Appendix to "Global convergence of neuron birth-death dynamics"

Grant Rotskoff [*1], Samy Jelassi[1,3], Joan Bruna [†1,2], and Eric Vanden-Eijnden [‡1]

[1]Courant Institute of Mathematical Sciences, New York University
[2]Center for Data Science, New York University
[3]Princeton University

May 13, 2019

A Generalizations of the birth-death PDE

Here we mention two ways in which we can modify (13) to certain advantages. For example, we can replace this equation with

$$\partial_t \mu_t = \nabla \cdot (\mu_t \nabla V) - \alpha f(V - \bar{V})\mu_t - \bar{f}\mu_t, \tag{1}$$

where $f : \mathbb{R} \to \mathbb{R}$ is some function and $\bar{f} = \int_D f(V - \bar{V}) d\mu_t$. As we will see in Proposition A.1, as long as $zf(z) \geq 0$ for all $z \in \mathbb{R}$, the additional term in (1) increase the rate of decay of the energy.

While the birth-death dynamics described above ensures convergence in the mean-field limit, when n is finite, particles can only be created in proportion to the empirical distribution $\mu^{(n)}$. In particular, such a birth process corresponds to "cloning" or creating identical replicas of existing particles. In practice, there may be an advantage to exploring parameter space with a distribution distinct from the instantaneous empirical particle distribution (7). To enable this exploration we introduce a birth term proportional to a distribution μ_{b} which we will assume has full support on D. In this case, the time evolution of the distribution is described by

$$\begin{aligned} \partial_t \mu_t = \nabla \cdot (\mu_t \nabla V) - \alpha (V - \bar{V})_+ \mu_t + \alpha \left(\textstyle\int_D (V - \bar{V})_+ d\mu_t \right) \frac{\mu_{\mathrm{b}} \mathbb{1}_{V \leq \bar{V}}}{\mu_b(V \leq \bar{V})} \\ + \alpha' (V - \bar{V})_- \mu_{\mathrm{b}} - \alpha' \left(\textstyle\int_D (V - \bar{V})_- d\mu_{\mathrm{b}} \right) \frac{\mu_t \mathbb{1}_{V > \bar{V}}}{\mu_t(V > \bar{V})}, \end{aligned} \tag{2}$$

where $\alpha, \alpha' > 0$, $(V - \bar{V})_+ = \max(V - \bar{V}, 0) \geq 0$, $(V - \bar{V})_- = \max(\bar{V} - V, 0) \geq 0$. That is, we kill particles in proportion to μ_t in region where $V > \bar{V}$ but create new particles from μ_{b} in regions where $V \leq \bar{V}$. We could also combine (1) with (2) to obtain other variants.

These alternative birth-death dynamical schemes also satisfy the consistency conditions of Proposition 3.1:

Proposition A.1 *Let μ_t be a solution of (1) with f such that $zf(z) \geq 0$ for all $z \in \mathbb{R}$ or (2), with $\mu_0 \in \mathcal{M}(D)$. Then, $\mu_t(D) = 1$ for all $t \geq 0$, and $E(t) = \mathcal{E}[\mu_t]$ satisfies*

$$\dot{E}(t) \leq - \int_D |\nabla V(\boldsymbol{\theta}, [\mu_t])|^2 \mu_t(d\boldsymbol{\theta}) \,. \tag{3}$$

[*]This work was partially supported by the James S. McDonnell Foundation.
[†]This work was partially supported by the Alfred P. Sloan Foundation and NSF RI-1816753.
[‡]This work was partially supported by the Materials Research Science and Engineering Center(MRSEC) program of the National Science Foundation(NSF) under award number DMR-1420073 and by NSF under award number DMS-1522767.

Proof: By considering again 1 and $V(\cdot,[\mu_t])$ as a test function in (1) or (2), we verify that $\partial_t\mu_t(D)=0$. In addition, (1) implies that

$$\begin{aligned}\dot{E}(t) &= \int_D V(\boldsymbol{\theta},[\mu_t])\partial_t\mu_t(d\boldsymbol{\theta}) \\ &= \int_D \left(V(\boldsymbol{\theta},[\mu_t]) - \bar{V}[\mu_t]\right)\partial_t\mu_t(d\boldsymbol{\theta}) \\ &= -\int_D |\nabla V|^2 d\mu_t - \alpha\int_D (V-\bar{V})f(V-\bar{V})d\mu_t\end{aligned}$$

which proves (3) for (1) since all the terms at the right hand side of this equation are negative individually if $zf(z)\geq 0$ for all $z\in\mathbb{R}$. Similarly, (2) implies that

$$\begin{aligned}\dot{E}(t) &= \int_D V(\boldsymbol{\theta},[\mu_t])\partial_t\mu_t(d\boldsymbol{\theta}) \\ &= \int_D \left(V(\boldsymbol{\theta},[\mu_t]) - \bar{V}[\mu_t]\right)\partial_t\mu_t(d\boldsymbol{\theta}) \\ &= -\int_D |\nabla V|^2 d\mu_t - \alpha\int_D (V-\bar{V})_+^2 d\mu_t - \alpha\frac{\int_D (V-\bar{V})_+ d\mu_t \int_D (V-\bar{V})_- d\mu_{\mathrm{b}}}{\mu_b(V\leq 0)} \\ &\quad - \alpha'\int_D (V-\bar{V})_-^2 d\mu_{\mathrm{b}} - \alpha'\frac{\int_D (V-\bar{V})_- d\mu_{\mathrm{b}} \int_D (V-\bar{V})_+ d\mu_t}{\mu_t(V>0)},\end{aligned}$$

which proves (3) for (2) since all the terms at the right hand side of this equation are negative. □

B Proximal formulation of birth-death dynamics

Following the framework of [JKO98], we can give an alternative interpretation to the birth-death PDE (13). First, we recall that the PDE (8) can be obtained as the time-continuous limit ($\tau\to 0$) of the proximal optimization scheme (also known as minimizing movement scheme [San17]) in which a sequence of distributions $\{\mu_k\}_{k\in\mathbb{N}_0}$ is constructed via the iteration: given an initial μ_0 such that $\mathcal{E}[\mu_0]<\infty$, set

$$\mu_{k+1}\in \operatorname{argmin}\left(\mathcal{E}[\mu] + \tfrac{1}{2}\tau^{-1}W_2^2(\mu,\mu_k)\right), \tag{4}$$

for $k=0,1,2,\ldots$ where $W_2(\mu,\mu_k)$ denotes the 2-Wasserstein distance between the probability measures μ and μ_k. Interestingly, the birth-death PDE relies on a different measure of "distance": the PDE

$$\partial_t\mu_t = -\alpha V\mu_t + \alpha\bar{V}\mu_t, \tag{5}$$

can be obtained as the time-continuous limit of the proximal optimization scheme: given an initial μ_0 such that $\mathcal{E}[\mu_0]<\infty$, set for $k=0,1,2,\ldots$

$$\mu_{k+1}\in \operatorname{argmin}\left(\mathcal{E}[\mu] + (\alpha\tau)^{-1}D_{\mathrm{KL}}(\mu||\mu_k)\right) \tag{6}$$

where the minimum is taken over all probability measures $\mu\in\mathcal{M}(D)$ and $D_{\mathrm{KL}}(\mu||\mu_k)$ is the Kullback-Leibler divergence

$$D_{\mathrm{KL}}(\mu||\mu_k) = \int_D \log\left(\frac{d\mu}{d\mu_k}\right)d\mu\ . \tag{7}$$

We verify this claim formally; notice that the Euler-Lagrange equation for the minimizer μ_{k+1}, obtained by zeroing the first variation of the objective function in (6), reads

$$V(\boldsymbol{\theta},[\mu_{k+1}]) + (\alpha\tau)^{-1}\log\left(\frac{d\mu_{k+1}}{d\mu_k}\right) + \lambda = 0 \tag{8}$$

where λ is a Lagrange multiplier added to enforce $\int_D d\mu_{k+1}=1$. (8) can be reorganized into

$$\mu_{k+1} = C^{-1}\mu_k\exp\left(-\alpha\tau V(\boldsymbol{\theta},[\mu_{k+1}])\right) \tag{9}$$

where C is adjusted so that $\int_D d\mu_{k+1} = 1$. (9) is the discrete equivalent of (14) If τ is small, we can expand the exponential to arrive at

$$\mu_{k+1} = C^{-1}\left(\mu_k - \alpha\tau V(\boldsymbol{\theta}, [\mu_{k+1}])\mu_k + O(\tau^2)\right) \tag{10}$$

Setting $\mu_{k+1} = \mu_k + O(\tau)$ in V and expanding again gives

$$\mu_{k+1} = \mu_k - \alpha\tau\left[V(\boldsymbol{\theta}, [\mu_k])\mu_k + \bar{V}[\mu_k]\right]\mu_k + O(\tau^2) \tag{11}$$

where we have also expanded C and solved for it explicitly at leading order in τ. Subtracting μ_k for both sides, dividing by τ, and letting $\tau \to 0$ gives (5). The full PDE (13) can be obtained by alternating (4) and (6).

Note that, under Assumption 4.2 below, the energy $\mathcal{E}[\mu]$ is convex and bounded below. As a result the augmented functionals to minimize in both (4) and (6) are strictly convex, which means that they admit a unique minimizer. This shows that the measures in the sequence $\{\mu_k\}_{k\in\mathbb{N}_0}$ are well-defined and such that $\mathcal{E}[\mu_{k+1}] \leq \mathcal{E}[\mu_k]$ whether we use (4), (6), or alternate between both. Because we discretize time in practice, solutions of (13) satisfying (14) for all $t > 0$ can be interpreted as implementations of the proximal scheme. Taking the limit $\tau \to 0$ with $k\tau$ large, however, requires ensuring well-definedness of the terms on the right hand side of (13). This proximal interpretation also enables the design of distinct algorithms for implementing this PDE at particle level, as discussed next.

B.1 Proximal Optimization

For concreteness we focus on the cases of neural networks—the ideas below can be easily adapted to the others situations treated in this paper. Assume that the neural representation at iterate k is

$$f_k^{(n)}(\boldsymbol{x}) = \frac{1}{n}\sum_{i=1}^{n} w_i^k \varphi(\boldsymbol{x}, \boldsymbol{\theta}_i^k) \tag{12}$$

where $\boldsymbol{\theta}_i^k$ denotes the parameter in the network and $w_i^k \geq 0$ are extra weights satisfying $n^{-1}\sum_{i=1}^n w_i^k = 1$—we will define a dynamics for these weights in a moment. Notice that (12) can be written as

$$f_k^{(n)}(\boldsymbol{x}) = \int_D \phi(\boldsymbol{x}, \boldsymbol{\theta}) d\mu_k^{(n)}(\boldsymbol{\theta}), \qquad d\mu_k^{(n)}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} w_i^k \delta_{\boldsymbol{\theta}_i^k}(d\boldsymbol{\theta}) \tag{13}$$

and the loss is given by

$$\begin{aligned}\ell(\boldsymbol{\theta}_1^k, \ldots, \boldsymbol{\theta}_n^k; w_1^k, \ldots, w_n^k) &= \tfrac{1}{2}\mathbb{E}_{y,\boldsymbol{x}}|y - f_k^{(n)}(\boldsymbol{x})|^2 \\ &= C_f + \frac{1}{n}\sum_{i=1}^{n} w_i^k F(\boldsymbol{\theta}_i^k) + \frac{1}{2n^2}\sum_{i,j=1}^{n} w_i^k w_j^k K(\boldsymbol{\theta}_i^k, \boldsymbol{\theta}_j^k)\end{aligned} \tag{14}$$

where $C_f = \frac{1}{2}\mathbb{E}_y y^2$ and $F(\boldsymbol{\theta})$ and $K(\boldsymbol{\theta}, \boldsymbol{\theta}')$ given in (4) and (5), respectively. The scheme we propose will update the $\boldsymbol{\theta}_i^k$ and the w_i^k separately, the first by usual gradient descent over the loss, the second by proximal gradient. That is, given $\{\boldsymbol{\theta}_i^k\}_{i=1}^n$ and $\{w_i^k\}_{i=1}^n$:

1. *Gradient step.* Evolve the parameters $\boldsymbol{\theta}_i^k$ by GD (or SGD if we need to use the empirical loss) with the weights w_i^k kept fixed. Do this for m steps of size Δt to obtain a new set of $\{\boldsymbol{\theta}_i^{k+1}\}_{i=1}^n$.

