Upsilon, \vartheta, q) - b(\Upsilon', \vartheta', q')\| \le K_b(\|\Upsilon - \Upsilon'\| + \|\vartheta - \vartheta'\| + \mathsf{W}_1(q, q')),$$

for all $\vartheta, \vartheta' \in \mathbb{R}^{d_\theta}$, $\Upsilon, \Upsilon' \in \mathbb{R}^{2d_\theta}$, and $q, q' \in \mathcal{P}(\mathbb{R}^{2d_x})$.

Proof. We begin with the definition, apply the triangle inequality, and use the fact that from the concavity of $\sqrt{\cdot}$ we have $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b > 0$ to obtain

$$\begin{aligned}
\|b(\Upsilon, \vartheta, q) - b(\Upsilon', \vartheta', q')\| &\le \eta_\theta(1+\gamma_\theta)\,\|m - m'\| + \|\nabla_\theta \mathcal{F}(\theta, q) - \nabla_\theta \mathcal{F}(\theta', q')\| \\
&\quad + \eta_x(1+\gamma_x)\|u - u'\| + \|\nabla_x \ell(\theta, x) - \nabla_x \ell(\theta', x')\| \\
&\le \eta_\theta(1+\gamma_\theta)\,\|m - m'\| + K_{\mathcal{F}}(\|\theta - \theta'\| + \mathsf{W}_1(q, q')) \\
&\quad + \eta_x(1+\gamma_x)\|u - u'\| + K(\|\theta - \theta'\| + \|x - x'\|) \\
&\le \sqrt{2}\max\{\eta_\theta(1+\gamma_\theta), K_{\mathcal{F}} + K\}\|\vartheta - \vartheta'\| \\
&\quad + \sqrt{2}\max\{\eta_x(1+\gamma_x), K, K_{\mathcal{F}}\}(\|\Upsilon - \Upsilon'\| + \mathsf{W}_1(q, q')),
\end{aligned}$$

where we use the Lipschitz assumption 1 on ℓ, and Proposition G.1. Hence we obtained as desired. □

Proof of Proposition 3.1. The proof is similar to Carmona (2016, Theorem 1.7) but with key generalizations to the product space.

Fix some $\nu \in C([0, T], \mathbb{R}^{2d_\theta} \times \mathcal{P}(\mathbb{R}^{2d_x}))$. We denoted by ν^ϑ and ν^Υ the projection to the $\mathbb{R}^{2d_\theta}$ and $\mathcal{P}(\mathbb{R}^{2d_x})$ components respectively. Consider substituting ν_t into (69) in place of the $\mathrm{Law}(\Upsilon_t)$ and ϑ_t, from Carmona (2016, Theorem 1.2), we have existence and uniqueness of the strong solution for some initial point $(\vartheta, \Upsilon)_0$. More explicitly, we have

$$(\vartheta, \Upsilon)^\nu_t = (\vartheta, \Upsilon)_0 + \int_0^t b(\Upsilon^\nu_t, \nu^\vartheta_s, \nu^\Upsilon_s)\mathrm{d}s + \int_0^t \beta \mathrm{d}W_t.$$

for $t \in [0, T]$. We define the operator $F_T : C([0, T], \mathbb{R}^{2d_\theta} \times \mathcal{P}(\mathbb{R}^{2d_x})) \to C([0, T], \mathbb{R}^{2d_\theta} \times \mathcal{P}(\mathbb{R}^{2d_x}))$ as

$$F_T : \nu \to (t \mapsto (\vartheta^\nu_t, \mathrm{Law}(\Upsilon^\nu_t))).$$

Clearly, if the process $(\vartheta, \Upsilon)_t$ is a solution to (69) then the function $t \mapsto (\vartheta_t, \mathrm{Law}(\Upsilon_t))$ is a fixed point to the operator F_T, and vice versa. Now we establish the existence and uniqueness of the fixpoint of the operator F_T.

We begin by endowing the space $\mathbb{R}^{2d_\theta} \times \mathcal{P}(\mathbb{R}^{2d_x})$ with the metric:

$$\mathsf{d}((\vartheta, q), (\vartheta', q')) = \sqrt{\|\vartheta - \vartheta'\|^2 + \mathsf{W}_2(q, q')^2}.$$

Note that the metric space $(\mathbb{R}^{2d_\theta} \times \mathcal{P}(\mathbb{R}^{2d_x}), d)$ is complete (Villani, 2009).

First note that using Jensen's inequality, b is Lipschitz, and the fact that $(a+b+c+d)^2 \le 4(a^2+b^2+c^2+d^2)$ we obtain

$$\begin{aligned}
\|\vartheta^\nu_t - \vartheta^{\nu'}_t\|^2 + \mathbb{E}[\|\Upsilon^\nu_t - \Upsilon^{\nu'}_t\|^2] &= \mathbb{E}\left[\left\|\int_0^t b(\Upsilon^\nu_s, \nu^\vartheta_s, \nu^\Upsilon_s) - b(\Upsilon^{\nu'}_s, {\nu'}^\vartheta_s, {\nu'}^\Upsilon_s)\mathrm{d}s\right\|^2\right] \\
&\le t\int_0^t \mathbb{E}\left[\left\|b(\Upsilon^\nu_s, \nu^\vartheta_s, \nu^\Upsilon_s) - b(\Upsilon^{\nu'}_s, {\nu'}^\vartheta_s, {\nu'}^\Upsilon_s)\right\|^2\right]\mathrm{d}s \\
&\le tC\int_0^t \mathbb{E}\left[\left\|\Upsilon^\nu_s - \Upsilon^{\nu'}_s\right\|^2\right] + \|\vartheta^\nu_s - \vartheta^{\nu'}_s\|^2 + \|\nu^\vartheta_s - {\nu'}^\vartheta_s\|^2 + \mathsf{W}_1^2(\nu^\Upsilon_s, {\nu'}^\Upsilon_s)\mathrm{d}s,
\end{aligned}$$

where $C = 4K_b^2$. Applying Grönwall's inequality, we obtain that

$$\|\vartheta^\nu_t - \vartheta^{\nu'}_t\|^2 + \mathbb{E}[\|\Upsilon^\nu_t - \Upsilon^{\nu'}_t\|^2] \le C(t)\int_0^t \left[\mathsf{W}_1^2(\nu^\Upsilon_s, {\nu'}^\Upsilon_s) + \|\nu^\vartheta_s - {\nu'}^\vartheta_s\|^2\right]\mathrm{d}s,$$

where $C(t) = Ct\exp(Ct^2)$. Then, using the fact that the LHS is an upper bound for the squared distance d and $\mathsf{W}_1 \leq \mathsf{W}_2$ (Villani, 2009, Remark 6.6), we have

$$\begin{aligned}\mathsf{d}^2(F_T(\nu)_t, F_T(\nu')_t) &\leq C(t)\int_0^t \left[\mathsf{W}_2^2(\nu_s^{\Upsilon}, {\nu'}_s^{\Upsilon}) + \|\nu_s^{\vartheta} - {\nu'}_s^{\vartheta}\|^2\right] \mathrm{d}s \\ &\leq C(T)\int_0^t \mathsf{d}^2(\nu_s, \nu'_s)\mathrm{d}s \leq tC(T) \sup_{s\in[0,T]} \mathsf{d}^2(\nu_s, \nu'_s), \end{aligned} \tag{70}$$

where we use the fact that $C(t) \leq C(T)$ since $t \in [0, T]$. We show that for $k \geq 1$ successive compositions of the map F_T denoted by F_T^k, we have the following inequality:

$$\mathsf{d}^2(F_T^k(\nu)_t, F_T^k(\nu')_t) \leq \frac{(tC(T))^k}{k!} \sup_{u\in[0,T]} \mathsf{d}^2(\nu_u, \nu'_u). \tag{71}$$

This can be proved inductively. The base case $k = 1$ follows immediately from (70), assume the inequality holds for $k - 1$, then for k we have

$$\begin{aligned}\mathsf{d}^2(F_T^k(\nu)_t, F_T^k(\nu')_t) &\leq C(T)\int_0^t \mathsf{d}^2(F_T^{k-1}(\nu)_s, F_T^{k-1}(\nu')_s)\mathrm{d}s \\ &\leq \frac{C(T)^k}{(k-1)!} \sup_{u\in[0,T]} \mathsf{d}^2(\nu_u, \nu'_u)\int_0^t t^{k-1}\mathrm{d}s \\ &\leq \frac{(tC(T))^k}{k!} \sup_{u\in[0,T]} \mathsf{d}^2(\nu_u, \nu'_u).\end{aligned}$$

Hence, we have shown that the inequality (71) holds. Taking the supremum, we obtain

$$\sup_{s\in[0,T]} \mathsf{d}^2(F_T^k(\nu)_s, F_T^k(\nu')_s) \leq \frac{(TC(T))^k}{k!} \sup_{s\in[0,T]} \mathsf{d}^2(\nu_s, \nu'_s).$$

Since $k \mapsto (TC(T))^k \in o(k!)$, there exists a large enough k such that there is a constant $0 < \alpha < 1$ for the following inequality holding:

$$\sup_{s\in[0,T]} \mathsf{d}^2(F_T^k(\nu)_s, F_T^k(\nu')_s) \leq \alpha \sup_{s\in[0,T]} \mathsf{d}^2(\nu_s, \nu'_s).$$

Hence, we have shown that the operator F_T^k is a contraction and from the Banach Fixed Point theorem and completeness of the space $(C([0,T], \mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})), \sup \mathsf{d})$, we have existence and uniqueness. $\square$

H. Space Discretization

In this section, we establish asymptotic pointwise propagation of chaos results. We are interested in justifying the use of a particle approximation in the flow (15). The flow (15) can be rewritten equivalently as follows. Let $(\Upsilon_0^i, W^i)_{i\in[M]}$ be i.i.d. copies of (Υ_0, W). We write $(\vartheta, \{\Upsilon^i\}_{i=1}^M)$ as solutions of (15) starting at $(\vartheta_0, \{\Upsilon_0^i\}_{i=1}^M)$ driven by the $\{W^i\}_{i=1}^M$. In other words, (15) satisfies

$$\begin{gathered}\mathrm{d}\vartheta_t = b_\vartheta\left(\vartheta_t, \mathrm{Law}(\Upsilon_t^1)\right)\mathrm{d}t, \\ \forall i \in [M] : \mathrm{d}\Upsilon_t^i = b_\Upsilon(\Upsilon_t^i, \vartheta_t)\mathrm{d}t + \boldsymbol{\beta}_\Upsilon\, \mathrm{d}W_t^i.\end{gathered}$$

where

$$b_\vartheta(\vartheta_t, q) = \begin{bmatrix} \eta_\theta m \\ -\gamma_\theta\eta_\theta m - \nabla_\theta \mathcal{F}(\theta', q)\end{bmatrix}, \quad b_\Upsilon(\Upsilon, \vartheta) = \begin{bmatrix} \eta_x u \\ -\gamma_x\eta_x u + \nabla_x \ell(\theta, x)\end{bmatrix}, \quad \boldsymbol{\beta}_\Upsilon := \sqrt{2}\mathrm{Diag}([0_{d_x}, 1_{d_x}]).$$

Clearly, for all $i, j \in [M]$, we have $\mathrm{Law}(\Upsilon_t^i) = \mathrm{Law}(\Upsilon_t^j)$.

