A Properties of Mean-field Transfer to Particle System

Lemma A.1. *The Assumptions 1–3 transfer seamlessly to the aggregated notions of drift and diffusion given in (7):*

1. *If $b(\cdot,\cdot)$ and $\sigma(\cdot,\cdot)$ satisfy Assumption 1, then $\boldsymbol{b}(\cdot)$ and $\boldsymbol{\sigma}(\cdot)$ are $L(\sqrt{N}+1)$-Lipschitz.*
2. *If $b(\cdot,\cdot)$ satisfies Assumption 2, then $\boldsymbol{b}(\cdot)$ satisfies the same condition with the constant $\sqrt{N}C_v$.*
3. *If $\sigma(\cdot,\cdot)$ satisfies Assumption 3, then the same holds for $\boldsymbol{\sigma}(\cdot)$ with constant $\sqrt{N}K$.*

Proof. We prove these statements separately:

1. Let $x, y \in (\mathbb{R}^d)^{\otimes N}$ and define $a_i = |x^i - y^i|$. First, notice that $\mathcal{W}_2^2(\mu_x, \mu_y) \le \frac{1}{N}\sum a_i^2$, as the average on the right-hand-side corresponds to the specific coupling of $x_i \leftrightarrow y_i$. Now, observe that

$$\begin{aligned}|\boldsymbol{b}(x) - \boldsymbol{b}(y)|^2 &= \sum_i |b(x^i, \mu_x) - b(y^i, \mu_y)|^2 \\ &\le L^2 \sum_i \big(a_i + \mathcal{W}_2(\mu_x, \mu_y)\big)^2 \\ &\le L^2 \sum_i \Big(a_i + \sqrt{\tfrac{1}{N}\textstyle\sum_j a_j^2}\Big)^2.\end{aligned}$$

Let $\boldsymbol{a} = (a_1, \dots, a_N)$, and notice that the last quantity above is equal to

$$L^2\left|\boldsymbol{a} + \sqrt{\tfrac{1}{N}}|\boldsymbol{a}| \cdot \mathbf{1}\right|^2 = L^2|\boldsymbol{a}|^2\left|\frac{\boldsymbol{a}}{|\boldsymbol{a}|} + \sqrt{\tfrac{1}{N}} \cdot \mathbf{1}\right|^2 \le L^2|\boldsymbol{a}|^2 N\left(1 + \sqrt{\tfrac{1}{N}}\right)^2.$$

This means that

$$|\boldsymbol{b}(x) - \boldsymbol{b}(y)| \le L(\sqrt{N}+1)|x - y|.$$

For the diffusion, it suffices to notice that

$$\|\boldsymbol{\sigma}(x) - \boldsymbol{\sigma}(y)\|_{\mathrm{F}}^2 = \sum_i \|\sigma(x^i, \mu_x) - \sigma(y^i, \mu_y)\|_{\mathrm{F}}^2.$$

The rest of the proof is similar to the one for the drift.

2. We have

$$\frac{1}{N}\langle x, \boldsymbol{b}(x)\rangle = \frac{1}{N}\sum_{i=1}^N \langle x^i, b(x^i, \mu_x)\rangle \le C_v\Big(\frac{1}{N}\sum |x^i| + 1\Big) \le C_v\Big(\frac{1}{\sqrt{N}}\sqrt{\textstyle\sum |x^i|^2} + 1\Big),$$

where in the last inequality, we used Cauchy-Schwarz. This implies $\langle x, \boldsymbol{b}(x)\rangle \le C_v\sqrt{N}(|x|+1)$.

3. It is easy to see that

$$\|\boldsymbol{\sigma}(x)\|_F^2 = \operatorname{tr}(\boldsymbol{\sigma}(x)^\top \boldsymbol{\sigma}(x)) = \sum_{i=1}^N \operatorname{tr}(\sigma(x^i, \mu_x)^\top \sigma(x^i, \mu_x)) \le NK^2. \qquad \blacksquare$$

Lemma A.2. *If $b(\cdot,\cdot)$ is (α,β)-dissipative on average, then $\boldsymbol{b}(\cdot)$ is $(\alpha, N\beta)$-dissipative in the usual sense, that is, for all $x \in (\mathbb{R}^d)^{\otimes N}$, $\langle x, \boldsymbol{b}(x)\rangle \le -\alpha|x|^2 + N\beta$.*