2. *Proximal step.* Evolve the weights w_i^k with the parameter $\boldsymbol{\theta}_i^{k+1}$ fixed using a proximal step based on the particle equivalent of (6), i.e.

$$\{w_i^{k+1}\}_{i=1}^n \in \operatorname{argmin}\left(\ell(\boldsymbol{\theta}_1^{k+1}, \ldots, \boldsymbol{\theta}_n^{k+1}; w_1, \ldots, w_n) + \frac{1}{\tau n}\sum_{i=1}^{n} w_i \log(w_i/w_i^k)\right) \tag{15}$$

where the minimization is done under the constraint that $n^{-1}\sum_{i=1}^n w_i = 1$. The equation for the minimizer w_i^{k+1} is the discrete equivalent of (10)

$$w_i^{k+1} = C^{-1} w_i^k \exp\left(-\tau\tilde{V}_i^{k+1}\right) \tag{16}$$

where C is a constant to be adjusted so that $n^{-1}\sum_{i=1}^n w_i^{k+1} = 1$ and

$$\tilde{V}_i^{k+1} = F(\boldsymbol{\theta}_i^{k+1}) + \frac{1}{n}\sum_{j=1}^n w_j^{k+1} K(\boldsymbol{\theta}_i^{k+1}, \boldsymbol{\theta}_j^{k+1}) \tag{17}$$

(16) is implicit in w_i^{k+1} and should be solved by iteration. Note that this proximal step is guaranteed to decrease the loss. In practice, this step could eventually lead to big variations of the weights. Should this happen, we add the additional step:

3. *Resampling step.* Resample the weights $\{w_i^{k+1}\}_{i=1}^n$ so as to keep them roughly equal to 1 each, that is: eliminate the ones that are too small and transfer their weights to the others: split the remaining (large) weights into bits of size roughly 1. There are standard ways to do this resampling step that are unbiased and preserve the population size exactly. This resampling step may increase the loss, though not to leading order. This step is the actual birth-death step in the scheme (and it is also the only random component of it if the exact loss is used).

If we set $\tau = \alpha m \Delta t$ and set $\Delta t \to 0$ and $n \to \infty$, the scheme above is formally consistent with the PDE

$$\partial_t \mu_t = \nabla \cdot (\nabla V \mu_t) - \alpha V \mu_t + \alpha \bar{V} \mu_t. \tag{18}$$

However, it is obviously not necessary to take either of these limits explicitly in practice, and, as explained above, the proximal step is guaranteed to decrease the loss. With a strict version of the the resampling step performed at every iteration, in which the weights are taken to be in $\{0, 1\}$ the scheme above recovers the one described in Algorithm 1. The main difference is that in Algorithm 1 the proximal step (16) is solved in one iteration, by substituting w_i^{k+1} by w_i^k at the right hand side of (16).

Finally notice that if we were to implement the proximal step only and skip both the gradient and the resampling steps, the scheme above is a naive implementation of the lazy training scheme discussed in [CB18]. This highlights again why using birth-death alone is not an efficient way to perform network optimization, and it should be combined with standard GD.

C Convergence and Rates in the Non-interacting Case

C.1 Non-interacting Case

We consider first the non-interacting case with $V = F$ and $D = \mathbb{R}^k$, under

Assumption C.1 *$F \in C^2(\mathbb{R}^k)$ is a Morse function, coercive, and with a single global minimum located at $\boldsymbol{\theta}^*$.*

With no loss of generality we set $F(\boldsymbol{\theta}^*) = 0$ since adding an offset to F in (13) does not affect the dynamics. We also denote by $H^* = \nabla\nabla F(\boldsymbol{\theta}^*)$ the Hessian of F at $\boldsymbol{\theta}^*$: recall that a Morse function is such that its Hessian is nondegenerate at all its critical points (where $\nabla F = 0$) and it is coercive if $\lim_{\boldsymbol{\theta}\to\infty} F(\boldsymbol{\theta}) = \infty$. Our main result is

Theorem C.2 (Global Convergence and Rate: Non-interacting Case) *Assume that the initial condition μ_0 of the PDE (12) has a density ρ_0 positive everywhere in $\mathbb{R}^k$ and is such that $\mathcal{E}[\mu_0] < \infty$. Then under Assumption C.1 the solution of (12) satisfies*

$$\mu_t \rightharpoonup \delta_{\boldsymbol{\theta}^*} \qquad as \quad t \to \infty. \tag{19}$$

In addition we can quantify the convergence rate: if $\bar{F}(t) = \int_{\mathbb{R}^k} F(\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta})$, then $\exists C > 0$ such that $\forall \epsilon > 0$, the time t_ϵ needed to reach $\mathcal{E}[\mu_{t_\epsilon}] \le \epsilon$ satisfies

$$t_\epsilon \le C\epsilon^{-(d+2)/2}. \tag{20}$$

Furthermore the rate of convergence becomes exponential in time asymptotically: for all $\delta > 0$, $\exists t_\delta$ such that

$$\bar{F}(t) \le \alpha^{-1} \operatorname{tr}\left(H^* e^{-2H^*(t-\delta)}\right) \quad if \quad t \ge t_\delta. \tag{21}$$

In fact we show that

$$\lim_{t\to\infty} \frac{\alpha \bar{F}(t)}{\operatorname{tr}\left(H^* e^{-2H^* t}\right)} = 1. \tag{22}$$

The theorem is proven in Appendix C This proof shows that the additional birth-death terms in the PDE (12) allow the measure to concentrate rapidly in the vicinity of $\boldsymbol{\theta}^*$; subsequently, the transport term takes over and leads to the exponential rate of energy decay in (21). The proof also shows that, if we remove the transportation term $\nabla \cdot (\mu_t \nabla V)$ in the PDE (12), the energy only decreases linearly in time asymptotically. This means that the combination of the transportation and the birth-death terms accelerates convergence. A similar theorem can be proven for the PDE (2).

C.2 Non-interacting Case without the Transportation Term

Let us look first at the PDE satisfied by the measure μ in the non-interacting case, i.e. with $V = F$ satisfying Assumption C.1, and without the transportation term:

$$\partial_t \mu_t = -\alpha F(\boldsymbol{\theta}) \mu_t + \alpha \bar{F}(t) \mu_t, \tag{23}$$

where $\bar{F}(t) = \int_{\mathbb{R}^k} F(\boldsymbol{\theta}) \mu_t(d\boldsymbol{\theta})$. This equation can be solved exactly. Assuming that μ_0 has a density everywhere positive on $\mathbb{R}^k$, μ_t has a density ρ_t given by

$$\rho_t(\boldsymbol{\theta}) = e^{\alpha \int_0^t \bar{F}(s) ds - \alpha t F(\boldsymbol{\theta})} \rho_0(\boldsymbol{\theta}). \tag{24}$$

The normalization condition $\mu_t(\mathbb{R}^k) = \int_{\mathbb{R}^k} \rho_t(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$ leads to:

$$\begin{aligned} & e^{\alpha \int_0^t \bar{F}(s) ds} \int_{\mathbb{R}^k} e^{-\alpha t F(\boldsymbol{\theta}')} \rho_0(\boldsymbol{\theta}') d\boldsymbol{\theta}' = 1 \\ \Leftrightarrow \quad & e^{-\alpha \int_0^t \bar{F}(s) ds} = \int_{\mathbb{R}^k} e^{-\alpha t F(\boldsymbol{\theta}')} \rho_0(\boldsymbol{\theta}') d\boldsymbol{\theta}'. \end{aligned}$$

Therefore, by plugging this last expression in equation (24), we obtain the explicit expression

$$\rho_t(\boldsymbol{\theta}) = \frac{e^{-\alpha t F(\boldsymbol{\theta})} \rho_0(\boldsymbol{\theta})}{\int_{\mathbb{R}^k} e^{-\alpha t F(\boldsymbol{\theta}')} \rho_0(\boldsymbol{\theta}') d\boldsymbol{\theta}'}. \tag{25}$$

We can use this equation to express the energy $\bar{F}(t) = \int_{\mathbb{R}^k} F(\boldsymbol{\theta}) \rho_t(\boldsymbol{\theta}) d\boldsymbol{\theta}$:

$$\bar{F}(t) = \frac{\int_{\mathbb{R}^k} F(\boldsymbol{\theta}) e^{-\alpha t F(\boldsymbol{\theta})} \rho_0(\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int_{\mathbb{R}^k} e^{-\alpha t F(\boldsymbol{\theta})} \rho_0(\boldsymbol{\theta}) d\boldsymbol{\theta}} = \frac{d}{d\alpha t} G(\alpha t), \tag{26}$$

where $G(\alpha t)$ is the function defined as:

$$G(\alpha t) = -\log \int_{\mathbb{R}^k} e^{-\alpha t F(\boldsymbol{\theta})} \rho_0(\boldsymbol{\theta}) d\boldsymbol{\theta}. \tag{27}$$

At late times, the factor $e^{-\alpha t F(\boldsymbol{\theta})}$ focuses all the mass in the vicinity of the global minimum of F. Therefore, we can neglect the influence of the density ρ_0 in this integral. More precisely a calculation using the Laplace method indicates that

$$\int_{\mathbb{R}^k} e^{-\alpha t F(\boldsymbol{\theta})} d\boldsymbol{\theta} \sim (2\pi)^{d/2} (\alpha t)^{-d/2} (\det(H^*))^{-1/2}. \tag{28}$$

where $H^* = \nabla\nabla F(\boldsymbol{\theta}^*)$ is the Hessian at the global minimum located at $\boldsymbol{\theta}^*$, and $\sim$ indicates that the ratio of both sides of the equation tend to 1 as $\alpha t \to \infty$. This shows that

$$\bar{F}(t) \sim \tfrac{1}{2} d (\alpha t)^{-1} \qquad \text{as} \quad \alpha t \to \infty \tag{29}$$