We will justify that we can replace the $\text{Law}(\Upsilon_t^1)$ with a particle approximation to obtain the approximate process (20), or equivalently as:

$$\mathrm{d}\vartheta_t^M = b_\vartheta\left(\vartheta_t^M, \frac{1}{M}\sum_{i=1}^M \delta_{\Upsilon_t^{i,M}}\right)\mathrm{d}t \tag{73a}$$

$$\forall i \in [M] : \mathrm{d}\Upsilon_t^{i,M} = b_\Upsilon(\Upsilon_t^{i,M}, \vartheta_t^M)\mathrm{d}t + \beta_\Upsilon \mathrm{d}W_t^i. \tag{73b}$$

Similarly to Proposition G.2, we can show that b_ϑ and b_Υ are both Lipschitz. We are now ready to prove Proposition 5.1 justifying that (20) is a good approximation to (15).

Proof of Proposition 5.1. This is equivalent to showing that

$$\lim_{M\to\infty} \mathbb{E}\left[\sup_{t\in[0,T]}\left\{\|\vartheta_t - \vartheta_t^M\|^2 + \frac{1}{M}\sum_{i=1}^M \|\Upsilon_t^i - \Upsilon_t^{i,M}\|^2\right\}\right] = 0.$$

More specifically, we will show that

$$\underbrace{\mathbb{E}\left[\sup_{t\in[0,T]}\|\vartheta_t - \vartheta_t^M\|^2\right]}_{(a)} + \underbrace{\mathbb{E}\left[\sup_{t\in[0,T]}\frac{1}{M}\sum_{i=1}^M \|\Upsilon_t^i - \Upsilon_t^{i,M}\|^2\right]}_{(b)} = o(1).$$

We begin with (a). From Jensen's inequality, we have

$$\begin{aligned}(a) &\le T\mathbb{E}\int_0^T \left\|b_\vartheta(\vartheta_s, \text{Law}(\Upsilon_s^1)) - b_\vartheta(\vartheta_s^M, \frac{1}{M}\sum_{i=1}^M \delta_{\Upsilon_s^{i,M}})\right\|^2 \mathrm{d}s \\ &\le C_\vartheta \int_0^T \mathbb{E}\|\vartheta_s - \vartheta_s^M\|^2 + \mathbb{E}\mathrm{W}_2^2\left(\text{Law}(\Upsilon_s^0), \frac{1}{M}\sum_{i=1}^M \delta_{\Upsilon_s^{i,M}}\right)\mathrm{d}s,\end{aligned}$$

where we use the fact that b_ϑ is K_{b_ϑ}–Lipschitz, $C_\vartheta = 2TK_{b_\vartheta}^2$, and $\mathbb{E}(a+b)^2 \le 2(\mathbb{E}a^2 + \mathbb{E}b^2)$ (known as the C_r–inequality (Loève, 1977, p157)).

Note that by the triangle inequality

$$\begin{aligned}\mathbb{E}\mathrm{W}_2^2\left(\text{Law}(\Upsilon_s^0), \frac{1}{M}\sum_{i=1}^M \delta_{\Upsilon_s^{i,M}}\right) \le\ & 2\mathbb{E}\mathrm{W}_2^2\left(\text{Law}(\Upsilon_s^0), \frac{1}{M}\sum_{i=1}^M \delta_{\Upsilon_s^i}\right) \\ &+ 2\mathbb{E}\mathrm{W}_2^2\left(\frac{1}{M}\sum_{i=1}^M \delta_{\Upsilon_s^i}, \frac{1}{M}\sum_{i=1}^M \delta_{\Upsilon_s^{i,M}}\right).\end{aligned}$$

The two terms can be bounded using Carmona (2016, Eq. (1.24) and Lemma 1.9) to produce the inequality

$$\mathbb{E}\mathrm{W}_2^2\left(\text{Law}(\Upsilon_s^0), \frac{1}{M}\sum_{i=1}^M \delta_{\Upsilon_s^{i,M}}\right) \le o(1) + \frac{2}{M}\sum_{i=1}^M \mathbb{E}\|\Upsilon_s^i - \Upsilon_s^{i,M}\|^2.$$

Hence, we have that

$$(a) \le C' \int_0^T \left\{\mathbb{E}\|\vartheta_s - \vartheta_s^M\|^2 + o(1) + \frac{1}{M}\sum_{i=1}^M \mathbb{E}\|\Upsilon_s^i - \Upsilon_s^{i,M}\|^2\right\}\mathrm{d}s.$$

where $C'_\vartheta = 2C_\vartheta$. We use the fact that

$$\|\vartheta_s - \vartheta_s^M\|^2 \le \sup_{s'\in[0,s]}\|\vartheta_{s'} - \vartheta_{s'}^M\|^2, \quad \frac{1}{M}\sum_{i=1}^M \|\Upsilon_s^i - \Upsilon_s^{i,M}\|^2 \le \sup_{s'\in[0,s]}\frac{1}{M}\sum_{i=1}^M \|\Upsilon_{s'}^i - \Upsilon_{s'}^{i,M}\|^2, \tag{74}$$

to obtain

$$(a) \leq C' \int_0^T \left\{ \mathbb{E} \sup_{s' \in [0,s]} \|\vartheta_{s'} - \vartheta_{s'}^M\|^2 + o(1) + \mathbb{E} \sup_{s' \in [0,s]} \frac{1}{M} \sum_{i=1}^M \|\Upsilon_{s'}^i - \Upsilon_{s'}^{i,M}\|^2 \right\} \mathrm{d}s \tag{75}$$

Similarly, for (b), we have

$$\begin{aligned}
(b) &\leq \frac{T}{M} \sum_{i=1}^M \mathbb{E} \int_0^T \|b_\Upsilon(\Upsilon_s^i, \vartheta_s) - b_\Upsilon(\Upsilon_s^{i,M}, \vartheta_s^M)\|^2 \mathrm{d}s \\
&\leq C_\Upsilon \mathbb{E} \int_0^T \|\vartheta_s - \vartheta_s^M\|^2 + \frac{1}{M} \sum_{i=1}^M \|\Upsilon_s^i - \Upsilon_s^{i,M}\|^2 \mathrm{d}s \\
&\leq C_\Upsilon \int_0^T \mathbb{E} \sup_{s' \in [0,s]} \|\vartheta_{s'} - \vartheta_{s'}^M\|^2 + \mathbb{E} \sup_{s' \in [0,s]} \frac{1}{M} \sum_{i=1}^M \|\Upsilon_{s'}^i - \Upsilon_{s'}^{i,M}\|^2, \mathrm{d}s,
\end{aligned}$$

where $C_\Upsilon := 2TK_{b_\Upsilon}^2$ for the last line we use the trick in (74).

Combining the bounds for (a) and (b), we obtain

$$(a) + (b) \leq C \int_0^T \left\{ \mathbb{E} \sup_{s' \in [0,s]} \|\vartheta_{s'} - \vartheta_{s'}^M\|^2 + o(1) + \mathbb{E} \sup_{s' \in [0,s]} \frac{1}{M} \sum_{i=1}^M \|\Upsilon_{s'}^i - \Upsilon_{s'}^{i,M}\|^2 \right\} \mathrm{d}s,$$

where $C = C_\Upsilon + C'_\vartheta$. Applying Gronwall's inequality, we obtain

$$\mathbb{E} \left[\sup_{t \in [0,T]} \|\vartheta_t - \vartheta_t^M\| \right] + \mathbb{E} \left[\sup_{t \in [0,T]} \frac{1}{M} \sum_{i=1}^M \|\Upsilon_t^i - \Upsilon_t^{i,M}\| \right] \leq o(1).$$

Taking the limit, we have as desired. □

I. Time Discretization

In this section, we are concerned with discretization schemes of various ODE/SDEs. The structure is as follows:

- Appendix I.1. We describe the Euler-Marayama discretization of PGD as described in Kuntz et al. (2023).
- Appendix I.2. We show we can obtain NAG as a discretization of the damped Hamiltonian.
- Appendix I.3. We show a discretization of MPD (described in Equation (15)) using a scheme replicating NAG as described in Appendix I.2 for the (θ, m)-components, and Cheng et al. (2018)'s for (x, u)-components.
- Appendix I.4. We derive the transition using a scheme inspired by Cheng et al. (2018) while incorporating NAG-style gradient correction as described in Sutskever et al. (2013).

I.1. PGD discretization

In order to obtain an implementable system, it is standard to then discretise the distribution q_t by representing it with a finite particle system, i.e. $q_t\left(\mathrm{d}x\right) \approx \frac{1}{M} \sum_{i \in [M]} \delta\left(X_t^i, \mathrm{d}x\right)$. Upon making this approximation, one obtains the system

$$\begin{aligned}
\dot{\theta}_t &= \frac{1}{M} \sum_{i \in [M]} \nabla_\theta \ell\left(\theta_t, X_t^i\right), \\
\text{for } i \in [M], \quad \mathrm{d}X_t^i &= \nabla_x \ell\left(\theta_t, X_t^i\right) \mathrm{d}t + \sqrt{2}\,\mathrm{d}W_t^i,
\end{aligned}$$

in which all terms are readily available. Discretising this process in time then yields the Particle Gradient Descent (PGD) algorithm of (Kuntz et al., 2023), i.e. for $k \geq 1$, iterate

$$\theta_k = \theta_{k-1} + h \left(\frac{1}{M} \sum_{i=1}^M \nabla_\theta \ell\left(\theta_{k-1}, X_{k-1}^i\right) \right),$$

$$\text{for } i \in [M], \quad X_k^i = X_{k-1}^i + h\nabla_x \ell\left(\theta_{k-1}, X_{k-1}^i\right) + \sqrt{2h}\epsilon_k^i,$$

where $\epsilon_k^i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0_{d_x}, I_{d_x})$, for some initialization $(\theta_0, \{X_0^i\}_{i=1}^M)$.