Proof. Observe that for $x \in (\mathbb{R}^d)^{\otimes N}$ we have

$$\begin{aligned}\frac{1}{N}\langle x, \boldsymbol{b}(x)\rangle = \frac{1}{N}\sum_{i=1}^N \langle x^i, b(x^i, \mu_x)\rangle &= \mathbb{E}_{\mu_x}[\langle y, b(y, \mu_x)\rangle] \le -\alpha\,\mathbb{E}_{\mu_x}|y|^2 + \beta \\ &= -\alpha\frac{1}{N}\sum_{i=1}^N |x^i|^2 + \beta = -\alpha\frac{1}{N}|x|^2 + \beta.\end{aligned}$$

This means that $\langle x, \boldsymbol{b}(x)\rangle \le -\alpha|x|^2 + N\beta$. $\blacksquare$

B The Main Theorem

B.1 Proof of Theorem 1

Recall the *Picard process* (Picard):

$$\Pi_s^{(t)} = X_t + \int_0^s \boldsymbol{b}(X_{t+u})\,du + \int_0^s \boldsymbol{\sigma}(X_{t+u})\,dW_u^{(t)}.$$

We break down the proof into four steps: first, we prove that the Picard process is close to the flow (Flow), and then we bound the distance between the Picard process and the interpolation (Int). We then conclude the proof of the WAPT property. Lastly, we prove that stability is implied by dissipativity.

§ Distance of Picard from Flow.

$$\begin{aligned}
\mathbb{E}|\Pi_s^{(t)} - \Phi_s^{(t)}|^2 &\le 2\,\mathbb{E}\left|\int_0^s \boldsymbol{b}(X_{t+u}) - \boldsymbol{b}(\Phi_u^{(t)})\,du\right|^2 + 2\,\mathbb{E}\left|\int_0^s \boldsymbol{\sigma}(X_{t+u}) - \boldsymbol{\sigma}(\Phi_u^{(t)})\,dW_u^{(t)}\right|^2 \\
&\le T\int_0^s \mathbb{E}|\boldsymbol{b}(\Phi_u^{(t)}) - \boldsymbol{b}(X_{t+u})|^2\,du + 2\,\mathbb{E}\int_0^s \|\boldsymbol{\sigma}(X_{t+u}) - \boldsymbol{\sigma}(\Phi_u^{(t)})\|_{\mathrm{F}}^2\,du \\
&\le 2(T+1)L^2\int_0^s \mathbb{E}|\Phi_u^{(t)} - X_{t+u}|^2
\end{aligned}$$

where we used Lipschitzness of $\boldsymbol{b}$ and $\boldsymbol{\sigma}$ (implied by Assumption 1 and Lemma A.1), Itô's isometry (see, e.g., [54, Lemma 3.4]), and Lemma A.1.

§ Distance of Picard to Interpolation. We place a bar above a symbol to denotes its piecewise constant interpolation.

$$\begin{aligned}
\mathbb{E}|\Pi_s^{(t)} - X_{t+s}|^2 &= \mathbb{E}\left|\int_t^{t+s} \boldsymbol{b}(X_u) - \boldsymbol{b}(\overline{X_u})\,du + \int_t^{t+s} \boldsymbol{\sigma}(X_u) - \boldsymbol{\sigma}(\overline{X_u})\,dW_u^{(t)} + \Delta_P(t,s)\right|^2 \\
&\le 3(T+1)L^2\int_t^{t+s} \mathbb{E}|X_u - \overline{X_u}|^2\,du + 3\,\mathbb{E}\left|\Delta_P(t,s)\right|^2,
\end{aligned}$$

where $\Delta_P(t,s)$ is the accumulated noise and bias from time t to time $t+s$, which is equal to

$$\Delta_P(t,s) \coloneqq \sum_{i=k}^{n-1} \gamma_{i+1}P_{i+1} + (t+s-\tau_n)\,\mathbb{E}[P_{n+1} \mid \mathscr{F}_{t+s}] - (t-\tau_k)\,\mathbb{E}[P_{k+1} \mid \mathscr{F}_t], \tag{B.1}$$

with $n = m(t+s)$ and $k = m(t)$. It is shown in [23] that $\lim_{t\to\infty} \mathbb{E}\left|\Delta_P(t,s)\right|^2 = 0$, a.s.