C.3 Non-interacting Case with Transportation and Birth-death

C.3.1 Proof of Theorem C.2

We first prove the following intermediate result

Lemma C.3 *Let $\delta > 0$ arbitrary, and define*

$$\phi_\delta(\boldsymbol{\theta}) = \max(0, 1 - \delta^{-1} F(\boldsymbol{\theta})) \ , \ f_\delta = \int_{\mathbb{R}^k} \phi_\delta(\boldsymbol{\theta}) \mu_0(d\boldsymbol{\theta}) \ .$$

Then

$$\forall t \ : \quad E(t) \leq \delta + \frac{1}{\alpha t f_\delta} \ . \tag{30}$$

Proof: By slightly abusing notation, we define

$$f_\delta(t) = \int_{\mathbb{R}^k} \phi_\delta(\boldsymbol{\theta}) \mu_t(d\boldsymbol{\theta}) \ .$$

We consider the following Lyapunov function:

$$\mathcal{L}_\delta(t) = \alpha t (E(t) - \delta) + \frac{1}{f_\delta(t)} \ . \tag{31}$$

Its time derivative is

$$\dot{\mathcal{L}}_\delta(t) = \alpha (E(t) - \delta) + \alpha t \dot{E}(t) - \frac{\dot{f}_\delta(t)}{f_\delta^2(t)} \ . \tag{32}$$

By definition, we have

$$\dot{E}(t) = - \int_{\mathbb{R}^k} |\nabla F(\boldsymbol{\theta})|^2 \mu_t(d\boldsymbol{\theta}) - \alpha \int_{\mathbb{R}^k} (F(\boldsymbol{\theta}) - F(t))^2 \mu_t(d\boldsymbol{\theta}) \leq 0 \ . \tag{33}$$

We also have

$$\begin{aligned}
\dot{f}_\delta(t) &= - \int_{\mathbb{R}^k} \langle \nabla \phi_\delta(\boldsymbol{\theta}), \nabla F(\boldsymbol{\theta}) \rangle \mu_t(d\boldsymbol{\theta}) - \alpha \int_{\mathbb{R}^k} \phi_\delta(\boldsymbol{\theta}) F(\boldsymbol{\theta}) \mu_t(d\theta) + \alpha E(t) f_\delta(t) \\
&\geq \delta^{-1} \int_{\mathbb{R}^k} |\nabla F(\boldsymbol{\theta})|^2 \mu_t(d\boldsymbol{\theta}) + \alpha (E(t) - \delta) f_\delta(t) \\
&\geq \alpha (E(t) - \delta) f_\delta(t) \ .
\end{aligned} \tag{34}$$

Observe that $0 \leq f_\delta(t) < 1$ because otherwise F would be flat (in which case the energy is 0). Also, we can assume wlog that $E(t) - \delta > 0$, since otherwise the statement of the lemma is trivially verified. By plugging (33) and (34) into (32) we have

$$\dot{\mathcal{L}}_\delta(t) \leq \alpha (E(t) - \delta) - \alpha (E(t) - \delta) f_\delta^{-1}(t) = \alpha (E(t) - \delta)(1 - f_\delta^{-1}(t)) \leq 0 \ . \tag{35}$$

Finally, since $f_\delta^{-1}(t) \geq 0$, we have

$$(E(t) - \delta) \leq \frac{\mathcal{L}_\delta(t)}{\alpha t} \leq \frac{\mathcal{L}_\delta(0)}{\alpha t} = \frac{1}{\alpha t f_\delta} \ ,$$

which concludes the proof of the Lemma. □

Proof of Theorem C.2: In order to prove (20), we apply the previous lemma for $\delta \to 0$. Let $\theta^* = \arg\min V(\theta)$, We have $F(\boldsymbol{\theta}^*) = 0$, and $\|\nabla \nabla F(\boldsymbol{\theta})\| \leq \beta$ for some $\beta > 0$. Then, for δ sufficiently small, the indicator function $\phi_\delta(\boldsymbol{\theta})$ is localized in the set

$$\left\{ \boldsymbol{\theta} \in \mathbb{R}^k ; \tfrac{1}{2} \langle (\boldsymbol{\theta} - \boldsymbol{\theta}^*), H^*(\boldsymbol{\theta} - \boldsymbol{\theta}^*) \leq \delta \right\} \supseteq \left\{ \boldsymbol{\theta} \in \mathbb{R}^d ; \|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2 \leq 2\beta^{-1}\delta \right\} \ .$$

where $H^* = \nabla\nabla F(\boldsymbol{\theta}^*)$. It follows that for sufficiently small δ,

$$\begin{aligned} f_\delta &= \int_{\mathbb{R}^k} \phi_\delta(\boldsymbol{\theta})\mu_0(d\boldsymbol{\theta}) \\ &\gtrsim \rho_0(\boldsymbol{\theta}^*) \int_{\|\boldsymbol{\theta}-\boldsymbol{\theta}^*\| \le \sqrt{2\beta^{-1}\delta}} \left(1 - \tfrac{1}{2}\delta^{-1}\langle(\boldsymbol{\theta}-\boldsymbol{\theta}^*), H^*(\boldsymbol{\theta}-\boldsymbol{\theta}^*)\rangle\right) d\boldsymbol{\theta} \\ &\sim \rho_0(\boldsymbol{\theta}^*) \left(2\beta^{-1}\delta\right)^{d/2} . \end{aligned} \tag{36}$$

By plugging (36) into (30) we obtain

$$\forall\, \delta, t > 0 \ : \qquad E(t) \le \delta + \frac{1}{\alpha t}\left(\frac{\beta}{2\delta}\right)^{d/2} \sim \delta + C\delta^{-d/2}t^{-1} ,$$

which implies that in order to reach an error ϵ, we need

$$t_\epsilon = O\left(\epsilon^{-(d+2)/2}\right) ,$$

which shows (20).

To obtain the asymptotic convergence rate in (21), note that by Lemma C.4 below the energy $\bar{F}(t) = \int_{\mathbb{R}^k} F(\boldsymbol{\theta})\rho_t(\boldsymbol{\theta})d\boldsymbol{\theta}$ can be written in terms of (43) as

$$\bar{F}(t) = \frac{\int_{\mathbb{R}^k} F(\boldsymbol{\theta}) \exp\left(\int_{-t}^0 (-\alpha F(\boldsymbol{\Theta}(s,\boldsymbol{\theta})) + \Delta F(\boldsymbol{\Theta}(s,\boldsymbol{\theta})))ds\right) \rho_0(\boldsymbol{\Theta}(-t,\boldsymbol{\theta}))d\boldsymbol{\theta}}{\int_{\mathbb{R}^k} \exp\left(\int_{-t}^0 (-\alpha F(\boldsymbol{\Theta}(s,\boldsymbol{\theta})) + \Delta F(\boldsymbol{\Theta}(s,\boldsymbol{\theta})))ds\right) \rho_0(\boldsymbol{\Theta}(-t,\boldsymbol{\theta}))d\boldsymbol{\theta}} \tag{37}$$

For large t, we can again use Laplace method to confirm that $\rho(t,\boldsymbol{\theta})$ concentrates near the absolute minimum of $F(\boldsymbol{\theta})$ located at $\boldsymbol{\theta}^*$. To see why notice that $\boldsymbol{\Theta}(t,\boldsymbol{\theta})$ converge, as $t \to \infty$, near local minima of F. Suppose that these minima are located at $\boldsymbol{\theta}_1^* = \boldsymbol{\theta}^*$, $\boldsymbol{\theta}_2^*$, etc. At these minima we have $\nabla F(\boldsymbol{\theta}_j^*) = 0$, and if in (45) we replace $F(\boldsymbol{\theta})$ by its quadratic approximation around any $\boldsymbol{\theta}_j^*$, $\frac{1}{2}\langle\boldsymbol{\theta}-\boldsymbol{\theta}_j^*, H_j^*(\boldsymbol{\theta}-\boldsymbol{\theta}_j^*)\rangle$ with $H_j^* = \nabla\nabla H(\boldsymbol{\theta}_j^*)$ positive definite, the solution to this equation reads

$$\boldsymbol{\Theta}^j_{\text{quad}}(t,\boldsymbol{\theta}) = \boldsymbol{\theta}_j^* + e^{-H^*t}(\boldsymbol{\theta}-\boldsymbol{\theta}_j^*) \tag{38}$$

from which we deduce

$$\int_{-t}^0 \Delta F(\boldsymbol{\Theta}^j_{\text{quad}}(s,\boldsymbol{\theta}))ds = \text{tr}(H_j^*)t, \tag{39}$$

and

$$\begin{aligned} -\alpha\int_{-t}^0 F(\boldsymbol{\Theta}^j_{\text{quad}}(s,\boldsymbol{\theta}))ds &= \alpha F(\boldsymbol{\theta}_j^*)t - \tfrac{1}{2}\alpha\int_{-t}^0 \langle\tilde{\boldsymbol{\theta}}_j, e^{-H^*s}H^*e^{-H^*s}\tilde{\boldsymbol{\theta}}_j\rangle \\ &= \alpha F(\boldsymbol{\theta}_j^*)t - \tfrac{1}{4}\alpha\langle\tilde{\boldsymbol{\theta}}_j, (e^{2H^*t} - \text{Id})\tilde{\boldsymbol{\theta}}_j\rangle. \end{aligned} \tag{40}$$

where $\tilde{\boldsymbol{\theta}}_j = \boldsymbol{\theta} - \boldsymbol{\theta}_j^*$. Since $F(\boldsymbol{\theta}_j^*) > 0$ except for the the global minimum $F(\boldsymbol{\theta}_1^*) = F(\boldsymbol{\theta}_1^*) = 0$, for large t, the only points that contribute to the integrals in (37) are those in a small region near $\boldsymbol{\theta}^*$ where we can replace $\boldsymbol{\Theta}(t,\boldsymbol{\theta})$ by $\boldsymbol{\Theta}^1_{\text{quad}}(t,\boldsymbol{\theta})$. As a result we can again neglect ρ_0 in these integrals, and evaluate them as if ρ_t was asymptotically the Gaussian density:

$$\rho_t(\boldsymbol{\theta}) \sim \mathcal{N}(\boldsymbol{\theta}^*, 2\alpha^{-1}e^{-2H^*t}). \tag{41}$$

This quantifies the late stages of the global convergence to the minimum and confirms the asymptotic decay rate in (21), thereby concluding the proof of Theorem C.2. □

Lemma C.4 *Denote by $\boldsymbol{\Theta}(t,\boldsymbol{\theta})$ the solution of the ODE*

$$\dot{\boldsymbol{\Theta}}(t,\boldsymbol{\theta}) = -\nabla F(\boldsymbol{\Theta}(t,\boldsymbol{\theta})), \qquad \boldsymbol{\Theta}(0,\boldsymbol{\theta}) = \boldsymbol{\theta} \tag{42}$$

Then under the conditions of Theorem C.2, the solution μ_t of the PDE (12) *has a density ρ_t given by*

$$\rho_t(\boldsymbol{\theta}) = \frac{\exp\left(\int_{-t}^0 G(\Theta(s,\boldsymbol{\theta}))ds\right)\rho_0(\boldsymbol{\Theta}(-t,\boldsymbol{\theta}))}{\int_D \exp\left(\int_{-t}^0 G(\Theta(s,\boldsymbol{\theta}'))ds\right)\rho_0(\boldsymbol{\Theta}(-t,\boldsymbol{\theta}'))d\boldsymbol{\theta}'} \tag{43}$$

where $G(\boldsymbol{\theta}) = \Delta F(\boldsymbol{\theta}) - \alpha F(\boldsymbol{\theta})$.