I.2. NAG as a discretization

Recall the momentum-enriched ODE is given by

$$\ddot{\theta}_t + \gamma\eta\dot{\theta}_t + \eta\nabla_\theta f(\theta_t) = 0.$$

Let $m_t = \frac{1}{\eta}\dot{\theta}_t$, then we can write the above equivalently as the following coupled first-order ODEs:

$$\dot{m}_t = -\gamma\eta m_t - \nabla f(\theta_t), \quad \dot{\theta}_t = \eta m_t.$$

When $\gamma = \frac{3}{t}$ and $\eta = 1$, we show that a particular discretization of the above is equivalent to Nesterov Accelerated Gradient (NAG) method (Nesterov, 1983). The argument is inspired by reversing the one of Su et al. (2014) who obtained the continuous limit of NAG.

Recall that NAG (Nesterov, 1983), for convex f (but not strongly convex), is defined by the following iteration:

$$\theta_k = y_{k-1} - h\nabla f\left(y_{k-1}\right), \quad y_k = \theta_k + \left(\frac{k-1}{k+2}\right)\left(\theta_k - \theta_{k-1}\right).$$

Since $\frac{k-1}{k+2} \approx (1 - \frac{3}{k})$, we have

$$\theta_k \approx y_{k-1} - h\nabla_\theta f\left(y_{k-1}\right), \quad y_k \approx \theta_k + \left(1 - \frac{3}{k}\right)\left(\theta_k - \theta_{k-1}\right). \tag{76}$$

We will now show how a particular discretization scheme will produce (76). It uses a combination of implicit Euler for $\dot{\theta}_t$, and use a semi-implicit Euler scheme for $\dot{m}_t$. More specifically, the semi-implicit scheme for $\dot{m}_t = -\frac{3}{t}m_t - \nabla_\theta f(\theta_t)$ uses an explicit approximation for the momentum $\frac{3}{t}m_t$, and implicit approximation for the gradient $\nabla_\theta f(\theta_t)$. In summary, we obtain the following discretization

$$m_t \approx m_{t-\sqrt{h}} - \sqrt{h}\left(\frac{3}{t-\sqrt{h}}m_{t-\sqrt{h}} + \nabla_\theta f(\theta_t)\right), \quad \theta_t \approx \theta_{t-\sqrt{h}} + \sqrt{h}m_t.$$

We can write the above equivalently through the map $t \mapsto \frac{t}{\sqrt{h}}$ and in terms of $k := \frac{t}{\sqrt{h}}$, to obtain

$$\begin{aligned} m_k &\approx \left(1 - \frac{3}{k-1}\right)m_{k-1} - \sqrt{h}\nabla_\theta f(\theta_k), \\ \theta_k &\approx \theta_{k-1} + \sqrt{h}m_k. \end{aligned}$$

Hence,

$$\theta_k \approx \theta_{k-1} + \sqrt{h}\left(1 - \frac{3}{k-1}\right)m_{k-1} - h\nabla_\theta f(\theta_k).$$

From our discretization, we have that $\sqrt{h}m_k \approx \theta_k - \theta_{k-1}$ then we obtain

$$\theta_k \approx \theta_{k-1} + \left(1 - \frac{3}{k-1}\right)(\theta_{k-1} - \theta_{k-2}) - h\nabla_\theta f(\theta_k).$$

We can write the θ update in terms of y_k (cf. (76)),

$$\theta_k \approx y_{k-1} - h\nabla_\theta f(\theta_k).$$

If f is Lipschitz, then we have

$$\|h\nabla_\theta f(\theta_k) - h\nabla_\theta f(y_{k-1})\| \le Lh\|\theta_k - y_{k-1}\| \le h^2 L\|\nabla_\theta f(\theta_k)\| \le o(h).$$

We replace the gradient $\nabla_\theta f(\theta_k)$ with $\nabla_\theta f(y_{k-1})$ to obtain as desired

$$\theta_k \approx y_{k-1} - h\nabla_\theta f(y_{k-1}), \quad y_k \approx \theta_k + \left(1 - \frac{3}{k}\right)(\theta_k - \theta_{k-1}).$$

Interestingly, for given step size h, the (approximate) NAG iterations approximate the flow for time $\sqrt{h}$ as opposed to h (Su et al., 2014, Section 3.4)

I.3. Nesterov and Cheng's discretization of MPD

In this section, we show how to discretize the MPD using a combination of Nesterov's (see Appendix I.2) and Cheng et al. (2018)'s discretization. Recall, we have

$$\begin{aligned}
\mathrm{d}\theta_t &= \eta_\theta m_t \,\mathrm{d}t, \\
\mathrm{d}m_t &= -\nabla_\theta \mathcal{E}(\theta_t, q^M_{t,X})\, dt - \gamma_\theta\eta_\theta m_t \,\mathrm{d}t, \\
\forall i \in [M] : \mathrm{d}X^i_t &= \eta_x U^i_t \mathrm{d}t, \\
\forall i \in [M] : \mathrm{d}U^i_t &= \nabla_x \ell(\theta_t, X^i_t)\mathrm{d}t - \gamma_x\eta_x U^i_t \mathrm{d}t + \sqrt{2\gamma_x}\,\mathrm{d}W^i_t.
\end{aligned}$$

In this section, we show how discretizing the θ component in the style of Nesterov (specifically, Sutskever et al. (2013)'s formulation) and q component in the style of Cheng et al. (2018) obtains the MPD-NC (Nesterov–Cheng) algorithm. The MPD-NC algorithm is described as follows: given previous values $(\theta_k, v_k, \{X^i_k, U^i_k\}^M_{i=1})$ and step-size $h > 0$, we iterate

$$\theta_{k+1} = \theta_k + v_k, \tag{77a}$$
$$v_{k+1} = \mu v_k - h^2 \nabla_\theta \mathcal{E}(\theta_k + \mu v_k, q^M_{k,X}), \tag{77b}$$
$$\forall i \in [M] : X^i_{k+1} = X^i_k + \frac{1}{\gamma_x}\left[(1-\omega_x(h))U^i_k + \nabla_x \ell(\theta_{k+1}, X^i_k)\left(h - \frac{1-\omega_x(h)}{\gamma_x\eta_x}\right)\right] + L^{XX}_\Sigma \xi^i_k, \tag{77c}$$
$$\forall i \in [M] : U^i_{k+1} = \omega_x(h)U^i_k + \frac{1-\omega_x(h)}{\gamma_x\eta_x}\nabla_x \ell(\theta_{k+1}, X^i_k) + L^{XU}_\Sigma \xi^i_k + L^{UU}_\Sigma \xi'^i_k, \tag{77d}$$

for all $i \in [N]$, where $\mu := 1 - h\gamma_\theta\eta_\theta$, $\omega_x(t) :=$; $L^{XX}_\Sigma, L^{XU}_\Sigma, L^{UU}_\Sigma$ is described in (83), and $\{\xi^i_k, \xi'^i_k\} \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{d_x})$. Each iteration corresponds to (approximately) solving (15) for time h.

In Appendix I.3.1, we show how Nesterov's discretization can be used to produce the update

$$\begin{aligned}
\theta_{k+1} &= \theta_k + v_k, \\
v_{k+1} &= \mu v_k - h^2 \nabla_\theta \mathcal{E}(\theta_k + \mu v_k, q^M_{k,X}).
\end{aligned}$$

In Appendix I.3.2, given $(\theta_{k+1}, \{X^i_k, U^i_k\}^M_{i=1})$, we show that the transition is described by

$$\begin{aligned}
\forall i \in [M] : X^i_{k+1} &= X^i_k + \frac{1}{\gamma_x}\left[(1-\omega_x(h))U^i_k + \nabla_x \ell(\theta_{k+1}, X^i_k)\left(h - \frac{1-\omega_x(h)}{\gamma_x\eta_x}\right)\right] + L^{XX}_\Sigma \xi^i_k, \\
\forall i \in [M] : U^i_{k+1} &= \omega_x(h)U^i_k + \frac{1-\omega_x(h)}{\gamma_x\eta_x}\nabla_x \ell(\theta_{k+1}, X^i_k) + L^{XU}_\Sigma \xi^i_k + L^{UU}_\Sigma \xi'^i_k.
\end{aligned}$$

I.3.1. Discretization of $(\dot\theta_t, \dot m_t)$

This discretization follows similarly to that in Appendix I.2. Similarly, let $k := \frac{t}{\sqrt{h}}$ with the time rescaling $t \mapsto \frac{t}{\sqrt{h}}$, then we consider the following discretization:

$$\begin{aligned}
m_k &= m_{k-1} - \sqrt{h}\left(\gamma_\theta\eta_\theta m_{k-1} + \nabla_\theta \mathcal{E}(\theta_k, q^M_{k-1,X})\right) \\
&= \left(1 - \sqrt{h}\gamma_\theta\eta_\theta\right) m_{k-1} - \sqrt{h}\nabla_\theta \mathcal{E}(\theta_k, q^M_{k-1,X}), \\
\theta_k &= \theta_{k-1} + \sqrt{h}\eta_\theta m_k,
\end{aligned}$$

Expanding m_k, we obtain

$$\begin{aligned}\theta_k &= \theta_{k-1} + \sqrt{h}\eta_\theta\left(1-\sqrt{h}\gamma_\theta\eta_\theta\right)m_{k-1} - h\nabla_\theta\mathcal{E}(\theta_k, q^M_{k-1,X})\\ &= \theta_{k-1} + \left(1-\sqrt{h}\gamma_\theta\eta_\theta\right)(\theta_{k-1}-\theta_{k-2}) - h\nabla_\theta\mathcal{E}(\theta_k, q^M_{k-1,X}),\end{aligned}$$