Continuing to bound the inside of the integral, we have

$$\mathbb{E}|X_t - x_k|^2 \le 3(t-\tau_k)^2(\mathbb{E}|\boldsymbol{b}(x_k)|^2 + \mathbb{E}|P_{k+1}|^2) + 3(t-\tau_k)\,\mathbb{E}\,\mathrm{tr}(\boldsymbol{\sigma}(x_k)^\top\boldsymbol{\sigma}(x_k))$$

where we used the fact that conditional expectation is a contraction in L^2, and

$$\begin{aligned}
\mathbb{E}|\boldsymbol{\sigma}(x_k)\,\xi_{k+1}|^2 &= \mathbb{E}\,\mathrm{tr}\big(\xi_{k+1}^\top\boldsymbol{\sigma}(x_k)^\top\boldsymbol{\sigma}(x_k)\,\xi_{k+1}\big) \\
&= \mathbb{E}\,\mathrm{tr}\big(\boldsymbol{\sigma}(x_k)^\top\boldsymbol{\sigma}(x_k)\,\xi_{k+1}\xi_{k+1}^\top\big) \\
&= \mathbb{E}\,\mathrm{tr}\big(\boldsymbol{\sigma}(x_k)^\top\boldsymbol{\sigma}(x_k)\,\mathbb{E}[\xi_{k+1}\xi_{k+1}^\top \mid \mathscr{F}_{\tau_k}]\big) \\
&= \mathbb{E}\,\mathrm{tr}\big(\boldsymbol{\sigma}(x_k)^\top\boldsymbol{\sigma}(x_k)\big).
\end{aligned}$$

Moreover, by Assumption 3 we have $\mathbb{E}\,\mathrm{tr}(\boldsymbol{\sigma}(x_k)^\top\boldsymbol{\sigma}(x_k)) = \mathcal{O}(1)$. We thus get by Lemma B.1

$$\mathbb{E}|X_t - x_k|^2 \le 3C\gamma_{k+1}^2(1/\gamma_{k+1} + 1) + 3C\gamma_{k+1} = \mathcal{O}(\gamma_{k+1}).$$

This implies

$$\sup_{s\in[0,T]} \mathbb{E}|\Pi_s^{(t)} - X_{t+s}|^2 \le CT^2L^2 \sup_{t\le u\le t+T} \overline{\gamma_u} + 3\,\mathbb{E}\left|\Delta_P(t,T)\right|^2 =: A_t,$$

with $A_t \to 0$ as $t \to \infty$, a.s.

§ Concluding the proof of APT. By Grönwall inequality,

$$\mathbb{E}|X_{t+s} - \Phi_s^{(t)}|^2 \le C \int_0^s \mathbb{E}|X_{t+u} - \Phi_u^{(t)}|^2 + A_t \le A_t \exp(sC) \le A_t \exp(TC) \to 0$$

as $t \to \infty$. Since

$$\mathcal{W}_2^2(\mathrm{law}(X_{t+s}), \mathrm{law}(\Phi_s^{(t)})) \le \mathbb{E}|X_{t+s} - \Phi_s^{(t)}|^2,$$

we get the desired result.

§ Stability. Lemma A.2 implies that under dissipativity on average, the iterates are stable, and as in [23, Theorem 3], we get the desired convergence result.

B.2 On Propagation of Chaos

Theorem 1 shows that the law μ_k of $x_k \in (\mathbb{R}^d)^{\otimes N}$ converges in the Wasserstein space to the limit-set (or the *internally chain-transitive* (ICT) set) $S \subset \mathscr{P}_2((\mathbb{R}^d)^{\otimes N})$ of the corresponding flow:

$$\lim_{k\to\infty} \inf_{\mu\in S} \mathcal{W}_2(\mu_k, \mu) = 0.$$

By looking only at the first particle of x_k, namely, x_k^1, and given that the dynamics is exchangeable, it follows that

$$\begin{aligned}
\mathcal{W}_2^2(\mu_k, \mu) &= \inf_{\pi} \int |x - y|^2 \, \pi(dx, dy) \\
&= \inf_{\pi} \left(\int |x^1 - y^1|^2 \, \pi^1(dx^1, dy^1) + \cdots + \int |x^N - y^N|^2 \, \pi^N(dx^N, dy^N) \right) \\
&\ge \inf_{\pi^1} \int |x^1 - y^1|^2 \, \pi^1(dx^1, dy^1) \\
&= \mathcal{W}_2^2(\mathrm{law}(x_k^1), \mathrm{marginal}_1(\mu)),
\end{aligned}$$

where $\pi^1(dx^1, dy^1) = \int \pi(x, y)\, dx^2 dy^2 \cdots dx^N dy^N$, and we used the exchangability in deducing that the law of y^i are the same as the first marginal of μ, for all $i = 1, \ldots, N$. Hence, as the limit-set S' of the first component of the SDE X_t^1 is a subset of the marginal of S,

$$\lim_{k\to\infty} \inf_{\nu\in S'} \mathcal{W}_2(\mathrm{law}(x_k^1), \nu) \le \lim_{k\to\infty} \inf_{\nu\in \mathrm{marginal}_1(S)} \mathcal{W}_2(\mathrm{law}(x_k^1), \nu) = 0.$$

This means that the first particle converges in law to the ICT sets of the corresponding SDE. Assuming a uniform propagation of chaos, we also know that the law of X_t^1 has a distance of $O(1/N)$ from the mean-field equation, and hence, we get that the law of the particles following the discrete algorithm have controllable distance from the mean-field dynamics.