Proof: Since the initial μ_0 has a density $\rho_0 > 0$, so does μ_t for all $t > 0$ (but not in the limit as $t \to \infty$) and its density satisfies

$$\partial_t \rho_t = \nabla \cdot (\rho_t \nabla F(\boldsymbol{\theta})) - \alpha F(\boldsymbol{\theta})\rho_t + \alpha \bar{F}(t)\rho(t), \tag{44}$$

If $\boldsymbol{\Theta}(t,\boldsymbol{\theta})$ satisfies

$$\dot{\boldsymbol{\Theta}}(t,\boldsymbol{\theta}) = -\nabla F(\boldsymbol{\Theta}(t,\boldsymbol{\theta})) \qquad \boldsymbol{\Theta}(0,\boldsymbol{\theta}) = \boldsymbol{\theta}. \tag{45}$$

we have

$$\begin{aligned}\frac{d}{dt}\rho_t(\boldsymbol{\Theta}(t,\boldsymbol{\theta})) &= \partial_t\rho_t(\boldsymbol{\Theta}(t,\boldsymbol{\theta})) + \dot{\boldsymbol{\Theta}}(t,\boldsymbol{\theta}))\cdot\nabla\rho_t(\boldsymbol{\Theta}(t,\boldsymbol{\theta})) \\ &= \Delta F(\boldsymbol{\Theta}(t,\boldsymbol{\theta}))\rho(t,\boldsymbol{\Theta}(t,\boldsymbol{\theta})) - \left(F(\boldsymbol{\Theta}(t,\boldsymbol{\theta})) - \alpha\bar{F}(t)\right)\rho_t(\boldsymbol{\Theta}(t,\boldsymbol{\theta})).\end{aligned} \tag{46}$$

Therefore

$$\rho_t(\boldsymbol{\Theta}(t,\boldsymbol{\theta})) = \exp\left(\int_0^t (-\alpha F(\boldsymbol{\Theta}(s,\boldsymbol{\theta})) + \alpha\bar{F}(s) + \Delta F(\boldsymbol{\Theta}(s,\boldsymbol{\theta})))ds\right)\rho_0(\boldsymbol{\theta}). \tag{47}$$

By using $\boldsymbol{\Theta}(t,\boldsymbol{\Theta}(s,\boldsymbol{\theta})) = \boldsymbol{\Theta}(t+s,\boldsymbol{\theta})$ and the normalization condition, this implies

$$\rho_t(\boldsymbol{\theta}) = \frac{\exp\left(\int_{-t}^0(-\alpha F(\boldsymbol{\Theta}(s,\boldsymbol{\theta})) + \Delta F(\boldsymbol{\Theta}(s,\boldsymbol{\theta})))ds\right)\rho_0(\boldsymbol{\Theta}(-t,\boldsymbol{\theta}))}{\int_{\mathbb{R}^k}\exp\left(\int_{-t}^0(-\alpha F(\boldsymbol{\Theta}(s,\boldsymbol{\theta}')) + \Delta F(\boldsymbol{\Theta}(s,\boldsymbol{\theta}')))ds\right)\rho_0(\boldsymbol{\Theta}(-t,\boldsymbol{\theta}'))d\boldsymbol{\theta}'}. \tag{48}$$

This is (43) and terminates the proof of the lemma. □

D Derivation of (17)

Let μ_* be a minimizer and compare its energy to that of any other probability measure μ. Since the energy minimum is unique by convexity, we must have $\mathcal{E}[\mu] \geq \mathcal{E}[\mu_*]$. A direct calculation shows that

$$\begin{aligned}\mathcal{E}[\mu] = \mathcal{E}[\mu_*] &+ \int_D V(\boldsymbol{\theta},[\mu_*])(\mu(d\boldsymbol{\theta}) - \mu_*(d\boldsymbol{\theta})) \\ &+ \tfrac{1}{2}\int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')(\mu(d\boldsymbol{\theta}) - \mu_*(d\boldsymbol{\theta}))(\mu(d\boldsymbol{\theta}') - \mu_*(d\boldsymbol{\theta}'))\end{aligned} \tag{49}$$

The last term at the right hand side is always non-negative. Focusing on the second term its positivity requires that

$$\int_D V(\boldsymbol{\theta},[\mu_*])\mu(d\boldsymbol{\theta}) \geq \int_D V(\boldsymbol{\theta},[\mu_*])\mu_*(d\boldsymbol{\theta})) \equiv \bar{V}[\mu_*] \tag{50}$$

Since this equation must hold for any $\mu \in \mathcal{M}(D)$, we can specialize to Dirac distributions to deduce that the second equation in (17) must hold everywhere in D. In turns, this implies the first equation in (17) must hold as well.

E Proof of Theorem 4.3

We begin by noting that, if (14) holds for al $t > 0$, then $\bar{V}[\mu_t] = -\alpha^{-1}d\log C(t)/dt$ must be well-defined at all times. From (15), this derivative is given by

$$\bar{V}[\mu_t] = -\alpha^{-1}\frac{d}{dt}\log C(t) = \frac{\int_D V(\boldsymbol{\Theta}(t,\boldsymbol{\theta}),[\mu_t])e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})}{\int_D e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})} \tag{51}$$

Differentiating one more times gives

$$\begin{aligned}
\frac{d}{dt}\bar{V}[\mu_t] &= -\alpha\frac{\int_D |V(\boldsymbol{\Theta}(t,\boldsymbol{\theta}),[\mu_t])|^2 e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})}{\int_D e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})} \\
&\quad + \alpha\left(\frac{\int_D V(\boldsymbol{\Theta}(t,\boldsymbol{\theta}),[\mu_t]) e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})}{\int_D e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}}\right)^2 \\
&\quad + \frac{\int_D \partial_t V(\boldsymbol{\Theta}(t,\boldsymbol{\theta}),[\mu_t]) e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})}{\int_D e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})} \\
&= -\alpha\frac{\int_D |V(\boldsymbol{\Theta}(t,\boldsymbol{\theta}),[\mu_t])|^2 e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})}{\int_D e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})} \\
&\quad + \alpha\left(\frac{\int_D V(\boldsymbol{\Theta}(t,\boldsymbol{\theta}),[\mu_t]) e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})}{\int_D e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}}\right)^2 \\
&\quad + \frac{\int_D \dot{\boldsymbol{\Theta}}(t,\boldsymbol{\theta})\cdot\nabla V(\boldsymbol{\Theta}(t,\boldsymbol{\theta}),[\mu_t]) e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})}{\int_D e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})} \\
&\quad + \frac{\int_{D\times D} K(\boldsymbol{\Theta}(t,\boldsymbol{\theta}),\boldsymbol{\theta}')\partial_t\mu_t(d\boldsymbol{\theta}') e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})}{\int_D e^{-\alpha\int_0^t V(\boldsymbol{\Theta}(s,\boldsymbol{\theta}),[\mu_s])ds}\mu_0(d\boldsymbol{\theta})}
\end{aligned} \tag{52}$$

Using (15) to replace $\dot{\boldsymbol{\Theta}}(t,\boldsymbol{\theta})$ by $-\nabla V(\boldsymbol{\Theta}(t,\boldsymbol{\theta}),[\mu_t])$ and (14) to express these integral as expectations against μ_t gives

$$\begin{aligned}
\frac{d}{dt}\bar{V}[\mu_t] &= -\alpha\int_D |V(\boldsymbol{\theta},[\mu_t])|^2\mu_t(d\boldsymbol{\theta}) + \alpha\left(\int_D V(\boldsymbol{\theta},[\mu_t])\mu_t(d\boldsymbol{\theta})\right)^2 \\
&\quad - \int_D |\nabla V(\boldsymbol{\theta},[\mu_t])|^2\mu_t(d\boldsymbol{\theta}) - \int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')\partial_t\mu_t(d\boldsymbol{\theta}')\mu_t(d\boldsymbol{\theta}) \\
&= -\alpha\int_D \left(V(\boldsymbol{\theta},[\mu_t]) - \bar{V}[\mu_t]\right)^2\mu_t(d\boldsymbol{\theta}) - \int_D |\nabla V(\boldsymbol{\theta},[\mu_t])|^2\mu_t(d\boldsymbol{\theta}) \\
&\quad - \tfrac{1}{2}\frac{d}{dt}\int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')\mu_t(d\boldsymbol{\theta}')\mu_t(d\boldsymbol{\theta})
\end{aligned} \tag{53}$$

Therefore the terms at right hand side of (16) must be well-defined and we must also have

$$\int_D |V(\boldsymbol{\theta},[\mu_t])|^2\mu_t(d\boldsymbol{\theta}) < \infty, \qquad \int_D |\nabla V(\boldsymbol{\theta},[\mu_t])|^2\mu_t(d\boldsymbol{\theta}) < \infty \qquad \int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')\mu_t(d\boldsymbol{\theta}')\mu_t(d\boldsymbol{\theta}) < \infty \tag{54}$$

Since $\mu_t \rightharpoonup \mu_* \in \mathcal{M}(D)$ by assumption, we can take the limit as $t\to\infty$ to deduce that

$$\begin{aligned}
\lim_{t\to\infty}\int_D V(\boldsymbol{\theta},[\mu_t])\mu_t(d\boldsymbol{\theta}) &= \int_D V(\boldsymbol{\theta},[\mu_*])\mu_*(d\boldsymbol{\theta}) \\
\lim_{t\to\infty}\int_D |V(\boldsymbol{\theta},[\mu_t])|^2\mu_t(d\boldsymbol{\theta}) &= \int_D |V(\boldsymbol{\theta},[\mu_*])|^2\mu_*(d\boldsymbol{\theta}) \\
\lim_{t\to\infty}\int_D |\nabla V(\boldsymbol{\theta},[\mu_t])|^2\mu_t(d\boldsymbol{\theta}) &= \int_D |\nabla V(\boldsymbol{\theta},[\mu_*])|^2\mu_*(d\boldsymbol{\theta})
\end{aligned} \tag{55}$$

We will use these properties below, along with

$$V(\boldsymbol{\theta},[\mu_t]) \to V(\boldsymbol{\theta},[\mu_*]) \quad \text{and} \quad \int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')\mu_t(d\boldsymbol{\theta}') \to \int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')\mu_*(d\boldsymbol{\theta}') \qquad \text{pointwise in } D \tag{56}$$

which is require in order that both $\bar{V}[\mu_t]$ and $\mathcal{E}[\mu_t]$ be well-defined at all $t>0$ and in the limit as $t\to\infty$.