Let $v_t = \theta_k - \theta_{k-1}$, then we have

$$\begin{aligned}\theta_k &= \theta_{k-1} + v_k\\ v_k &= \bar{\mu}v_{k-1} - h\nabla_\theta\mathcal{E}(\theta_k, q^M_{k-1,X}),\end{aligned}$$

where $\bar{\mu} := 1-\sqrt{h}\gamma_\theta\eta_\theta$. Then, as before, using the approximation $h\nabla\mathcal{E}(\theta_k, q^M_{k-1,X}) \approx h\nabla\mathcal{E}(\theta_{k-1}+\bar{\mu}v_{k-1}, q^M_{k-1,X})+o(h)$, we obtain that

$$\begin{aligned}\theta_k &= \theta_{k-1} + v_k,\\ v_k &= \bar{\mu}v_{k-1} - h\nabla_\theta\mathcal{E}(\theta_{k-1}+\bar{\mu}v_{k-1}, q^M_{k-1,X}).\end{aligned}$$

Similarly to Appendix I.2 this approximates the flow for time $\sqrt{h}$ instead of h. Hence, we have to apply the appropriate rescaling to obtain the following iteration with the desired behaviour:

$$\begin{aligned}\theta_k &= \theta_{k-1} + v_k,\\ v_k &= \mu v_{k-1} - h^2\nabla_\theta\mathcal{E}(\theta_{k-1}+\mu v_{k-1}, q^M_{k-1,X}),\end{aligned}$$

where $\mu := 1 - h\gamma_\theta\eta_\theta$. This is exactly that of Sutskever et al. (2013)'s characterization of NAG (see Equations 3 and 4 in their paper).

I.3.2. Discretization of $(\dot{X}_t, \dot{U}_t)$

We describe the discretization scheme of Cheng et al. (2018). For simplicity, we derive the transition of a single particle (X_t, U_t) since there are no interactions between the particles given θ. Furthermore, for brevity and without loss in generality, we derive the transition for some step size $h > 0$ given initial values (θ, X_0, U_0), which can be easily generalized to future transitions.

Consider a time interval $t \in (0, h]$ and given (X_0, U_0), we first approximate the gradient $\nabla_x\ell(\theta, X_t)$ with $\nabla_x\ell(\theta, X_0)$ to arrive at the following linear SDE:

$$\mathrm{d}\begin{pmatrix}X_t\\ U_t\end{pmatrix} = \left[\begin{pmatrix}0\\ \nabla_x\ell(\theta, X_0)\end{pmatrix} + \begin{pmatrix}0 & \eta_x I\\ 0 & -\gamma_x\eta_x I\end{pmatrix}\begin{pmatrix}X_t\\ U_t\end{pmatrix}\right]\mathrm{d}t + \sqrt{2\gamma_x}\begin{pmatrix}0\\ 1\end{pmatrix}\mathrm{d}W_t. \tag{78}$$

A $2d_x$-dimensional linear SDE is given by:

$$\mathrm{d}\boldsymbol{X}_t = (\boldsymbol{A}\boldsymbol{X}_t + \boldsymbol{\alpha})\,\mathrm{d}t + \boldsymbol{\beta}\,\mathrm{d}W_t, \tag{79}$$

where $\boldsymbol{A} \in \mathbb{R}^{2d_x\times 2d_x}$ and $\boldsymbol{\alpha}, \boldsymbol{\beta} \in \mathbb{R}^{2d_x}$ are fixed matrices. It is clear that if we set

$$\boldsymbol{X}_t = \begin{pmatrix}X_t\\ U_t\end{pmatrix}, \quad \boldsymbol{A} = \begin{pmatrix}0_{d_x} & \eta_x I_{d_x}\\ 0_{d_x} & -\gamma_x\eta_x I_{d_x}\end{pmatrix}, \quad \boldsymbol{\alpha} = \begin{pmatrix}0_{d_x\times 1}\\ \nabla_x\ell(\theta, X_0)\end{pmatrix}, \quad \boldsymbol{\beta} = \sqrt{2\gamma_x}\begin{pmatrix}0_{d_x\times 1}\\ 1_{d_x\times 1}\end{pmatrix},$$

then the discretized underdamped Langevin SDE of (78) falls within the class of linear SDE of the form specified in (79). These SDEs admits the following explicit solution (Platen and Bruti-Liberati, 2010, see pages 48, 101),

$$\boldsymbol{X}_t = \boldsymbol{\Psi}_t\left(\boldsymbol{X_0} + \int_0^t \boldsymbol{\Psi}_s^{-1}\boldsymbol{\alpha}\,\mathrm{d}s + \int_0^t \boldsymbol{\Psi}_s^{-1}\boldsymbol{\beta}\,\mathrm{d}W_s\right), \tag{80}$$

where

$$\boldsymbol{\Psi}_t := \exp\left(\boldsymbol{A}t\right),$$

with exp to be understood as the matrix exponential. In our case, the matrix exponential and its inverse is given by

$$\boldsymbol{\Psi}_t = I_{2d_x} + \sum_{i=1}^{\infty} \frac{1}{k!} \begin{bmatrix} 0_{d_x} & \frac{(-\gamma_x \eta_x t)^k}{-\gamma_x} I_{d_x} \\ 0_{d_x} & (-\gamma_x \eta_x t)^k I_{d_x} \end{bmatrix} = \begin{pmatrix} I_{d_x} & \frac{1-\omega_x(t)}{\gamma_x} I_{d_x} \\ 0_{d_x} & \omega_x(t) I_{d_x} \end{pmatrix}, \tag{81a}$$

$$\boldsymbol{\Psi}_t^{-1} = \begin{pmatrix} I_{d_x} & \frac{1-\omega_x(-t)}{\gamma_x} I_{d_x} \\ 0_{d_x} & \omega_x(-t) I_{d_x} \end{pmatrix}, \tag{81b}$$

where $\omega_x(t) := \exp(-\gamma_x \eta_x t)$. It can be verified that they are indeed the inverse of each other, i.e., $\boldsymbol{\Psi}_t \boldsymbol{\Psi}_t^{-1} = I_{2d_x}$. Plugging in (81) into (80), we obtain the following:

$$\begin{pmatrix} X_t \\ U_t \end{pmatrix} = \begin{pmatrix} I_{d_x} & \frac{1-\omega_x(t)}{\gamma_x} I_{d_x} \\ 0_{d_x} & \omega_x(t) I_{d_x} \end{pmatrix} \left(\begin{bmatrix} X_0 \\ U_0 \end{bmatrix} + \int_0^t \begin{pmatrix} \frac{1-\omega_x(-s)}{\gamma_x} \nabla_x \ell(\theta, X_0) \\ \omega_x(-s) \nabla_x \ell(\theta, X_0) \end{pmatrix} \mathrm{d}s + \sqrt{2} \int_0^t \begin{pmatrix} \frac{1-\omega_x(-s)}{\sqrt{\gamma_x}} 1_{d_x \times 1} \\ \sqrt{\gamma_x} \omega_x(-s) 1_{d_x \times 1} \end{pmatrix} \mathrm{d}W_s \right).$$

Or, equivalently, we have

$$\begin{aligned} U_t &= \omega_x(t) U_0 + \nabla_x \ell(\theta, X_0) \int_0^t \omega_x(t-s)\, \mathrm{d}s + \sqrt{2\gamma_x} \int_0^t \omega_x(t-s)\, \mathrm{d}W_s \\ &= \omega_x(t) U_0 + \frac{1-\omega_x(t)}{\gamma_x \eta_x} \nabla_x \ell(\theta, X_0) + \sqrt{2\gamma_x} \int_0^t \omega_x(t-s)\, \mathrm{d}W_s, \\ X_t &= X_0 + \frac{1}{\gamma_x} \left[(1-\omega_x(t)) U_0 + \nabla_x \ell(\theta, X_0) \left[\int_0^t 1 - \omega_x(t-s)\, \mathrm{d}s \right] + \sqrt{\frac{2}{\gamma_x}} \int_0^t 1 - \omega_x(t-s)\, \mathrm{d}W_s \right] \\ &= X_0 + \frac{1}{\gamma_x} \left[(1-\omega_x(t)) U_0 + \nabla_x \ell(\theta, X_0) \left[t - \frac{1-\omega_x(t)}{\gamma_x \eta_x} \right] + \sqrt{\frac{2}{\gamma_x}} \int_0^t 1 - \omega_x(t-s)\, \mathrm{d}W_s \right]. \end{aligned}$$

As noted by Cheng et al. (2018), this is a Gaussian transition. To characterize it, we need to calculate the first two moments: For the first moments, we have

$$\begin{aligned} \mu_U(X_0, U_0, t) :=& \mathbb{E}[U_t] \\ =& \omega_x(t) U_0 + \frac{1-\omega_x(t)}{\gamma_x \eta_x} \nabla_x \ell(\theta, X_0), \\ \mu_X(X_0, U_0, t) :=& \mathbb{E}[X_t] \\ =& X_0 + \frac{1}{\gamma_x} \left[(1-\omega_x(t)) U_0 + \nabla_x \ell(\theta, X_0) \left[t - \frac{1-\omega_x(t)}{\gamma_x \eta_x} \right] \right]. \end{aligned}$$