B.3 Supporting Lemmas

Lemma B.1. *Suppose Assumptions 1–5 hold. One has* $\mathbb{E}|\boldsymbol{b}(x_k)|^2 = \mathcal{O}(1/\gamma_{k+1})$, $\mathbb{E}|\varepsilon_{k+1}|^2 = \mathcal{O}(\gamma_{k+1})$, *and* $\mathbb{E}|P_{k+1}|^2 = \mathcal{O}(1)$.

Proof. We repeatedly use the fact that $\mathbb{E}|\boldsymbol{b}(x_k)|^2 \le 2L^2\,\mathbb{E}|x_k|^2 + \mathbb{E}|\boldsymbol{b}(x_0)|^2 =: 2L^2\,\mathbb{E}|x_k|^2 + C_0$. By Assumption 4, $\mathbb{E}|\varepsilon_{k+1}|^2 \le \mathcal{O}(\gamma_{k+1}^2)a_k + \mathcal{O}(\gamma_{k+1})$, and we have

$$\mathbb{E}|P_{k+1}|^2 \le 2\,\mathbb{E}|\varepsilon_{k+1}|^2 + 2\,\mathbb{E}|U_{k+1}|^2 = \mathcal{O}(\gamma_{k+1}^2)a_k + \mathcal{O}(1). \tag{B.2}$$

Moreover, as $\sqrt{p+q} \le \sqrt{p} + \sqrt{q}$, we have

$$\sqrt{\mathbb{E}|P_{k+1}|^2} \le \mathcal{O}(\gamma_{k+1})\sqrt{a_k} + \mathcal{O}(1). \tag{B.3}$$

Assumption 3 also implies that $\mathbb{E}|\boldsymbol{\sigma}(x_k)\xi_{k+1}|^2 \le C_\sigma$.

Define $a_k := \mathbb{E}|x_k|^2$. Then,

$$\begin{aligned} a_{k+1} - a_k &= \gamma_{k+1}^2 \,\mathbb{E}|\boldsymbol{b}(x_k) + P_{k+1}|^2 + \gamma_{k+1}\,\mathbb{E}|\boldsymbol{\sigma}(x_k)\xi_{k+1}|^2 + 2\gamma_{k+1}\,\mathbb{E}\langle x_k, \boldsymbol{b}(x_k) + P_{k+1}\rangle \\ &\quad + 2\gamma_{k+1}^{1/2}\,\mathbb{E}\langle x_k, \boldsymbol{\sigma}(x_k)\xi_{k+1}\rangle + 2\gamma_{k+1}^{3/2}\,\mathbb{E}\langle \boldsymbol{b}(x_k) + P_{k+1}, \boldsymbol{\sigma}(x_k)\xi_{k+1}\rangle \\ &\le 2L^2\gamma_{k+1}^2 a_k + \gamma_{k+1}^2 C_0 + 2\gamma_{k+1}^2\,\mathbb{E}|P_{k+1}|^2 + \gamma_{k+1}C_\sigma + 2\gamma_{k+1}C_v(\sqrt{a_k}+1) \\ &\quad + 2\gamma_{k+1}\sqrt{a_k}\sqrt{\mathbb{E}|P_{k+1}|^2} + 2\gamma_{k+1}^{3/2}\sqrt{C_\sigma}\sqrt{\mathbb{E}|P_{k+1}|^2} \end{aligned} \tag{B.4}$$

Plugging the bounds from (B.2) and (B.3) into (B.4) gives

$$\begin{aligned} a_{k+1} - a_k &\le \mathcal{O}(\gamma_{k+1}^2)a_k + \mathcal{O}(\gamma_{k+1})\sqrt{a_k} + \mathcal{O}(\gamma_{k+1}) \\ &=: P\gamma_{k+1}^2 a_k + Q\gamma_{k+1}\sqrt{a_k} + R\gamma_{k+1}, \end{aligned}$$

for some $P, Q, R > 0$ that do not depend on k.