With these preliminaries, we now recall that the argument given after Theorem 4.3 implies that any fixed point μ_* of the PDE (13) must satisfy the first equation in (17). That is, we must have

$$V(\boldsymbol{\theta}, [\mu_*]) = \bar{V}[\mu_*] \qquad \forall \boldsymbol{\theta} \in \operatorname{supp} \mu_* \tag{57}$$

Therefore, to prove Theorem 4.3, it remains to show that the second equation in (17) must be satisfied as well. We will argue by contradiction: Let $D_* = \operatorname{supp} \mu_*$, assume $D_*^c \neq \emptyset$, and suppose that there exists a region $N \subseteq D_*^c$ where $V(\boldsymbol{\theta}, [\mu_*]) < \bar{V}[\mu_*]$. If it exists, this region must have nonzero Hausdorff measure in D since, by Assumption 4.2, $V(\boldsymbol{\theta}, [\mu_t]) \in C^2(D)$ for all $t \geq 0$ and $V(\boldsymbol{\theta}, [\mu_*]) \in C^2(D)$. $V(\boldsymbol{\theta}, [\mu_*]) - \bar{V}[\mu_*]$ must also reach a minimum value inside D even if D is open, for otherwise (15) would eventually carry mass towards infinity, which contradicts $\mu_t \rightharpoonup \mu_*$. This implies that, if we pick $\delta \in (0, \bar{V}[\mu_*] - \min_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, [\mu_*]))$ and let

$$N_\delta = \{\boldsymbol{\theta} : \delta \leq \bar{V}[\mu_*] - V(\boldsymbol{\theta}, [\mu_*])\} \subset N, \tag{58}$$

then N_δ is not empty. Since $V(\boldsymbol{\theta}, [\mu_*])$ is twice differentiable in $\boldsymbol{\theta}$, for δ close enough to $\bar{V}[\mu_*] - \min_{\boldsymbol{\theta}} V(\boldsymbol{\theta}, [\mu_*])$, N_δ is also compact and such that

$$\forall \boldsymbol{\theta} \in \partial N_\delta \ : \qquad |\nabla V(\boldsymbol{\theta}, [\mu_*])| > 0. \tag{59}$$

Given any solution μ_t of the PDE (13) that is supposed to converge to μ_* as $t \to \infty$, consider

$$f_\delta(t) = \mu_t(N_\delta) \tag{60}$$

Since μ_t is positive everywhere at any finite time, we must have $f_\delta(t) > 0$ for $t \in (0, \infty)$ However, since $\mu_t \to \mu_*$, we must also have

$$\lim_{t \to \infty} f_\delta(t) = 0. \tag{61}$$

From (13), $f_\delta(t)$ satisfies

$$\dot{f}_\delta(t) = \int_{\partial N_\delta} \hat{n} \cdot \nabla V d\sigma_t - \alpha \int_{N_\delta} (V - \bar{V}) d\mu_t \tag{62}$$

where $\hat{n}(\boldsymbol{\theta})$ is the inward pointing unit normal to ∂N_δ at $\boldsymbol{\theta}$ and σ_t is the probability measure on ∂N_δ obtained by restricting μ_t on this boundary: If $\phi_\epsilon \in C_c^\infty(D)$ is a sequence of test functions with $\operatorname{supp} \phi_\epsilon = N_\delta$ and converging towards the indicator set of N_δ as $\epsilon \to 0$, σ_t is defined as

$$\lim_{\epsilon \to 0} \int_{N_\delta} \nabla \phi_\epsilon(\boldsymbol{\theta}) \cdot \nabla V(\boldsymbol{\theta}, [\mu_t]) \mu_t(d\boldsymbol{\theta}) = \int_{\partial N_\delta} \hat{n}(\boldsymbol{\theta}) \cdot \nabla V(\boldsymbol{\theta}, [\mu_t]) d\sigma_t(\boldsymbol{\theta}) \tag{63}$$

Since

$$\lim_{t \to \infty} \hat{n}(\boldsymbol{\theta}) \cdot \nabla V(\boldsymbol{\theta}, [\mu_t]) = |\nabla V(\boldsymbol{\theta}, [\mu_*])| > 0, \tag{64}$$

there exists $t_+ > 0$ such that

$$\forall t > t_+ \ : \qquad \int_{\delta N_\delta} \hat{n} \cdot \nabla V d\nu_t > 0. \tag{65}$$

Restricting ourselves to $t > t_+$, we therefore have

$$\dot{f}_\delta(t) > -\alpha \int_{N_\delta} (V - \bar{V}) d\mu_t \tag{66}$$

Let us analyze the remaining integral in this equation. Denoting $\tilde{V}(\boldsymbol{\theta}, [\mu_t]) = V(\boldsymbol{\theta}, [\mu_t]) - \bar{V}[\mu_t]$, we have

$$\begin{aligned} -\alpha \int_{N_\delta} \tilde{V}(\boldsymbol{\theta}, [\mu_t]) \mu_t(d\boldsymbol{\theta}) &= -\alpha \int_{N_\delta} \tilde{V}(\boldsymbol{\theta}, [\mu_*]) \mu_t(d\boldsymbol{\theta}) \\ &\quad - \alpha \int_{N_\delta} \left(\tilde{V}(\boldsymbol{\theta}, [\mu_t]) - \tilde{V}(\boldsymbol{\theta}, [\mu_*]) \right) \mu_t(d\boldsymbol{\theta}) \\ &\geq \alpha \delta f_\delta(t) - \alpha \int_{N_\delta} \left(\tilde{V}(\boldsymbol{\theta}, [\mu_t]) - \tilde{V}(\boldsymbol{\theta}, [\mu_*]) \right) \mu_t(d\boldsymbol{\theta}) \end{aligned} \tag{67}$$

where we used the definition of N_δ. Looking at the last term, we can assess its magnitude using

$$\begin{aligned}
&\left|\int_{N_\delta}\left(\tilde{V}(\boldsymbol{\theta},[\mu_t])-\tilde{V}(\boldsymbol{\theta},[\mu_*])\right)\mu_t(d\boldsymbol{\theta})\right| \\
&\leq \tfrac{1}{2}\int_{N_\delta}\left|\int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')\left(\mu_t(d\boldsymbol{\theta}')-\mu_*(d\boldsymbol{\theta}')\right)\right|\mu_t(d\boldsymbol{\theta})+\left|\bar{V}[\mu_t]-\bar{V}(\mu_*)\right| f_\delta(t) \\
&\leq M(t)f_\delta(t)
\end{aligned} \tag{68}$$

where (using the compactness of N_δ)

$$M(t)=\max_{N_\delta}\left|\int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')(\mu_t(d\boldsymbol{\theta}')-\mu_*(d\boldsymbol{\theta}'))\right|+|\bar{V}[\mu_t]-\bar{V}(\mu_*)|<\infty \tag{69}$$

Summarizing, we have deduced that

$$\dot{f}_\delta(t)>\alpha\delta f_\delta(t)+R(t) \tag{70}$$

with

$$|R(t)|\leq M(t)f_\delta(t) \tag{71}$$

Since we work under the assumption that $\mu_t \rightharpoonup \mu_*$, $M(t)$ must tend to 0 as $t\to\infty$. As a result, $\exists t_\delta>0$ such $\forall t>t_\delta$ we have $N(t)<\delta$, which, from (70), implies that $\forall t>\max(t_+,t_\delta)$ we have $\dot{f}_\delta(t)>0$, a contradiction with (61). Therefore the only fixed points accessible by the PDE (13) are those for which both equations in (17) hold, which proves the theorem.

F Proof of Theorem 4.4

Let $\mu_*=\lim_{t\to\infty}\mu_t$ be the stationary point reached by the solution of (13) and denote $E(t)=\mathcal{E}[\mu_t]-\mathcal{E}[\mu_*]\geq 0$. Then

$$\begin{aligned}
\frac{d}{dt}E^{-1} &= -E^{-2}\int_D V\partial_t\mu_t \\
&= E^{-2}\int_D\left(|\nabla V|^2+\alpha|V-\bar{V}|^2\right)d\mu_t \\
&\geq \alpha E^{-2}\int_D|V-\bar{V}|^2 d\mu_t
\end{aligned} \tag{72}$$

where we used $\int_D V^2 d\mu_t-\bar{V}^2=\int_D|V-\bar{V}|^2 d\mu_t$. By convexity

$$\begin{aligned}
\mathcal{E}[\mu_*] &\geq \mathcal{E}[\mu]-\int_D V(d\mu-d\mu_*) \\
&= \mathcal{E}[\mu]-\bar{V}+\int_D V d\mu_* \\
&= \mathcal{E}[\mu]+\int_D(V-\bar{V})d\mu_*
\end{aligned} \tag{73}$$

As a result

$$0\leq E\leq\int_D(\bar{V}-V)d\mu_* \tag{74}$$

and hence

$$0\leq E^2\leq\left|\int_D(V-\bar{V})d\mu_*\right|^2 \tag{75}$$

Using this inequality in (72) gives

$$\frac{d}{dt}E^{-1}\geq\alpha\frac{\int_D|V-\bar{V}|^2 d\mu_t}{\left|\int_D(V-\bar{V})d\mu_*\right|^2} \tag{76}$$

In Lemma F.1 below we show that $\exists t_+ > 0$ such that

$$\forall t > t_+ \ : \ \frac{\int_D |V - \bar{V}|^2 d\mu_t}{\left|\int_D (V - \bar{V}) d\mu_*\right|^2} \geq C > 0 \tag{77}$$

As a result, $dE^{-1}/dt \geq \alpha$ for $t > t_+$. Integrating this relation in time on $[t_0, t]$ with $t_+ < t_0 \leq t$ gives

$$E^{-1}(t) \geq E^{-1}(t) - E^{-1}(t_0) \geq \alpha C(t - t_0) \tag{78}$$

and hence

$$\lim_{t\to\infty} tE(t) \leq (\alpha C)^{-1} \tag{79}$$

which proves the theorem. □

Note that the proof only takes into account the effects of birth-death terms; adding transport may accelerate the rate.

Lemma F.1 *There exist $t_+ > 0$ such that (77) holds.*

Proof: Let $\nu_t = \mu_t - \mu_*$ and for future reference note that ν_t is a signed measure on $D_* = \operatorname{supp}\mu_*$ but $\nu_t \geq 0$ on D_c^*. Denote

$$V = V(\boldsymbol{\theta}, [\mu_t]), \qquad \bar{V} = \int_D V(\boldsymbol{\theta}, [\mu_t]) d\mu_t, \qquad V_* = V(\boldsymbol{\theta}, [\mu_*]), \qquad \bar{V}_* = \int_D V(\boldsymbol{\theta}, [\mu_*]) d\mu_* \tag{80}$$

We have

$$\begin{aligned} V &= F(\boldsymbol{\theta}) + \int_D K(\boldsymbol{\theta}, \boldsymbol{\theta}')(\mu_*(d\boldsymbol{\theta}') + \nu_t(d\boldsymbol{\theta})') \\ &= V_* + \int_D K(\boldsymbol{\theta}, \boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}') \end{aligned} \tag{81}$$

and hence

$$\int_D V d\mu_* = \bar{V}_* + \int_{D\times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}')\mu_*(d\boldsymbol{\theta}) \tag{82}$$

Recall that $V_* = \bar{V}_*$ on $\operatorname{supp}\mu_*$. As a result

$$\begin{aligned} \bar{V} &= \int_D F(\boldsymbol{\theta})(\mu_*(d\boldsymbol{\theta}) + \nu_t(d\boldsymbol{\theta})) + \int_{D\times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')(\mu_*(d\boldsymbol{\theta}) + \nu_t(d\boldsymbol{\theta}))(\mu_*(d\boldsymbol{\theta}') + \nu_t(d\boldsymbol{\theta}')) \\ &= \bar{V}_* + \int_D F(\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}) + 2\int_{D\times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\mu_*(d\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}') + \int_{D\times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}') \end{aligned} \tag{83}$$

We can combine these two equations to obtain

$$\begin{aligned} \int_D (\bar{V} - V) d\mu_* &= \int_D F(\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}) + \int_{D\times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\mu_*(d\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}') + \int_{D\times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}') \\ &= \int_D V_* d\nu_t + \int_{D\times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}') \\ &= \int_D (V_\star - \bar{V}_*) d\nu_t + \int_{D\times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}') \\ &= \int_{D_*^c} (V_\star - \bar{V}_*) d\nu_t + \int_{D\times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}') \end{aligned} \tag{84}$$

where we used $\int_D \bar{V}_* d\nu_t = \bar{V}_* \int_D (d\mu_t - d\mu_*) = 0$ to get the penultimate equality and $V_\star - \bar{V}_* = 0$ on D_* to get the last.