For the second moments, we have

$$\begin{aligned} \Sigma_{UU}(t) &:= \mathbb{E}[(U_t - \mathbb{E}[U_t])(U_t - \mathbb{E}[U_t])^\top] \\ &= 2\gamma_x \mathbb{E}\left[\left(\int_0^t \omega_x(t-s)\, \mathrm{d}W_s \right) \left(\int_0^t \omega_x(t-s)\, \mathrm{d}W_s \right)^\top \right] \\ &= 2\gamma_x \left(\int_0^t \omega_x(2(t-s))\, \mathrm{d}s \right) I_{d_x} \\ &= \left(\frac{1-\omega_x(2t)}{\eta_x} \right) I_{d_x}, \\ \Sigma_{XX}(t) &:= \mathbb{E}[(X_t - \mathbb{E}[X_t])(X_t - \mathbb{E}[X_t])^\top] \\ &= \frac{2}{\gamma_x} \mathbb{E}\left[\left(\int_0^t 1 - \omega_x(t-s)\, \mathrm{d}W_s \right) \left(\int_0^t 1 - \omega_x(t-s)\, \mathrm{d}W_s \right)^\top \right] \end{aligned}$$

$$
\begin{aligned}
&= \frac{2}{\gamma_x}\left(\int_0^t [1-\omega_x(t-s)]^2 \, \mathrm{d}s\right) I_{d_x} \\
&= \frac{1}{\gamma_x}\left[2t - \frac{\omega_x(2t)}{\gamma_x\eta_x} + \frac{4\omega_x(t)}{\gamma_x\eta_x} - \frac{3}{\eta_x\gamma_x}\right] I_{d_x}, \\
\Sigma_{UX}(t) &:= \mathbb{E}[(U_t - \mathbb{E}[U_t])(X_t - \mathbb{E}[X_t])^\top] \\
&= 2\mathbb{E}\left[\left(\int_0^t \omega_x(t-s)\,\mathrm{d}W_s\right)\left(\int_0^t 1-\omega_x(t-s)\,\mathrm{d}W_s\right)^\top\right] \\
&= 2\left(\int_0^t \omega_x(t-s)(1-\omega_x(t-s))\,\mathrm{d}s\right) I_{d_x} \\
&= \frac{1}{\gamma_x\eta_x}\left(1 - 2\omega_x(t) + \omega_x(2t)\right) I_{d_x}.
\end{aligned}
$$

Hence, the transition can be described as $\begin{pmatrix} X_t \\ U_t \end{pmatrix} \sim \mathcal{N}\left(\mu(t), \Sigma(t)\right)$ where

$$
\mu(t) = \begin{pmatrix} \mu_X(X_k, U_k, t) \\ \mu_U(X_k, U_k, t) \end{pmatrix}, \quad \Sigma(t) = \begin{pmatrix} \Sigma_{XX}(t) & \Sigma_{UX}(t) \\ \Sigma_{UX}(t) & \Sigma_{UU}(t) \end{pmatrix}.
$$

Therefore, given some point (X_k, U_k) and some step-size $h > 0$, we have that

$$
\begin{pmatrix} X_{k+1} \\ U_{k+1} \end{pmatrix} \sim \mathcal{N}\left(\mu(h), \Sigma(h)\right). \tag{82}
$$

Sampling from the transition. Using samples from a standard Gaussian $z \sim \mathcal{N}(0_d, I_d)$, one may produce samples from of a multivariate distribution $\mathcal{N}(\mu, \Sigma)$ by using the fact that

$$
X = \mu + Lz \sim \mathcal{N}(\mu, \Sigma),
$$

where L is a lower triangular matrix with positive diagonal entries obtained via the Cholesky decomposition of Σ, i.e, $\Sigma = LL^\top$.

The Cholesky decomposition of the covariance matrix of (82) will be described here. Consider the LDL decomposition of a symmetric matrix

$$
\Sigma = \begin{bmatrix} A & B \\ B^\top & C \end{bmatrix} = \begin{bmatrix} I & 0 \\ B^\top A^{-1} & I \end{bmatrix} \begin{bmatrix} A & 0 \\ 0 & S \end{bmatrix} \begin{bmatrix} I & A^{-1}B \\ 0 & I \end{bmatrix},
$$

where $S = D - B^\top A^{-1} B$ is the Schur complement. Given Cholesky factorization of A and B, written as $A = L_A L_A^\top$ and $S = L_S L_S^\top$ respectively, we can write the Cholesky decomposition of Σ as

$$
\Sigma = \begin{bmatrix} L_A & 0 \\ B^\top L_A^{-\top} & L_S \end{bmatrix} \begin{bmatrix} L_A^\top & L_A^{-1}B \\ 0 & L_S^\top \end{bmatrix} = L_\Sigma L_\Sigma^\top.
$$

In our case, we compute the Cholesky decomposition of Σ as follows:

$$
L_\Sigma = \begin{pmatrix} L_\Sigma^{XX} I_{d_x} & 0 \\ L_\Sigma^{XU} I_{d_x} & L_\Sigma^{UU} I_{d_x} \end{pmatrix},
$$

where the constants are defined as

$$
L_\Sigma^{XX} = \sqrt{\frac{1}{\gamma_x}\left[2h - \frac{\omega_x(2h)}{\gamma_x\eta_x} + \frac{4\omega_x(h)}{\gamma_x\eta_x} - \frac{3}{\eta_x\gamma_x}\right]}, \tag{83a}
$$

$$L_{\Sigma}^{XU} = \frac{\frac{1}{\gamma_x \eta_x}\left(1 - 2\omega_x(h) + \omega_x(2h)\right)}{\sqrt{\frac{1}{\gamma_x}\left[2h - \frac{\omega_x(2h)}{\gamma_x \eta_x} + \frac{4\omega_x(h)}{\gamma_x \eta_x} - \frac{3}{\eta_x \gamma_x}\right]}}, \tag{83b}$$

$$L_{\Sigma}^{UU} = \sqrt{\frac{1 - \omega_x(2h)}{\eta_x} - \frac{\left(\frac{1}{\gamma_x \eta_x}\left(1 - 2\omega_x(h) + \omega_x(2h)\right)\right)^2}{\frac{1}{\gamma_x}\left[2h - \frac{\omega_x(2h)}{\gamma_x \eta_x} + \frac{4\omega_x(h)}{\gamma_x \eta_x} - \frac{3}{\eta_x \gamma_x}\right]}}. \tag{83c}$$

Therefore, we the transition can be written as follows:

$$\forall i \in [M] : X_{k+1}^i = X_k^i + \frac{1}{\gamma_x}\left[(1 - \omega_x(h))U_k + \nabla_x \ell(\theta, X_k^i)\left[h - \frac{1 - \omega_x(h)}{\gamma_x \eta_x}\right]\right] + L_{\Sigma}^{XX} \xi_k^i,$$
$$\forall i \in [M] : U_{k+1}^i = \omega_x(h) U_k^i + \frac{1 - \omega_x(h)}{\gamma_x \eta_x} \nabla_x \ell(\theta, X_k^i) + L_{\Sigma}^{XU} \xi_k^i + L_{\Sigma}^{UU} \xi_k'^{,i},$$

where $\{\xi_k^i, \xi_k'^{,i}\}_{i \in [M], k} \stackrel{\text{i.i.d.}}{\sim} \mathcal{N}(0_{d_x}, I_{d_x})$.

I.4. Proposed discretization of (21)

Recall, our approximating SDE in (21) is given by:

$$\begin{aligned}
&\mathrm{d}\tilde{\theta}_t = \eta_\theta \tilde{m}_t \,\mathrm{d}t \\
&\mathrm{d}\tilde{m}_t = -\gamma_\theta \eta_\theta \tilde{m}_t \,\mathrm{d}t - \nabla_\theta \mathcal{E}\left(\bar{\theta}_0, \tilde{q}_{0,X}^M\right) \mathrm{d}t \\
\text{for } i \in [M], \quad &\mathrm{d}\tilde{X}_t^i = \eta_x \tilde{U}_t^i \,\mathrm{d}t \\
\text{for } i \in [M], \quad &\mathrm{d}\tilde{U}_t^i = -\gamma_x \eta_x \tilde{U}_t^i \,\mathrm{d}t + \nabla_x \ell\left(\bar{\bar{\theta}}_0, \tilde{X}_0^i\right) \mathrm{d}t + \sqrt{2\gamma_x}\, \mathrm{d}W_t^i.
\end{aligned}$$

For simplicity and without loss in generality, we assume there is a single particle, i.e., $M = 1$, and derive the transition for a single step. Clearly, (21) can be written as an linear $2d_x + 2d_\theta$ SDE:

$$\mathrm{d}\boldsymbol{X}_t = (\boldsymbol{A}\boldsymbol{X}_t + \boldsymbol{\alpha})\,\mathrm{d}t + \boldsymbol{\beta}\,\mathrm{d}W_t,$$

where

$$\boldsymbol{X}_t = \begin{pmatrix} \tilde{\theta}_t \\ \tilde{m}_t \\ \tilde{X}_t \\ \tilde{U}_t \end{pmatrix}, \quad \boldsymbol{A} = \begin{pmatrix} \begin{matrix} 0_{d_\theta} & \eta_\theta I_{d_\theta} \\ 0_{d_\theta} & -\eta_\theta \gamma_\theta I_{d_\theta} \end{matrix} & 0_{2d_\theta \times 2d_x} \\ 0_{2d_x \times 2d_\theta} & \begin{matrix} 0_{d_x} & \eta_x I_{d_x} \\ 0_{d_x} & -\eta_x \gamma_x I_{d_x} \end{matrix} \end{pmatrix},$$

$$\boldsymbol{\alpha} = \begin{pmatrix} 0_{d_\theta \times 1} \\ -\nabla_\theta \mathcal{E}(\bar{\theta}_0, \tilde{q}_{0,X}^M) \\ 0_{d_x \times 1} \\ \nabla_x \ell(\bar{\bar{\theta}}_0, \tilde{X}_0) \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} 0_{d_\theta \times 1} \\ 0_{d_\theta \times 1} \\ 0_{d_x \times 1} \\ 1_{d_x \times 1} \end{pmatrix}.$$