We now prove $a_k \le M/\gamma_{k+1}$ for some fixed $M > 0$ via induction. Suppose this is the case for k. For $k+1$ we have

$$\begin{aligned} a_{k+1} &\le (P\gamma_{k+1}^2 + 1)a_k + Q\gamma_{k+1}\sqrt{a_k} + R\gamma_{k+1} \\ &\le M(P\gamma_{k+1} + 1/\gamma_{k+1}) + \sqrt{M}Q\sqrt{\gamma_{k+1}} + R\gamma_{k+1} \\ &\overset{!}{\le} M/\gamma_{k+2}. \end{aligned}$$

The last inequality is equivalent to the fact that the following quadratic equation (in $\sqrt{M}$) has a bounded largest root (and the bound shall not depend on k):

$$M(P\gamma_{k+1} + 1/\gamma_{k+1} - 1/\gamma_{k+2}) + \sqrt{M}Q\sqrt{\gamma_{k+1}} + R\gamma_{k+1}$$

Notice that by Assumption 5, the leading coefficient is negative, and the larger root is computed as

$$\begin{aligned} &\frac{Q\sqrt{\gamma_{k+1}} + Q\sqrt{\gamma_{k+1}} + \sqrt{4R(\gamma_{k+1}/\gamma_{k+2} - P\gamma_{k+1}^2 - 1)}}{2(1/\gamma_{k+2} - P\gamma_{k+1} - 1/\gamma_{k+1})} \\ &\le \frac{2Q\sqrt{\gamma_{k+1}} + 2\sqrt{R}\sqrt{\gamma_{k+1}/\gamma_{k+2}}}{2\gamma_{k+1}/\gamma_{k+2}} \le 2Q\sqrt{\gamma_{k+2}} + \sqrt{R}\sqrt{\gamma_{k+2}/\gamma_{k+1}} < 2Q + \sqrt{R} =: M. \end{aligned}$$

The claim for $\boldsymbol{b}$ follows by Lipschitzness, from which the claim for the bias and perturbation follows directly. ∎

C Proofs of Results for Applications

C.1 Two-Layer Neural Networks and Mean-field Langevin

Proof of Corollary 1. **Smoothness of drift:** We start by showing Lipschitzness with respect to the measure parameter of the drift. First, observe that $\nabla_\theta W(\theta, \cdot)$ is Lipschitz:

$$\begin{aligned} |\nabla_\theta W(\theta, \theta') - \nabla_\theta W(\theta, \theta'')| &= |\mathbb{E}_{z\sim\mathcal{D}}[(\varphi(z, \theta') - \varphi(z, \theta''))\,\nabla_\theta\varphi(z, \theta)]| \\ &= |\mathbb{E}_{z\sim\mathcal{D}}[(\kappa(\langle z, \theta'\rangle) - \kappa(\langle z, \theta''\rangle))\,\kappa'(\langle z, \theta\rangle)\,z]| \\ &\le C|\theta' - \theta''|, \end{aligned}$$

due to the boundedness of κ' and z, and Lipschitzness of κ.

Now, consider $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ and let π be the optimal coupling (in $\mathcal{W}_2$ sense) between them. Then, for a fixed $\theta \in \mathbb{R}^d$,

$$\begin{aligned} |b(\theta, \mu) - b(\theta, \nu)|^2 &= \left|\int \nabla_\theta W(\theta, p) - \nabla_\theta W(\theta, q)\,\pi(dp, dq)\right|^2 \\ &\le \int |\nabla_\theta W(\theta, p) - \nabla_\theta W(\theta, q)|^2\,\pi(dp, dq) \\ &\le \int C^2|p - q|^2\,\pi(dp, dq) \\ &= C^2 W_2^2(\mu, \nu). \end{aligned}$$

Next, we show for a fixed measure $\mu \in \mathscr{P}_2(\mathbb{R}^d)$, $b(\cdot, \mu)$ is Lipschitz in the first input.

$$|b(\theta,\mu) - b(\theta',\mu)| \le \left| \int \nabla_\theta W(\theta,p) - \nabla_\theta W(\theta',p)\, \mu(dp) \right| + |\nabla V(\theta) - \nabla V(\theta')|.$$

Let us treat each term separately. We have

$$\begin{aligned}
|\nabla_\theta W(\theta,p) - \nabla_\theta W(\theta',p)| &= |\mathbb{E}_{z\sim\mathcal{D}}[\varphi(z,p)(\nabla_\theta \varphi(z,\theta) - \nabla_\theta \varphi(z,\theta'))]| \\
&= |\mathbb{E}_{z\sim\mathcal{D}}[\kappa(\langle p, z\rangle)(\kappa'(\langle z,\theta\rangle) - \kappa'(\langle z,\theta'\rangle))\, z]| \\
&\le C\, \mathbb{E}_{z\sim\mathcal{D}}[|z||\theta-\theta'|\, z] \\
&\le C|\theta-\theta'|.
\end{aligned}$$

Similarly,

$$|\nabla V(\theta) - \nabla V(\theta')| = \left|\mathbb{E}_{(y,z)\sim\mathcal{D}}[yz(\kappa'(\langle\theta,z\rangle) - \kappa'(\langle\theta',z\rangle))\, \mathcal{D}(dy,dz)]\right| \le C|\theta-\theta'|.$$

Thus,

$$|b(\theta,\mu) - b(\theta',\nu)| \le |b(\theta,\mu) - b(\theta,\nu)| + |b(\theta,\nu) - b(\theta',\nu)| \le L(|\theta-\theta'| + \mathcal{W}_2(\mu,\nu)),$$

showing b satisfies Assumption 1.