Proceeding similarly using again $V_* = \bar{V}_*$ on $\operatorname{supp}\mu_*$ as well as $\int_D d\nu_t = \int_D (d\mu_t - d\mu_*) = 0$, we can also obtain

$$\int_D |V - \bar{V}|^2 d\mu_* = \int_D \left(\int_D K(\boldsymbol{\theta}, \boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}')\right)^2 \mu_*(d\theta) + R^2 - 2R\int_{D\times D} K(\boldsymbol{\theta}, \boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}')\mu_*(d\boldsymbol{\theta}) \tag{85}$$

and

$$\begin{aligned}\int_D |V-\bar{V}|^2 d\mu_t &= \int_{D_*^c} |V_*-\bar{V}_*|^2 d\nu_t + \int_D \left(\int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}')\right)^2 (\mu_*(d\theta)+\nu_t(d\boldsymbol{\theta})) + R^2 \\ &\quad - 2R\int_{D_*^c}(V_*-\bar{V}_*)d\nu_t - 2R\int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}')(\mu_*(d\boldsymbol{\theta})+\nu_t(d\boldsymbol{\theta})) \\ &\quad + 2\int_{D_*^c\times D}(V_*-\bar{V}_*)\nu_t(d\boldsymbol{\theta})K(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}')\end{aligned} \tag{86}$$

where we denote

$$\begin{aligned}R &= \bar{V}-\bar{V}_* \\ &= \int_D F(\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}) + 2\int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')\mu_*(d\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}') + \int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}')\end{aligned} \tag{87}$$

Let us now compare the square of (84) to (86). Since $V_*-\bar{V}_*\geq 0$ and $\nu_t\geq 0$ on D_*^c, we have

$$\int_{D_*^c}(V_\star-\bar{V}_*)d\nu_t \geq 0. \tag{88}$$

We distinguish two cases:

Case 1: $\int_{D_*^c}(V_*-\bar{V}_*)d\nu_t>0$ (which requires $D_*^c\neq\emptyset$). Since $\nu_t \rightharpoonup 0$ as $t\to\infty$ the last term in (82) is higher order. As a result, for any $\delta>0$, $\exists t_1>0$ such that

$$\forall t>t_1 \ : \ \int_D(\bar{V}-V)d\mu_* \leq (1+\delta)\int_{D_*^c}(V_\star-\bar{V}_*)d\nu_t \tag{89}$$

which also implies that (using again $\nu_t\geq 0$ on D_*^c)

$$\begin{aligned}\forall t>t_1 \ : \left|\int_D (V-\bar{V})d\mu_*\right|^2 &\leq (1+\delta)^2\left|\int_{D^*}(V_*-\bar{V}_*)d\nu_t\right|^2 \\ &\leq (1+\delta)^2\nu_t(D_*^c)\int_{D^*}|V_\star-\bar{V}_*|^2 d\nu_t\end{aligned} \tag{90}$$

Similarly, the first term at the right hand side of (86) dominates all the other ones as $t\to\infty$ in the sense that, for any $\delta>0$, $\exists t_2>0$ such that

$$\forall t>t_2 \ : \ \int_D |V-\bar{V}|^2 d\mu_t \geq (1-\delta)\int_{D_*^c}|V_*-\bar{V}_*|^2 d\nu_t \tag{91}$$

Taken together, (90) and (91) imply the statement of the lemma with any $C>0$ (since $\nu_t(D_*^c)\to 0$ as $t\to\infty$). As a result $\lim_{t\to\infty} tE(t)=0$ in this case since $\int_D|V-\bar{V}|^2 d\mu_t/|\int_D(V-\bar{V})d\mu_*|^2\to\infty$.

Case 2: $\int_{D_*^c}(V_*-\bar{V}_*)d\nu_t=0$ (i.e. $D_*^c=\emptyset$ or $V_*=\bar{V}_*$ on D_*^c as well as D_*). In this case it is easier to use (85) via the inequality

$$\left|\int_D(V-\bar{V})d\mu_*\right|^2 \leq \int_D|V-\bar{V}|^2 d\mu_* \tag{92}$$

We also have that (86) reduces to

$$\begin{aligned}\int_D |V-\bar{V}|^2 d\mu_t &= \int_D\left(\int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}')\right)^2(\mu_*(d\theta)+\nu_t(d\boldsymbol{\theta}))+R^2 \\ &\quad - 2R\int_{D\times D}K(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}')(\mu_*(d\boldsymbol{\theta})+\nu_t(d\boldsymbol{\theta})) \\ &= \int_D\left(\int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}')\right)^2(\mu_*(d\boldsymbol{\theta})+\nu_t(d\boldsymbol{\theta}))-R^2\end{aligned} \tag{93}$$

where we use the fact that R reduces to (using $V_* = \bar{V}_*$ and $\int_D V_* d\nu_t = \bar{V}_* \int_D (d\mu_t - d\mu_*) = 0$)

$$\begin{aligned} R &= \int_D V_* d\nu_t + \int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')\mu_*(d\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}') + \int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta})\nu_t(d\boldsymbol{\theta}') \\ &= \int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta})(\mu_*(d\boldsymbol{\theta}') + \nu_t(d\boldsymbol{\theta}')) \end{aligned} \tag{94}$$

Since $\int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}') \neq 0$ on D_*, the leading order terms in $\int_D |V - \bar{V}|^2 d\mu_*$ and $\int_D |V - \bar{V}|^2 d\mu_t$ are the same and given by

$$A = \int_D \left(\int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}') \right)^2 d\mu_* - \left(\int_{D\times D} K(\boldsymbol{\theta},\boldsymbol{\theta}')\nu_t(d\boldsymbol{\theta}')\mu_*(d\boldsymbol{\theta}) \right)^2 > 0 \tag{95}$$

That is, for any $\delta > 0$, $\exists t_3 > 0$ such that

$$\forall t > t_3 \ : \ \int_D |V - \bar{V}|^2 d\mu_* \leq (1+\delta)A, \qquad \int_D |V - \bar{V}|^2 d\mu_t \geq (1-\delta)A \tag{96}$$

Together with (92), this implies the statement of the lemma with $C = 1$. □

G Proof of Propositions 5.1 and 5.2

Here we give formal proofs Propositions 5.1 and 5.2 using tools from the theory of measure-valued Markov processes [Daw06].

To begin, recall that the evolution of $\mu_t^{(n)} = n^{-1}\sum_{i=1}^n \delta_{\boldsymbol{\theta}_i(t)}$ is Markovian since that of the particles $\boldsymbol{\theta}_i(t)$ is and these particles are interchangeable. To study this measure-valued Markov process and in particular analyze its properties when $n \to \infty$, it is useful to write its infinitesimal generator, i.e. the operator whose action on a functional $\Phi : \mathcal{M}(\mathbb{R}^k) \to \mathbb{R}$ evaluated on $\mu^{(n)}$ is defined via

$$(\mathcal{L}_n \Phi)[\mu^{(n)}] = \lim_{t\to 0+} t^{-1} \left(\mathbb{E}^{\mu_0^{(n)} = \mu^{(n)}} \Phi[\mu_t^{(n)}] - \Phi[\mu^{(n)}] \right) \tag{97}$$

where $\mathbb{E}^{\mu_0^{(n)} = \mu^{(n)}}$ denotes the expectation along the trajectory $\mu_t^{(n)}$ taken conditional on $\mu_0^{(n)} = \mu^{(n)}$ for some given $\mu^{(n)}$. To compute the limit in (97), notice that if particle $\boldsymbol{\theta}_i(t)$ gets killed at time t and particle $\boldsymbol{\theta}_j(t)$ gets duplicated, the changes this induces on $\mu_t^{(n)}$ is

$$\mu_t^{(n)} = \mu_{t-}^{(n)} + n^{-1}\left(\delta_{\boldsymbol{\theta}_j} - \delta_{\boldsymbol{\theta}_i}\right). \tag{98}$$

where $\mu_{t-}^{(n)} = \lim_{\epsilon\to 0+} \mu_{t-\epsilon}^{(n)}$ Similarly if particle $\boldsymbol{\theta}_i(t)$ gets duplicated at time t and particle $\boldsymbol{\theta}_j(t)$ gets killed, the change this induces on $\mu_t^{(n)}$ is

$$\mu_t^{(n)} = \mu_{t-}^{(n)} - n^{-1}\left(\delta_{\boldsymbol{\theta}_j} - \delta_{\boldsymbol{\theta}_i}\right). \tag{99}$$

A particle swap occurs with rates dictated by $\tilde{V}$, so we define

$$\mu_t^{(n)}\{\boldsymbol{\theta} \leftrightarrow \boldsymbol{\theta}'\} = \mu_t^{(n)} + n^{-1}\sigma(\boldsymbol{\theta})\left(\delta_{\boldsymbol{\theta}} - \delta_{\boldsymbol{\theta}'}\right) \tag{100}$$

where $\sigma(\boldsymbol{\theta}_i) = \operatorname{sign}\tilde{V}(\boldsymbol{\theta}_i)$ determines the direction of the swap. If we account for the rate at which these events occur, as well as the effect of transport by GD, we can explicitly compute the generator defined in (97) and arrive at the expression

$$\begin{aligned} (\mathcal{L}_n\Phi)[\mu^{(n)}] = &-\frac{1}{n}\sum_{i=1}^n \int_D \nabla V(\boldsymbol{\theta}_i, [\mu^{(n)}])\delta_{\boldsymbol{\theta}_i}(d\boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}_i} D_{\mu^{(n)}}\Phi(\boldsymbol{\theta}_i) \\ &+ \frac{\alpha}{n}\sum_{i,j=1}^n \int_{D\times D} |\tilde{V}(\boldsymbol{\theta}_i)|\delta_{\boldsymbol{\theta}_i}(d\boldsymbol{\theta})\delta_{\boldsymbol{\theta}_j}(d\boldsymbol{\theta}') \left(\Phi[\mu_t^{(n)}\{\boldsymbol{\theta}_i \leftrightarrow \boldsymbol{\theta}_j\}] - \Phi[\mu^{(n)}] \right) \end{aligned} \tag{101}$$

where the functional derivative $D_\mu \Phi$ is the function from D to $\mathbb{R}$ defined via: for any $\omega \in \mathcal{M}_s(D)$, the space of signed distributions such that $\int_D \omega(d\boldsymbol{\theta}) = 0$,

$$\lim_{\varepsilon \to 0} \varepsilon^{-1} \left(\Phi[\mu + \epsilon \omega] - \Phi[\mu] \right) = \int_D D_\mu \Phi(\boldsymbol{\theta}) \omega(d\boldsymbol{\theta}) \tag{102}$$