Hence it admits the following explicit solution (Platen and Bruti-Liberati, 2010, see page 48, 101):

$$\boldsymbol{X}_t = \boldsymbol{\Psi}_t \left(\boldsymbol{X_0} + \int_0^t \boldsymbol{\Psi}_s^{-1} \boldsymbol{\alpha}\,\mathrm{d}s + \int_0^t \boldsymbol{\Psi}_s^{-1} \boldsymbol{\beta}\,\mathrm{d}W_s\right),$$

where

$$\boldsymbol{\Psi}_t := \exp\left(\boldsymbol{A}t\right),$$

with $\exp$ to be understood as the matrix exponential. In our case, similarly to (81), we have

$$\mathbf{\Psi}_t = \begin{pmatrix} I_{d_\theta} & \frac{1-\omega_\theta(t)}{\gamma_\theta} I_{d_\theta} & \multicolumn{2}{c}{0_{2d_\theta \times 2d_x}} \\ 0_{d_\theta} & \omega_\theta(t) I_{d_\theta} & & \\ \multicolumn{2}{c}{0_{2d_x \times 2d_\theta}} & I_{d_x} & \frac{1-\omega_x(t)}{\gamma_x} I_{d_x} \\ & & 0_{d_x} & \omega_x(t) I_{d_x} \end{pmatrix},$$

$$\mathbf{\Psi}_t^{-1} = \begin{pmatrix} I_{d_\theta} & \frac{1-\omega_\theta(-t)}{\gamma_\theta} I_{d_\theta} & \multicolumn{2}{c}{0_{2d_x \times 2d_\theta}} \\ 0_{d_\theta} & \omega_\theta(-t) I_{d_\theta} & & \\ \multicolumn{2}{c}{0_{2d_\theta \times 2d_x}} & I_{d_x} & \frac{1-\omega_x(-t)}{\gamma_x} I_{d_x} \\ & & 0_{d_x} & \omega_x(-t) I_{d_x} \end{pmatrix},$$

where $\omega_x(t) := \exp(-\gamma_x \eta_x t)$, and similarly $\omega_\theta(t) := \exp(-\gamma_\theta \eta_\theta t)$. Hence, we have

$$\begin{aligned}
\tilde{\theta}_t &= \tilde{\theta}_0 + \frac{1}{\gamma_\theta}\left[(1-\omega_\theta(t))\tilde{m}_0 - \left(t - \frac{1-\omega_\theta(t)}{\gamma_\theta \eta_\theta}\right)\nabla_\theta \mathcal{E}(\bar{\theta}_0, \tilde{q}^M_{0,X})\right], \\
\tilde{m}_t &= \omega_\theta(t)\tilde{m}_0 - \frac{1-\omega_\theta(t)}{\gamma_\theta \eta_\theta}\nabla_\theta \mathcal{E}(\bar{\theta}_0, \tilde{q}^M_{0,X}), \\
\forall i \in [M] : \tilde{X}^i_t &= \tilde{X}^i_0 + \frac{1}{\gamma_x}\left[(1-\omega_x(t))\tilde{U}^i_0 + \left(t - \frac{1-\omega_x(t)}{\gamma_x \eta_x}\right)\nabla_x \ell(\bar{\bar{\theta}}_0, \tilde{X}^i_0)\right] + L^{XX}_\Sigma \xi, \\
\forall i \in [M] : \tilde{U}^i_t &= \omega_x(t)\tilde{U}^i_0 + \frac{1-\omega_x(t)}{\gamma_x \eta_x}\nabla_x \ell(\bar{\bar{\theta}}_0, \tilde{X}^i_0) + L^{XU}_\Sigma \xi + L^{UU}_\Sigma \xi',
\end{aligned}$$

where the constants are exactly the same as those from Equation (83), and $\xi, \xi' \sim \mathcal{N}(0, I_{d_x})$.

It can be seen that taking setting $t = h$, the general iterations are given by

$$\begin{aligned}
\tilde{\theta}_{k+1} &= \tilde{\theta}_k + \frac{1}{\gamma_\theta}\left[(1-\omega_\theta(k))\tilde{m}_k - \left(t - \frac{1-\omega_\theta(k)}{\gamma_\theta \eta_\theta}\right)\nabla_\theta \mathcal{E}(\bar{\theta}_k, \tilde{q}^M_{k,X})\right], \\
\tilde{m}_{k+1} &= \omega_\theta(k)\tilde{m}_k - \frac{1-\omega_\theta(k)}{\gamma_\theta \eta_\theta}\nabla_\theta \mathcal{E}(\bar{\theta}_k, \tilde{q}^M_{k,X}), \\
\forall i \in [M] : \tilde{X}^i_{k+1} &= \tilde{X}^i_k + \frac{1}{\gamma_x}\left[(1-\omega_x(k))\tilde{U}^i_k + \left(t - \frac{1-\omega_x(k)}{\gamma_x \eta_x}\right)\nabla_x \ell(\bar{\bar{\theta}}_k, \tilde{X}^i_k)\right] + L^{XX}_\Sigma \xi^i_k, \\
\forall i \in [M] : \tilde{U}^i_{k+1} &= \omega_x(k)\tilde{U}^i_k + \frac{1-\omega_x(k)}{\gamma_x \eta_x}\nabla_x \ell(\bar{\bar{\theta}}_k, \tilde{X}^i_k) + L^{XU}_\Sigma \xi^i_k + L^{UU}_\Sigma \xi'^i_k,
\end{aligned}$$

where $\{\xi^i_k, \xi'^i_k\}_{k,i\in[M]} \sim \mathcal{N}(0, I_{d_x})$.

J. Practical Concerns and Experimental Details

Here, we describe practical considerations and experiment details. First, in Appendix J.1, we describe the RMSProp precondition (Tieleman and Hinton, 2012; Staib et al., 2019) used in our experiments. Then, in Appendix J.2, we describe the subsampling procedure used for our image generation task. After, in Appendix J.3, we discuss the heuristic we introduced for choosing momentum parameters $(\gamma_\theta, \gamma_x, \eta_\theta, \eta_x)$. Finally, in Appendix J.4, we detail all the parameters and models used in our experiment for reproducibility.

J.1. RMSProp Preconditioner

Given $0 < \beta < 1$, the RMSProp preconditioner G_k (Tieleman and Hinton, 2012; Staib et al., 2019) at iteration k is defined by the following iteration:

$$G_{k+1} = \beta G_k + (1-\beta)\nabla_\theta \mathcal{E}(\bar{\theta}_k, q^M_{k,X})^2,$$

where $G_0 = 0_{d_\theta}$ and x^2 denotes element-wise square.

J.2. Subsampling

In the presence of a large dataset, it is common to develop computationally cheaper implementations by appropriate using of data sub-sampling. Such a scheme was devised in Kuntz et al. (2023, Appendix E.4). We utilize the same subsampling scheme for (θ, m)-components. However, we found it necessary to devise a new scheme for (x, u)-components. We found that this alternative subsampling scheme substantially improved the performance of MPD but did not noticeably affect the PGD algorithm.

It is desirable to update the whole particle cloud. However, in cases where each sample has its own posterior cloud approximation, as in the generative modelling task, the dataset can become prohibitively large for updating the whole cloud. In Kuntz et al. (2023), the subsampling scheme only updated the portion of the particle cloud associated with the mini-batch, which neglects the remainder of the posterior approximation. As such, we propose to update the subsampled cloud at the beginning of each iteration at a step proportional to the "time"/steps it has missed. This is described more succinctly in Algorithm 1.

Algorithm 1 A single subsampled step. In pink, we indicate the existing subsampling scheme of Kuntz et al. (2023). We indicate our proposed additions in teal.

Require: Subsampled indices $\mathcal{I}$, subsampled data $\{y^i\}_{i\in\mathcal{I}}$, step-size h, previous iterates $(\theta_k, m_k, \{X_k^i, U_k^i\}_{i\in\mathcal{I}})$
  // Updated cloud for missed time
  $\{X_{k+\epsilon}^i, U_{k+\epsilon}^i\}_{i\in\mathcal{I}} \leftarrow$ Solve Equations (21c) and (21d) with step-size $(h \times missed[\mathcal{I}])$ and $(\theta_k, m_k, \{X_k^i, U_k^i\}_{i\in\mathcal{I}})$.
  // Reset time missed
  $missed[\mathcal{I}] = 0$
  // Take a step with step-size h
  $\theta_{k+1}, m_{k+1} \leftarrow$ Solve Equations (21a) and (21b) with step-size h with $(\theta_k, m_k, \{X_{k+\epsilon}^i, U_{k+\epsilon}^i\}_{i\in\mathcal{I}})$.
  $X_{k+1}^{\mathcal{I}}, u_{k+1}^{\mathcal{I}} \leftarrow$ Solve Equations (21c) and (21d) with step-size h with $(\theta_k, m_k, \{X_{k+\epsilon}^i, U_{k+\epsilon}^i\}_{i\in\mathcal{I}})$..
  // Increment missed time for particles that are not updated
  $missed[\text{not in } \mathcal{I}] = missed[\text{not in } \mathcal{I}] + 1$.

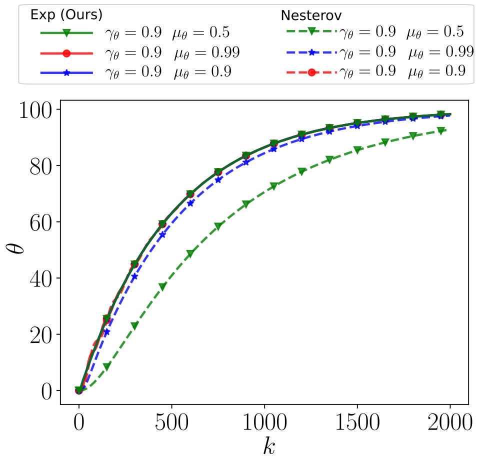
J.3. Heuristic

We (partially) circumvent the choice of momentum parameters by leveraging the relationship between Equation (5) and NAG to define a heuristic. For completeness, we briefly describe the heuristic here. For a given step-size h, one can define the "momentum coefficient" $\mu_\theta = 1 - h\gamma_\theta\eta_\theta$ (and similarly, for μ_x). Since, in NAG, μ has typical values, we can use say $\mu_\theta = 0.9$ with a fixed value of γ_θ to find a suitable value of η_θ (and similarly, for η_x, γ_x). In our experiments, we found that $\gamma_\theta \in [\frac{1}{2}, 5]$ performed well. Another possible approach to handling hyperparameters is to borrow inspiration from adaptive restart methods (O'Donoghue and Candes, 2015). While some practical heuristics exist, it seems that to a large extent, the problem of tuning these hyperparameters remains open; we leave this topic for future work.