Growth control: First, let us calculate

$$\begin{aligned}
\int \langle\theta, b(\theta,\mu)\rangle\, \mu(d\theta) &= \iint \varphi(z,\theta')\langle\theta, \nabla_\theta\varphi(z,\theta)\rangle\, \mathcal{D}(dz)\mu(d\theta')\mu(d\theta) \\
&\quad - \iint y\, \langle\theta, \nabla_\theta\varphi(z,\theta)\rangle\, \mathcal{D}(dy,dz)\mu(d\theta) \\
&\quad - \lambda \int |\theta|^2\, \mu(d\theta) \\
&= \iint \varphi(z,\theta')\kappa'(\langle z,\theta\rangle)\langle\theta, z\rangle\, \mathcal{D}(dz)\mu(d\theta')\mu(d\theta) \\
&\quad - \iint y\, \kappa'(\langle z,\theta\rangle)\langle\theta, z\rangle\, \mathcal{D}(dy,dz)\mu(d\theta) \\
&\quad - \lambda \int |\theta|^2\, \mu(d\theta).
\end{aligned}$$

As φ, κ', and $\mathrm{supp}(\mathcal{D})$ are bounded, we can see that

$$\left|\int \langle\theta, b(\theta,\mu)\rangle\, \mu(d\theta)\right| \le C \iint |\theta||z|\, \mathcal{D}(dz)\mu(d\theta) + C' \int |\theta||z|\, \mathcal{D}(dz)\mu(d\theta) \le C \int |\theta|\, \mu(d\theta),$$

thus, satisfying Assumption 2.

Dissipativity on average: Here we use the extra assumption that $|a\,\kappa'(a)|$ is bounded. We directly bound the terms $\kappa'(\langle\theta,z\rangle)\langle\theta,z\rangle$ above and obtain

$$\int \langle\theta, b(\theta,\mu)\rangle\, \mu(d\theta) \le -\lambda \int |\theta|^2\, \mu(d\theta) + C. \qquad \blacksquare$$

C.2 Stein Variational Gradient Descent

Proof of Corollary 2. While the first term in the drift is standard to work with (see Section 4.4), it is the second term in the drift that makes it difficult to analyze. Specifically, we prove the dissipativity on average only for empirical measures. While this would be enough for our purposes (and Theorem 1 goes through), it is an interesting future direction to see when does dissipativity hold in a more general setup. Moreover, for simplicity, we only consider the case where the kernel K is of the form $K(x,y) = h(x-y)$, for some function h.

Below, we first prove that b is dissipative on average, which implies that the law of the iterates will be in a compact subset of $\mathscr{P}_2(\mathbb{R}^d)$. Then, we show that b is smooth on this compact subset.

Dissipativity on average: Due to K being symmetric, $\nabla_2 K(x,y) = -\nabla_2 K(y,x)$. We thus have

$$\begin{aligned}
&\iint \langle x-y, \nabla_2 K(x,y)\rangle \,\mu(dx)\mu(dy)\\
&= \iint \langle x, \nabla_2 K(x,y)\rangle \,\mu(dx)\mu(dy) - \iint \langle y, \nabla_2 K(x,y)\rangle \,\mu(dx)\mu(dy)\\
&= \iint \langle x, \nabla_2 K(x,y)\rangle \,\mu(dx)\mu(dy) + \iint \langle y, \nabla_2 K(y,x)\rangle \,\mu(dx)\mu(dy)\\
&= 2\iint \langle x, \nabla_2 K(x,y)\rangle \,\mu(dx)\mu(dy).
\end{aligned}$$