We can use the properties of the Dirac distribution to rewrite the generator in (101) as

$$\begin{aligned} (\mathcal{L}_n \Phi)[\mu^{(n)}] = &- \int_D \nabla V(\boldsymbol{\theta}, [\mu^{(n)}]) \mu^{(n)}(d\boldsymbol{\theta}) \cdot \nabla D_{\mu^{(n)}} \Phi(\boldsymbol{\theta}) \\ &+ n\alpha \int_{D \times D} |\tilde{V}(\boldsymbol{\theta}, [\mu^{(n)}])| \mu^{(n)}(d\boldsymbol{\theta}) \mu^{(n)}(d\boldsymbol{\theta}') \left(\Phi[\mu_t^{(n)}\{\boldsymbol{\theta} \leftrightarrow \boldsymbol{\theta}'\}] - \Phi[\mu^{(n)}] \right) \end{aligned} \tag{103}$$

and σ in (100) is evaluated on

$$\tilde{V}(\boldsymbol{\theta}, [\mu]) = F(\boldsymbol{\theta}) + \int_D K(\boldsymbol{\theta}, \boldsymbol{\theta}') \mu(d\boldsymbol{\theta}') - \int_D \left(F(\boldsymbol{\theta}') + \int_D K(\boldsymbol{\theta}', \boldsymbol{\theta}'') \mu(d\boldsymbol{\theta}'') \right) \mu(d\boldsymbol{\theta}'). \tag{104}$$

The operator in (103) is now defined for any $\mu \in \mathcal{M}(D)$, and we will use it in this form in our developments below.

The generator (103) can be used to write an evolution equation for the expectation of functionals evaluated on $\mu_t^{(n)}$. That is, if we define

$$\Phi_t[\mu^n] = \mathbb{E}^{\mu_0^{(n)} = \mu^{(n)}} \Phi[\mu_t^n] \tag{105}$$

then this time-dependent functional satisfies the backward Kolmogorov equation (BKE)

$$\partial_t \Phi_t[\mu^n] = (\mathcal{L}_n \Phi_t)[\mu^{(n)}], \qquad \Phi_{t=0}[\mu^n] = \Phi[\mu^n]. \tag{106}$$

The proof of Proposition 5.1 is based on analyzing the properties of this equation in the limit as $n \to \infty$, which we expand upon in Appendix G.1. The proof of Proposition 5.2 is based on writing a similar equation for an extended process in which we magnify the dynamics of $\mu_t^{(n)}$ around its limit, as shown in Appendix G.2.

G.1 Proof of Proposition 5.1

If we take the limit of $(\mathcal{L}_n \Phi)[\mu^{(n)}]$ as $n \to \infty$ on a sequence such that $\mu^{(n)} \rightharpoonup \mu$, we deduce that $(\mathcal{L}_n \Phi)[\mu^{(n)}] \to (\mathcal{L}\Phi)[\mu]$ with

$$(\mathcal{L}\Phi)[\mu] = - \int_D \nabla V(\boldsymbol{\theta}, [\mu]) \mu(d\boldsymbol{\theta}) \cdot \nabla_{\boldsymbol{\theta}} D_\mu \Phi(\boldsymbol{\theta}) - \alpha \int_D \tilde{V}(\boldsymbol{\theta}, [\mu]) \mu(d\boldsymbol{\theta}) D_\mu \Phi(\boldsymbol{\theta}) \tag{107}$$

Correspondingly, in this limit the BKE (106) becomes

$$\partial_t \Phi_t[\mu] = (\mathcal{L}\Phi_t)[\mu], \qquad \Phi_{t=0}[\mu] = \Phi[\mu]. \tag{108}$$

Since (107) is precisely the generator of process defined by the PDE (13), this shows that, if $\mu_{t=0}^{(n)} = \mu^{(n)} \rightharpoonup \mu$ as $n \to \infty$, then

$$\lim_{n \to \infty} \Phi_t[\mu^{(n)}] = \Phi_t[\mu] \quad \Leftrightarrow \quad \lim_{n \to \infty} \mathbb{E}^{\mu_0^{(n)} = \mu^{(n)}} \Phi[\mu_t^n] = \Phi[\mu_t] \tag{109}$$

where μ_t solves the PDE (13) for the initial condition $\mu_{t=0} = \mu$. This proves the weak version of the LLN stated in Proposition 5.1.

G.2 Proof of Proposition 5.2

To quantify the fluctuations around the LLN, let μ_t be the limit of $\mu_t^{(n)}$ (i.e. the solution to the PDE (13)) and define

$$\omega_t^{(n)} = \sqrt{n} \left(\mu_t^{(n)} - \mu_t \right) \in \mathcal{M}_s(D) \tag{110}$$

We can write down the generator of the joint process $(\mu_t, \omega_t^{(n)})$. To do so, we consider its action on a functional, $\hat{\Phi} : \mathcal{M}(D) \times \mathcal{M}_s(D) \to \mathbb{R}$ is given by (using $\mu^{(n)} = \mu + n^{-1/2}\omega^{(n)}$)

$$
\begin{aligned}
&(\hat{\mathcal{L}}_n\hat{\Phi})[\mu,\omega^{(n)}] \\
&= n^{1/2}\int_D \nabla V(\boldsymbol{\theta},[\mu+n^{-1/2}\omega^{(n)}]))\left(\mu(d\boldsymbol{\theta})+n^{-1/2}\omega^{(n)}(d\boldsymbol{\theta})\right)\cdot\nabla D_{\omega^{(n)}}\hat{\Phi}(\boldsymbol{\theta}) \\
&+ n\alpha\int_{D\times D}\sigma(\boldsymbol{\theta},[\mu+n^{-1/2}\omega^{(n)}])\tilde{V}(\boldsymbol{\theta},[\mu+n^{-1/2}\omega^{(n)}])\left(\mu(d\boldsymbol{\theta})+n^{-1/2}\omega^{(n)}(d\boldsymbol{\theta})\right)\left(\mu(d\boldsymbol{\theta}')+n^{-1/2}\omega^{(n)}(d\boldsymbol{\theta}')\right) \\
&\qquad\times\left(\hat{\Phi}[\mu,\omega^{(n)}+n^{-1/2}\sigma(\boldsymbol{\theta},[\mu+n^{-1/2}\omega^{(n)}])(\delta_{\boldsymbol{\theta}'}-\delta_{\boldsymbol{\theta}})]-\hat{\Phi}[\mu,\omega^{(n)}]\right) \\
&-\int_D \nabla V(\boldsymbol{\theta},[\mu]))\mu(d\boldsymbol{\theta})\cdot\nabla_{\boldsymbol{\theta}}\left(D_\mu\hat{\Phi}(\boldsymbol{\theta})-n^{1/2}D_{\omega^{(n)}}\hat{\Phi}(\boldsymbol{\theta})\right) \\
&-\alpha\int_D \tilde{V}(\boldsymbol{\theta},[\mu])\mu(d\boldsymbol{\theta})\left(D_\mu\hat{\Phi}(\boldsymbol{\theta})-n^{1/2}D_{\omega^{(n)}}\hat{\Phi}(\boldsymbol{\theta})\right).
\end{aligned}
\tag{111}
$$

Proceeding similarly as we did to derive (107), we can take the limit of $(\hat{\mathcal{L}}_n\hat{\Phi})[\mu,\omega^{(n)}]$ as $n\to\infty$ on a sequence such that $\omega^{(n)} \rightharpoonup \omega \in \mathcal{M}_s(D)$. A direct calculation using $\int_D \tilde{V}(\boldsymbol{\theta},[\mu])d\mu = 0$, $\int_D d\omega = 0$, and $\int_D \tilde{V}(\boldsymbol{\theta},[\mu])d\mu = 1$ indicates that $(\hat{\mathcal{L}}_n\hat{\Phi})[\mu,\omega^{(n)}] \to (\hat{\mathcal{L}}\hat{\Phi})[\mu,\omega]$ with

$$
\begin{aligned}
(\hat{\mathcal{L}}\hat{\Phi})[\mu,\omega] = &-\int_D \nabla V(\boldsymbol{\theta},[\mu])\omega(d\boldsymbol{\theta})\cdot\nabla D_\omega\Phi(\boldsymbol{\theta}) - \int_{D\times D}\nabla K(\boldsymbol{\theta},\boldsymbol{\theta}')\omega(d\boldsymbol{\theta}')\mu(d\boldsymbol{\theta})\cdot\nabla D_\omega\Phi(\boldsymbol{\theta}) \\
&-\alpha\int_D \tilde{V}(\boldsymbol{\theta},[\mu])\omega(d\boldsymbol{\theta})D_\omega\hat{\Phi}(\boldsymbol{\theta}) - \alpha\int_{D\times D}K(\boldsymbol{\theta},\boldsymbol{\theta}')\omega(d\boldsymbol{\theta}')\mu(d\boldsymbol{\theta})D_\omega\hat{\Phi}(\boldsymbol{\theta}) \\
&+\alpha\int_{D\times D}\tilde{V}(\boldsymbol{\theta}',[\mu])\omega(d\boldsymbol{\theta}')\mu(d\boldsymbol{\theta})D_\omega\hat{\Phi}(\boldsymbol{\theta}) \\
&+\alpha\int_{D\times D\times D}K(\boldsymbol{\theta}',\boldsymbol{\theta}'')\omega(d\boldsymbol{\theta}')\mu(d\boldsymbol{\theta}'')\mu(d\boldsymbol{\theta})D_\omega\hat{\Phi}(\boldsymbol{\theta}) \\
&+\alpha\int_{D\times D}|\tilde{V}(\boldsymbol{\theta},[\mu])|\mu(d\boldsymbol{\theta})\mu(d\boldsymbol{\theta}')\left(D_\omega^2\hat{\Phi}(\boldsymbol{\theta},\boldsymbol{\theta})+D_\omega^2\hat{\Phi}(\boldsymbol{\theta}',\boldsymbol{\theta}')-2D_\omega^2\hat{\Phi}(\boldsymbol{\theta},\boldsymbol{\theta}')\right) \\
&-\int_D \nabla V(\boldsymbol{\theta},[\mu])\mu(d\boldsymbol{\theta})\cdot\nabla_{\boldsymbol{\theta}}D_\mu\Phi(\boldsymbol{\theta}) - \alpha\int_D\tilde{V}(\boldsymbol{\theta},[\mu])\mu(d\boldsymbol{\theta})D_\mu\Phi(\boldsymbol{\theta})
\end{aligned}
\tag{112}
$$

where the second order functional derivative $D_\mu^2\hat{\Phi}$ is the function from $D\times D$ to $\mathbb{R}$ defined via: for any $\nu,\nu' \in \mathcal{M}_s(D)$,