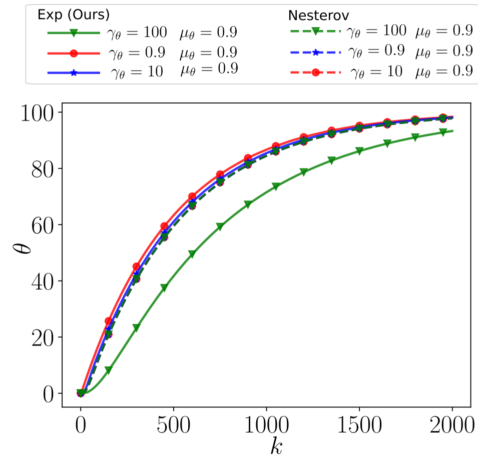
In Appendix I.3.1, we show how we can discretize Equation (15) to obtain the following update in the (θ, m)-components:

$$\begin{aligned}
\theta_k &= \theta_{k-1} + v_k, \\
v_k &= \mu v_{k-1} - h^2 \nabla_\theta \mathcal{E}(\theta_{k-1} + \mu v_{k-1}, q_{k-1,X}^M),
\end{aligned}$$

where $\mu_\theta = 1 - h\gamma_\theta\eta_\theta$, and $h > 0$ is the step size. We refer to the resulting iterations as MPD-NAG-Cheng. This is exactly the update used in Sutskever et al. (2013)'s characterization of NAG with step-size h^2. Since there are some well-accepted choices for choosing μ_θ in NAG (e.g., Nesterov (1983) advocating for $\mu_t = 1 - 3/(t + 5)$), we can leverage the formula $\mu_\theta = 1 - h\gamma_\theta\eta_\theta$ to define a suitable choice of η_θ for a fixed γ_θ. One can, of course, fix η_θ instead, but we found that suitable values for γ_θ are easier to choose from than η_θ. We note that for MPD-NAG-Cheng, it depends only on the value of μ_θ which is not the case for our discretization. In Figure 4, we show the effect of varying the momentum coeffect μ_θ, while keeping the damping parameter γ_θ fixed (and vice versa). In Figure 4a, it can be seen that for a fixed damping parameter γ_θ and varying the momentum coefficient MPD-NAG-Cheng changes significantly while ours does not. As for the vice versa case, in Figure 4b, it can be observed that MPD-NAG-Cheng does not change while ours varies significantly instead. This suggests that the momentum coefficient in our discretization no longer has the same interpretation as that in MPD-NAG-Cheng. Nonetheless, it can be observed that for a suitably chosen γ_θ and momentum parameter μ_θ, our discretization performs better than MPD-NAG-Cheng. Hence, we use this as a heuristic when choosing momentum parameters $(\gamma_\theta, \gamma_x, \eta_\theta, \eta_x)$.



(a) Fixed γ_θ, varying μ_θ.



(b) Fixed μ_θ, varying γ_θ.

Figure 4: The effect of the damping parameter γ_θ and momentum coefficient μ_θ on MPD.

J.4. Experimental Details

Here, we outline for reproducibility the settings: for the Toy Hierarchical Model (Appendix J.4.1); image generation experiment (Appendix J.4.2); and, additional figures are in Appendix J.6. Much akin to Kuntz et al. (2023, Section 2), we found it beneficial to consider separate time-scales for the (θ_t, m_t) and (X_t, U_t) which we denote h_θ and h_x respectively.

J.4.1. Toy Hierarchical Model

Model. The model is described in Kuntz et al. (2023, Example 1). For completeness, we describe it here. The model is defined by

$$p_\theta(x, y) = \prod_{i=1}^{d_x} \mathcal{N}(y_i; x_i, 1)\mathcal{N}(x_i; \theta, \sigma^2). \tag{84}$$

Thus, for the log-likelihood with $\sigma^2 = 1$, we have

$$\log p_\theta(x, y) = C - \frac{1}{2}\sum_{i=1}^{d_x} \left((x_i - \theta)^2 + (y_i - x_i)^2\right),$$

where C is a constant independent of θ and x.

Figure 1a. For the different regimes, we use the following momentum parameters:

- **Underdamped**: $\gamma_\theta = 0.1, \eta_\theta = 403.96, \gamma_x = 0.1, \eta_x = 403.96$.
- **Overdamped**: $\gamma_\theta = 1, \eta_\theta = 403.96, \gamma_x = 1, \eta_x = 403.96$.
- **Critically Damped** $\gamma_\theta = 0.7, \eta_\theta = 403.96, \gamma_x = 0.7, \eta_x = 403.96$.

The step-sizes are $h_\theta = 10^{-2}/N = 10^{-4}$, $h_x = 10^{-2}$, with number of samples $N = 100$, and number of particles $M = 100$.

Figure 1b. We compare the discretization outlined in Appendix I.3, and Section 5. We keep all parameters equal except for changing the momentum coefficient μ. For step-sizes, we have $h_\theta = 10^{-5/2} \approx 5.8 \times 10^{-3}, h_x = 10^{-3}, \gamma_x = \gamma_\theta = 0.5$, with the momentum coefficient set to $\mu_\theta = \mu_x = \mu$ where $\mu \in \{0.9, 0.8, 0.5\}$.

Lipschitz Constant and Strong log concavity. For the toy Hierarchical model with $\sigma^2 = 1$, one can show that it satisfies our Lipschitz assumption, as well as being strongly log-concave. We have

$$\nabla^2_{(\theta,x)} \log p_\theta = \begin{pmatrix} -d_x & 1_{1\times d_x} \\ 1_{d_x\times 1} & -2I_{d_x} \end{pmatrix}.$$

The characteristic equation can be written as

$$\det(\nabla^2_{(\theta,x)} \log p_\theta - lI_{d_x+1}) = 0.$$

The determinant can be evaluated by expanding the minors. Thus, we obtain

$$\begin{aligned}\det(\nabla^2_{(\theta,x)} \log p_\theta - lI_{d_x+1}) =&(-d_x - l)(-2-l)^{d_x} - d_x(-2-l)^{d_x-1}\\ =&(-2-l)^{d_x-1}((2+l)(d_x+l)-d_x)\\ =&(-2-l)^{d_x-1}(l^2+(2+d_x)l+d_x).\end{aligned}$$

Hence, the characteristic equation is satisfied when $l \in \{-2, \frac{-(2+d_x)\pm\sqrt{d_x^2+4}}{2}\}$. Note that $\frac{-(2+d_x)-\sqrt{d_x^2+4}}{2} \leq -2$.

From the above calculations, it can be seen that $\nabla^2_{(\theta,x)} \log p_\theta$ is strongly-log concave, i.e., we have $\nabla^2_{(\theta,x)} \log p_\theta \preceq -2I$.

In order to calculate its Lipschitz constant of Assumption 1, by Nesterov (2003, Lemma 1.2.2), this is equivalent to finding a constant $K_{hm} > 0$ such that

$$\|\nabla^2_{(\theta,x)} \log p_\theta\| \leq K_{hm}.$$

Since $\nabla^2_{(\theta,x)} \log p_\theta$ is symmetric, then the matrix norm is given by the largest absolute value of the eigenvalue of $\nabla^2_{(\theta,x)} \log p_\theta$ (i.e., it's spectral radius).

Hence, we have $K_{hm} = \frac{(2+d_x)+\sqrt{d_x^2+4}}{2}$.

J.4.2. Image Generation

The dataset of both MNIST and CIFAR-10 is processed similarly. We use $N = 5000$ images for training and 3000 for testing. The model is updated 6280 times using subsampling with a batch size of 32. Hence, it completes 40 epochs. The model is defined as:

$$p_\theta(y,x) = \prod_{i=1}^N p_\theta(y^i, x^i),$$

where:

- The datum $(y^i, x^i) \in \mathbb{R}^{d_y} \times \mathbb{R}^{d_x}$ denotes a single image and its corresponding latent variable. For MNIST, we have $d_y = 28 \times 28 = 784$ and for CIFAR, we have $d_y = 32 \times 32 \times 3 = 3072$. For both datasets, we set $d_x = 64$. Thus, we have that $(y, x) \in \mathbb{R}^{d_y \times N} \times \mathbb{R}^{d_x \times N}$.
- For the VAE model, we have $p(y^i, x^i) = \mathcal{N}(y^i | g_\psi(x^i), \sigma^2 I)p_\phi(x^i)$, where ψ and ϕ are parameters of the generator and prior respectively. Thus, the parameter of the model is given by $\theta = \{\psi, \phi\}$. We set $\sigma^2 = 0.1$. The following specifies the details regarding the generator g_ψ and prior p_ϕ:
  - **Generator** g_ψ. For CIFAR, we use a Convolution Transpose network (as in Pang et al. (2020)) shown in Table 2. For MNIST, we use a fully connected network specified in Table 3 with $d_{in} = d_x$ and $d_{out} = d_y$.
  - **Prior** $p_\phi(x)$**.** The prior is inspired by the VampPrior (Tomczak and Welling, 2018) and is defined as:

  $$p_\phi(x) = \frac{1}{K}\sum_{i=1}^K \mathcal{N}(x|\mu_\nu(z_i), \sigma^2_\nu(z_i)I_{d_z}),$$

  where $\mu_\nu, \sigma^2_\nu : \mathbb{R}^{d_z} \to \mathbb{R}^{d_x}$ are neural networks (with parameters ν) that parameterized the mean and variance; pseudo-points $\{z_i\}_{i=1}^K$ are optimized; and so the prior parameters are $\phi := \{\nu\} \cup \{z_i\}_{i=1}^K$. The architecture can be found in Table 4 where $d_{in} = d_z$ and $d_{out} = d_z$

Layers	Size	Stride	Pad
Input	1x1x128	-	-
8x8 ConvT(ngf x 8), LReLU	8x8x(ngf x 8)	1	0
4x4 ConvT(ngf x 4), LReLU	16x16x(ngf x 4)	2	1
4x4 ConvT(ngf x 2), LReLU	32x32x(ngf x 2)	2	1
3x3 ConvT(3), Tanh	32x32x3	1	1

Table 2: Generator network for the VAE used for CIFAR-10 (ngf = 256).

Layers	Size
Input	d_x
Linear(d_x, 512), LReLU	512
Linear(512,512), LReLU	512
Linear(512,512), LReLU	512
Linear(512,d_y), Tanh	d_y

Table 3: Generator network for the VAE used for MNIST.

Layers	Size
Input	d_{in}
Linear(d_{in}, 512), LReLU	512
Linear(512,512), LReLU	512
Linear(512,512), LReLU	512
Linear(512,$d_{out} \times 2$)	$d_{out} \times 2$
Output: Id([:d_{out}]), Softplus([d_{out}:])	d_{out}, d_{out}

Table 4: Neural network parametrization of the mean and variance of a Gaussian Distribution used in VI and VAMPprior. "Id" is the identity function, and "[:d_{out}]" is a standard pseudo-code notation that refers to the operation that extracts the first d_{out} dimensions and similarly for [d_{out}:].