Thus,

$$\iint \langle x, \nabla_2 K(x,y)\rangle \,\mu(dx)\mu(dy) = \frac{1}{2}\iint \langle x-y, \nabla_2 K(x,y)\rangle \,\mu(dx)\mu(dy) \le \eta$$

by Cauchy-Schwarz and the assumption that $|\nabla_2 K(x,y)| \le \eta/|x-y|$. With similar arguments, and using dissipativity of V, we have

$$\begin{aligned}
&\iint \langle x, \nabla V(y)\rangle K(x,y)\,\mu(dx)\mu(dy)\\
&= \iint \langle x, \nabla V(x)\rangle K(x,y)\,\mu(dx)\mu(dy) - \frac{1}{2}\iint \langle x-y, \nabla V(x) - \nabla V(y)\rangle K(x,y)\,\mu(dx)\mu(dy)\\
&\ge \alpha \iint |x|^2 K(x,y)\,\mu(dx)\mu(dy) - \beta\|K\|_\infty\\
&\quad - \frac{1}{2}L\iint |x-y|^2 K(x,y)\,\mu(dx)\mu(dy)\\
&\ge \alpha \iint |x|^2 K(x,y)\,\mu(dx)\mu(dy) - \beta\|K\|_\infty + \frac{L\eta}{2}.
\end{aligned}$$

As $K(x,x) = h(0)$ and $K(x,y) > 0$ for all x, y, and that μ is an empirical measure $\mu = \frac{1}{N}\sum_{i=1}^N \delta_{x_i}$, the last quantity is equal to

$$\frac{1}{N^2}\sum_i |x_i|^2 \sum_j K(x_i,x_j) \ge \frac{1}{N^2}\sum_i |x_i|^2 K(x_i,x_i) \ge \frac{h(0)}{N}\int |x|^2\,\mu(dx).$$

In total, we derive that b is dissipative on average.

Smoothness of the drift: We have, for μ in a compact set of $\mathscr{P}_2(\mathbb{R}^d)$

$$\begin{aligned}
&|b(x,\mu) - b(x',\mu)|\\
&\le \left|\int \nabla_2 K(x,y) - \nabla_2 K(x',y)\,\mu(dy)\right| + \left|\int (K(x,y) - K(x',y))\nabla V(y)\,\mu(dy)\right|\\
&\le L|x-x'|\left(1 + \int |\nabla V(y)|\,\mu(dy)\right)\\
&\le L|x-x'|\left(1 + C\int (1+|y|^2)\,\mu(dy)\right) < L'|x-x'|.
\end{aligned}$$

Moreover, take μ, ν in the same compact set, and let π be the optimal coupling (in $\mathcal{W}_2$ sense). Then,

$$\begin{aligned}
&|b(x,\mu) - b(x,\nu)|^2\\
&\le 2\left|\int \nabla_2 K(x,y) - \nabla_2 K(x,z)\,\pi(dy,dz)\right|^2\\
&\qquad + 2\left|\int K(x,y)\nabla V(y) - K(x,z)\nabla V(z)\,\pi(dy,dz)\right|^2\\
&\le 2L^2\mathcal{W}_2^2(\mu,\nu)\\
&\qquad + 2\left|\int K(x,y)\nabla V(y) - K(x,z)\nabla V(y) + K(x,z)\nabla V(y) - K(x,z)\nabla V(z)\,\pi(dy,dz)\right|^2\\
&\le 2L^2\mathcal{W}_2^2(\mu,\nu) + 4(L^2 + L'^2)\mathcal{W}_2^2(\mu,\nu).
\end{aligned}$$

■

C.3 Two-player Zero-sum Continuous Games

Proof of Corollary 3. Recall that

$$b(q,\mu) \coloneqq \int \begin{pmatrix} -\nabla_x K(q_1, q_2') \\ \alpha\, \nabla_y K(q_1', q_2) \end{pmatrix} \mu(dq'), \quad q = (q_1, q_2).$$

Smoothness of drift: For a fixed $\mu \in \mathscr{P}_2(\mathbb{R}^d)\mathbb{R}^{2d}$ and $q, r \in \mathbb{R}^{2d}$ we have

$$\begin{aligned} |b(q,\mu) - b(r,\mu)|^2 &\le \int \left| \begin{pmatrix} -\nabla_x K(q_1, q_2') \\ \alpha\, \nabla_y K(q_1', q_2) \end{pmatrix} - \begin{pmatrix} -\nabla_x K(r_1, q_2') \\ \alpha\, \nabla_y K(q_1', r_2) \end{pmatrix} \right|^2 \mu(dq') \\ &\le L^2 |q_1 - r_1|^2 + \alpha^2 L^2 |q_2 - r_2|^2 \\ &\le L^2 |q - r|^2, \end{aligned}$$

where we used L-Lipschitzness of $\nabla_x K$ and $\nabla_y K$.