$$
\begin{aligned}
&\lim_{\varepsilon,\varepsilon'\to 0}(\varepsilon\varepsilon')^{-1}\left(\hat{\Phi}[\mu+\varepsilon\nu+\varepsilon'\nu',\omega]-\hat{\Phi}[\mu+\varepsilon\nu,\omega]-\hat{\Phi}[\mu+\varepsilon'\nu',\omega]+\hat{\Phi}[\mu,\omega]\right) \\
&\qquad=\int_{D\times D}D_\mu^2\hat{\Phi}(\boldsymbol{\theta},\boldsymbol{\theta}')\nu(d\boldsymbol{\theta})\nu(d\boldsymbol{\theta}'),
\end{aligned}
\tag{113}
$$

and similarly for $D_\omega^2\hat{\Phi}$. The operator in μ in (111) is the same as in (107), confirming the LLN; the operator in ω is a second order operator, i.e. it is the generator of a stochastic differential equation. That is, we have established that, as $n\to\infty$,

$$
\omega_t^{(n)} \equiv \sqrt{n}\left(\mu_t^{(n)}-\mu_t\right) \rightharpoonup \omega_t \qquad \text{in law} \tag{114}
$$

where $\omega_t(d\boldsymbol{\theta})$ is Gaussian random distribution whose equation can be obtained from the generator in (112) Formally

$$
\begin{aligned}
\partial_t\omega_t = &\nabla\cdot\left(\nabla V(\boldsymbol{\theta},[\mu_t])\omega_t+\int_D\nabla K(\boldsymbol{\theta},\boldsymbol{\theta}')\omega_t(d\boldsymbol{\theta}')\mu_t\right) \\
&-\alpha\tilde{V}(\boldsymbol{\theta},[\mu_t])\omega_t-\alpha\int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')\omega_t(d\boldsymbol{\theta}')\mu_t \\
&+\alpha(\int_{D\times D}K(\boldsymbol{\theta}',\boldsymbol{\theta}'')\mu_t(d\boldsymbol{\theta}')\omega_t(d\boldsymbol{\theta}''))\mu_t+\sqrt{2}\eta(t),
\end{aligned}
\tag{115}
$$

where $\eta(t)$ is a white-noise term with covariance consistent with (112):

$$\begin{aligned}\mathbb{E}\eta(t)\eta(t') &= \alpha|\tilde{V}(\boldsymbol{\theta},[\mu_t])|\mu_t(d\boldsymbol{\theta})\delta_{\boldsymbol{\theta}}(d\boldsymbol{\theta}')\delta(t-t') \\ &\quad - \alpha\left(|\tilde{V}(\boldsymbol{\theta},[\mu_t])| + |\tilde{V}(\boldsymbol{\theta}',[\mu_t])|\right)\mu_t(d\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta}')\delta(t-t')\end{aligned} \tag{116}$$

Since ω_t is Gaussian with zero mean, all its information is contained in its covariance $\Sigma_t(d\boldsymbol{\theta}, d\boldsymbol{\theta}') = \mathbb{E}\omega_t(d\boldsymbol{\theta})\omega_t(d\boldsymbol{\theta}')$, for which we can derive the equation

$$\begin{aligned}\partial_t\Sigma_t &= \nabla_{\boldsymbol{\theta}}\cdot\left(\nabla V(\boldsymbol{\theta},[\mu_t])\Sigma_t + \int_D \nabla K(\boldsymbol{\theta},\boldsymbol{\theta}'')\Sigma_t(d\boldsymbol{\theta},d\boldsymbol{\theta}'')\mu_t(d\boldsymbol{\theta})\right) \\ &\quad + \nabla_{\boldsymbol{\theta}'}\cdot\left(\nabla V(\boldsymbol{\theta}',[\mu_t])\Sigma_t + \int_D \nabla K(\boldsymbol{\theta}',\boldsymbol{\theta}'')\Sigma_t(d\boldsymbol{\theta}',d\boldsymbol{\theta}'')\mu_t(d\boldsymbol{\theta}')\right) \\ &\quad - \alpha\left(\tilde{V}(\boldsymbol{\theta},[\mu_t]) + \tilde{V}(\boldsymbol{\theta}',[\mu_t])\right)\Sigma_t \\ &\quad - \alpha\mu_t(d\boldsymbol{\theta})\int_D K(\boldsymbol{\theta},\boldsymbol{\theta}'')\Sigma_t(d\boldsymbol{\theta}'',d\boldsymbol{\theta}') - \alpha\mu_t(d\boldsymbol{\theta}')\int_D K(\boldsymbol{\theta}',\boldsymbol{\theta}'')\Sigma_t(d\boldsymbol{\theta}'',d\boldsymbol{\theta}) \\ &\quad + \alpha\mu_t(d\boldsymbol{\theta})\int_D \tilde{V}(\boldsymbol{\theta}'',[\mu_t])\Sigma_t(d\boldsymbol{\theta}'',d\boldsymbol{\theta}) + \alpha\mu_t(d\boldsymbol{\theta}')\int_D \tilde{V}(\boldsymbol{\theta}'',[\mu])\Sigma_t(d\boldsymbol{\theta}'',d\boldsymbol{\theta}') \\ &\quad + \alpha\mu_t(d\boldsymbol{\theta})\int_{D\times D} K(\boldsymbol{\theta}''',\boldsymbol{\theta}'')\mu_t(d\boldsymbol{\theta}''')\Sigma_t(d\boldsymbol{\theta}'',d\boldsymbol{\theta}') + \alpha\mu_t(d\boldsymbol{\theta}')\int_D K(\boldsymbol{\theta}''',\boldsymbol{\theta}'')\mu_t(d\boldsymbol{\theta}''')\Sigma_t(d\boldsymbol{\theta}'',d\boldsymbol{\theta}) \\ &\quad + \alpha|\tilde{V}(\boldsymbol{\theta},[\mu_t])|\mu_t(d\boldsymbol{\theta})\delta_{\boldsymbol{\theta}}(d\boldsymbol{\theta}') - \alpha(|\tilde{V}(\boldsymbol{\theta},[\mu_t])| + |\tilde{V}(\boldsymbol{\theta}',[\mu_t])|)\mu_t(d\boldsymbol{\theta})\mu_t(d\boldsymbol{\theta}')\end{aligned} \tag{117}$$

This equation should also be interpreted in the weak sense by testing it against some $\phi \in C_c^\infty(D\times D)$, and it can be seen that it conserves mass in the sense that $\Sigma_t(d\boldsymbol{\theta}, D) = \Sigma_t(D, d\boldsymbol{\theta}') = 0$ for all $t > 0$ since this is true initially and $\partial_t\Sigma_t(d\boldsymbol{\theta}, D) = \partial_t\Sigma_t(D, d\boldsymbol{\theta}') = 0$.

We can also analyze the effect of the fluctuations at long times. Since $|\tilde{V}(\boldsymbol{\theta},[\mu_t])|\mu_t(d\boldsymbol{\theta}) \rightharpoonup 0$ as $t\to\infty$, the noise terms in (115) and (117) converge to zero—a property we refer to as self-quenching—and these equations reduce respectively to

$$\begin{aligned}\partial_t\omega_t &= \nabla\cdot\left(\nabla V(\boldsymbol{\theta},[\mu_*])\omega_t + \int_D \nabla K(\boldsymbol{\theta},\boldsymbol{\theta}')\omega_t(d\boldsymbol{\theta}')\mu_*\right) \\ &\quad - \alpha\tilde{V}(\boldsymbol{\theta},[\mu_*])\omega_t - \alpha\int_D K(\boldsymbol{\theta},\boldsymbol{\theta}')\omega_t(d\boldsymbol{\theta}')\mu_* \\ &\quad + \alpha(\int_D V(\boldsymbol{\theta}',[\mu_*])d\omega_t(\boldsymbol{\theta}'))\mu_* + \alpha(\int_{D\times D} K(\boldsymbol{\theta}',\boldsymbol{\theta}'')\mu_*(d\boldsymbol{\theta}')\omega_t(d\boldsymbol{\theta}''))\mu_*\end{aligned} \tag{118}$$

and

$$\begin{aligned}\partial_t\Sigma_t &= \nabla_{\boldsymbol{\theta}}\cdot\left(\nabla V(\boldsymbol{\theta},[\mu_*])\Sigma_t + \int_D \nabla K(\boldsymbol{\theta},\boldsymbol{\theta}'')\Sigma_t(d\boldsymbol{\theta},d\boldsymbol{\theta}'')\mu_*(d\boldsymbol{\theta})\right) \\ &\quad + \nabla_{\boldsymbol{\theta}'}\cdot\left(\nabla V(\boldsymbol{\theta}',[\mu_*])\Sigma_t + \int_D \nabla K(\boldsymbol{\theta}',\boldsymbol{\theta}'')\Sigma_t(d\boldsymbol{\theta}',d\boldsymbol{\theta}'')\mu_*(d\boldsymbol{\theta}')\right) \\ &\quad - \alpha\left(\tilde{V}(\boldsymbol{\theta},[\mu_*]) + \tilde{V}(\boldsymbol{\theta}',[\mu_*])\right)\Sigma_t \\ &\quad - \alpha\mu_t(d\boldsymbol{\theta})\int_D K(\boldsymbol{\theta},\boldsymbol{\theta}'')\Sigma_t(d\boldsymbol{\theta}'',d\boldsymbol{\theta}') - \alpha\mu_*(d\boldsymbol{\theta}')\int_D K(\boldsymbol{\theta}',\boldsymbol{\theta}'')\Sigma_t(d\boldsymbol{\theta}'',d\boldsymbol{\theta}) \\ &\quad + \alpha\mu_*(d\boldsymbol{\theta})\int_D \tilde{V}(\boldsymbol{\theta}'',[\mu])\Sigma_t(d\boldsymbol{\theta}'',d\boldsymbol{\theta}) + \alpha\mu_*(d\boldsymbol{\theta}')\int_D \tilde{V}(\boldsymbol{\theta}'',[\mu_*])\Sigma_t(d\boldsymbol{\theta}'',d\boldsymbol{\theta}') \\ &\quad + \alpha\mu_*(d\boldsymbol{\theta})\int_{D\times D^2} K(\boldsymbol{\theta}''',\boldsymbol{\theta}'')\mu_*(d\boldsymbol{\theta}''')\Sigma_t(d\boldsymbol{\theta}'',d\boldsymbol{\theta}') + \alpha\mu_*(d\boldsymbol{\theta}')\int_{D^2} K(\boldsymbol{\theta}''',\boldsymbol{\theta}'')\mu_*(d\boldsymbol{\theta}''')\Sigma_t(d\boldsymbol{\theta}'',d\boldsymbol{\theta})\end{aligned} \tag{119}$$

Since $\tilde{V}(\boldsymbol{\theta},[\mu_*]) \geq 0$, the fixed points of these equations are $\omega_t = 0$ and $\Sigma_t = 0$. That is, the effect of the fluctuations disappear as $t\to\infty$, and in particular they do not impede in the particle system the convergence observed at mean field level.

References

[CB18] Lénaïc Chizat and Francis Bach. A Note on Lazy Training in Supervised Differentiable Programming. working paper or preprint, December 2018.

[Daw06] Donald Dawson. Measure-valued Markov processes. In *École d'Été de Probabilités de Saint-Flour XXI—1991*, pages 1–260. Springer Berlin Heidelberg, Berlin, Heidelberg, September 2006.

[JKO98] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

[San17] Filippo Santambrogio. Euclidean, metric, and Wasserstein gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, March 2017.