J.5. Hyperparameter settings

Unless specified otherwise, we use the same parameters for both MNIST and CIFAR datasets. Our approach to selecting the step size involved first setting it to 10^{-3}, then we would decrease it appropriately if instabilities arise.

- **MPD**. For step sizes, we have $h_\theta = h_x = 10^{-4}$. The number of particles is set to $M = 5$. For the momentum parameters, we use $\gamma_\theta = \gamma_x = 0.9$ with the momentum coefficient $\mu_\theta = 0.95, \mu_x = 0$ (or equivalently, $\eta_\theta \approx 556, \eta_x \approx 11,111$)for MNIST and $\mu_\theta = 0.95, \mu_x = 0.5$ (or equivalently, $\eta_\theta \approx 556, , \eta_x \approx 55,555$) for CIFAR. We use the RMSProp preconditioner (see Appendix J.1) with $\beta = 0.9$.

- **PGD**. We have $h_\theta = 10^{-4}, h_x = 10^{-3}$. The number of particles is set to $M = 5$. We use the RMSProp preconditioner (see Appendix J.1) with $\beta = 0.9$.

- **ABP**. We use SGD optimizer with step size $h_\theta = 10^{-4}$, and run a length 5 ULA chain 5 with step-size $h_x = 10^{-3}$. We found that RMSProp worked well in the transient period but failed to achieve better performance than SGD.

- **SR**. We used RMSProp optimizer with step size $h_\theta = 10^{-4}$, and ran five 20-length ULA chain with step-size $h_x = 10^{-3}$

- **VI**. We use RMSProp optimizer with a step size of $h = 10^{-5}$. We found that this resulted in the best performance. For the encoder, we use a Mean Field approximation as in Kingma and Welling (2014). The neural network is a 3-layer fully connected network with Leaky Relu activation functions. In the last layer, we use a softmax to parameterize the variance. See Table 4.

J.6. Additional Figures

(a) MPD. (b) PGD. (c) VI. (d) ABP.

Figure 5: **MNIST**. Samples generated from various algorithms.

(a) MPD. (b) PGD. (c) VI. (d) ABP.

Figure 6: **CIFAR-10**. Samples generated from various algorithms.

K. Additional Experiments

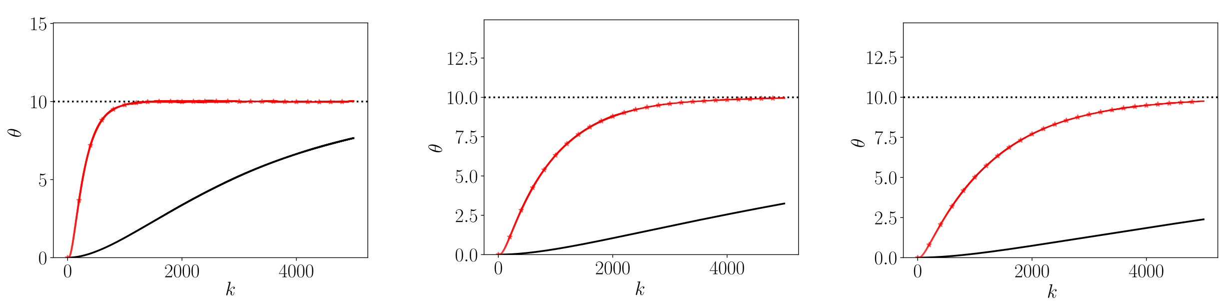
In this section, we show additional experiments particularly concerned with comparing MPD with PGD and other variants of MPD that only accelerated one component of the space instead of two. We say X-enriched to indicate algorithms with momentum incorporated in X with either momentum included/excluded in θ. To separate the two cases, we call X-only-enriched to refer to the algorithm with gradient descent in θ. We use similar terminology for θ-enriched and θ-only-enriched.

We consider two settings: in Appendix K.1, further results with the toy Hierarchical model (see Appendix J.4.1); and, in Appendix K.2 a density estimation task using VAEs on a Mixture of Gaussian dataset.

K.1. Toy Hierarchical Model

Figure 7. As the model, we consider a Toy Hierarchical model for different σ values in (84) where σ is chosen to be the same as the data generating process. The data is generated from a toy Hierarchical model with parameters $(\theta = 10, \sigma)$ where $\sigma = \{5, 10, 12\}$. As noted by Kuntz et al. (2023, Eq. 51), the marginal maximum likelihood of θ is the empirical average of the observed samples, i.e., $\frac{1}{d_x}\sum_{i=1}^{d_x} y_i$. For each trial, we process the data $(y_i)_{i=1}^{d_x}$ to have an empirical average of 10 for simplicity.

Figure 2 (a), (b), (c). We use a toy Hierarchical model with $\sigma = 12$, and, similarly, the data-generating processing is a toy Hierarchical model with $(\theta = 10, \sigma = 12)$. We are interested in how the initialization of the particle cloud affects MPD, PGD and other algorithms that only momentum-enriched one component. Each subplot shows the evolution of the parameter θ for the initialization from $\mathcal{N}(\mu, I)$ where $\mu \in \{-5, -20, -100\}$. This example serves as an illustration of the typical case where the cloud is initialized badly. It can be seen that all methods are affected by poor initialization. For X-enriched, the algorithms can recover faster to achieve almost identical performances (at the later iterations) between different initializations. In other words, for X-enriched algorithms transient phase is short. Methods without such X-enrichment take longer to recover and are unable to achieve similar performances compared to the cases where they are better-initialized. Overall, it can be seen that MPD perform better than PGD.



(a) $\sigma = 5$. (b) $\sigma = 10$. (c) $\sigma = 12$.

Figure 7: Comparison between MPD and PGD on the Toy Hierarchical Model for different choices of σ. The plot shows the average and standard deviation of θ across iterations computed over 10 independent trials. MPD is shown in red and PGD in black. The dashed line shows the true value.

K.2. Density Estimation

In this problem, we show the result of training a VAE with VAMP Prior on a 1-d Mixture of Gaussians (MoG) dataset with PDF shown in Figure 8. The dataset is composed of 100 samples. In Figure 2 (d), we show the performance of various acceleration algorithms. The empirical Wasserstein-1 distance (denoted by $\hat{\mathrm{W}}_1$) is estimated from the empirical CDF produced from 1000 samples across each iteration. The results show the average (with standard error bars which are barely noticeable) computed over 100 trials. PGD is shown in black, and MPD is shown in green. In this case, all accelerated algorithms perform better than PGD, and MPD performs best of all.

Hyperparameters: No subsampling used. We have $d_y = 1, d_x = 10$. We set $\gamma_x = \gamma_\theta = 0.4$, and use the momentum heuristic with $\mu_\theta = \mu_x = 0.1$ (see Appendix J.3). We use the same VAE architecture in Appendix J.4.2 with likelihood noise $\sigma = 0.01$, an MLP decoder (see Table 4), and VAMP prior with $d_z = 2$ and $K = 20$.

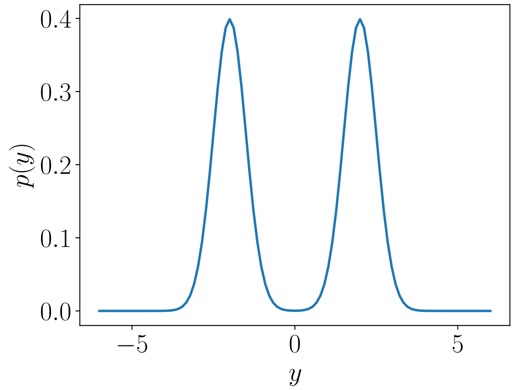


Figure 8: The true density of the MoG dataset.

K.3. Sensitivity Analysis

We are interested in investigating the sensitivity of MPD to different momentum parameters. One way to quantify whether you are in a regime where MPD is an improvement to PGD is by computing an approximation to the (signed and weighted) area between PGD and the MPD curves called *Area Between Curves* (ABC) More specifically, we compute

$$\mathrm{ABC} := \frac{1}{K} \sum_{k=1}^{K} w(k)[\mathrm{PGD}(k) - \mathrm{MPD}(k)],$$

where $\mathrm{PGD} : \mathbb{Z}_+ \to \mathbb{R}$ and $\mathrm{MPD} : \mathbb{Z}_+ \to \mathbb{R}$ denote the loss/metric curves of PGD and MPD respectively, and $w(k)$ is a weighting function such that $\sum_{k=1}^{K} w(k) = 1$. The role of the factor $\frac{1}{K}$ is to normalize and $w(k)$ is to reweight the sum to have more importance at the end of the run than the beginning (we use $w(k) = \frac{2}{K(1+K)} \cdot k/K$). If MPD is in a desirable regime and performs better than PGD then the curve of MPD will lower bound the curve of PGD. In this case, ABC will obtain large positive values and in the converse case, ABC will have small negative, hence a higher ABC indicates better hyperparameters. Figure 9 shows the ABC for different momentum parameters γ and η while keeping all other

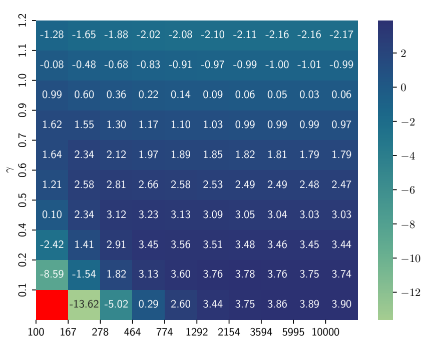


Figure 9: The Area Between the Curve of PGD and MPD for different momentum parameters.

hyperparameters the same between PGD and MPD. We use the toy Hierarchical model which allows us to have access to the true parameter θ, so ABC is computed from curves MPD and PGD that returns the absolute difference between the model and true θ. We note that a similar heatmap can be produced using the loss curve. This heatmap supports our recommendation that if we set η to be sufficiently large, we can easily tune γ to achieve better performance than PGD. Note that η-axis is in log scale and that this plot only applies to toy Hierarchical Model.