Now, for a fixed $q \in \mathbb{R}^{2d}$, and $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)\mathbb{R}^{2d}$ with optimal coupling π, we have

$$\begin{aligned} |b(q,\mu) - b(q,\nu)|^2 &\le \iint \left| \begin{pmatrix} -\nabla_x K(q_1, r_2) \\ \alpha\, \nabla_y K(r_1, q_2) \end{pmatrix} - \begin{pmatrix} -\nabla_x K(q_1, r_2') \\ \alpha\, \nabla_y K(r_1', q_2) \end{pmatrix} \right|^2 \pi(dr, dr') \\ &\le L^2 \iint |r_2 - r_2'|^2 + |r_1 - r_1'|^2 \, \pi(dr, dr') \\ &= L^2 \mathcal{W}_2^2(\mu, \nu). \end{aligned}$$

Average dissipativity of drift: Suppose $\nabla_x K$ and $-\nabla_y K$ are (a, β)-dissipative. Then

$$\begin{aligned} \int \langle q, b(q,\mu) \rangle \, \mu(dq) &= \iint \langle q_1, -\nabla_x K(q_1, q_2') \rangle + \alpha \langle q_2, \nabla_y K(q_1', q_2) \rangle \, \mu(dq')\mu(dq) \\ &\le \int -a\alpha (|q_1|^2 + |q_2|^2) \, \mu(dq) + 2\beta, \end{aligned}$$

implying that $b(\cdot,\cdot)$ is $(a\alpha, 2\beta)$-dissipative on average.

If, on the other hand, the domains $\mathcal{X}$ and $\mathcal{Y}$ are bounded, observe that by Cauchy-Schwarz

$$\left| \int \langle q, b(q,\mu) \rangle \, \mu(dq) \right| \le \iint |q_1| |\nabla_x K(q_1, q_2')| + \alpha |q_2| |\nabla_y K(q_1', q_2)| \, \mu(dq')\mu(dq) \le M,$$

where $M = \sup_{q_1 \in \mathcal{X}, q_2 \in \mathcal{Y}} |q_1| |\nabla_x K(q_1, q_2')| + \alpha |q_2| |\nabla_y K(q_1', q_2)|$. Also denoting by $R = \sup_{q \in \mathcal{X} \times \mathcal{Y}} |q|^2$, we see that for any $\alpha > 0$, $b(\cdot,\cdot)$ is $(\alpha, M + \alpha N)$-dissipative on average, as

$$\int \langle q, b(q,\mu) \rangle \, \mu(dq) + \alpha \int |q|^2 \, \mu(dq) \le M + \alpha N.$$

Optimistic algorithm fits Assumption 4: Recall the iterates

$$q_{k+1}^i = q_k^i + \gamma_{k+1} \big(2b(q_k^i, \widehat{\mu}_k) - b(q_{k-1}^i, \widehat{\mu}_{k-1}) \big) + \sqrt{2\gamma_{k+1}}\, \sigma\, \Xi_{k+1}^i,$$

where $\Xi_{k+1}^i = (\xi_{k+1}^i, \zeta_{k+1}^i)$. Notice that the bias of this iteration is

$$\varepsilon_{k+1}^i = b(q_k^i, \widehat{\mu}_k) - b(q_{k-1}^i, \widehat{\mu}_{k-1}).$$

For brevity, let us write $\mathscr{F}_k \coloneqq \mathscr{F}_{\tau_k}$. We have

$$\begin{aligned} \mathbb{E}[|\varepsilon_{k+1}^i|^2 \,|\, \mathscr{F}_k] &= \mathbb{E}[|q_{k+1}^i - q_k^i - \gamma_{k+1} b(q_k^i, \widehat{\mu}_k) - \sqrt{2\gamma_{k+1}}\, \sigma\, \Xi_{k+1}^i|^2 \,|\, \mathscr{F}_k] \\ &\le 3\, \mathbb{E}[|q_{k+1}^i - q_k^i|^2 \,|\, \mathscr{F}_k] + 3\gamma_{k+1}^2\, \mathbb{E}[|b(q_k^i, \widehat{\mu}_k)|^2 \,|\, \mathscr{F}_k] + 6\gamma_{k+1}\tau(1+\alpha)d. \end{aligned}$$

Moreover, we have

$$\begin{aligned} \mathbb{E}[|q_{k+1}^i - q_k^i|^2 \,|\, \mathscr{F}_k] &\le 2\gamma_{k+1}^2\, \mathbb{E}[|2b(q_k^i, \widehat{\mu}_k) - b(q_{k-1}^i, \widehat{\mu}_{k-1})|^2 \,|\, \mathscr{F}_k] + 2\gamma_{k+1}\tau(1+\alpha)d \\ &= 2\gamma_{k+1}^2\, \mathbb{E}[|b(q_k^i, \widehat{\mu}_k) + \varepsilon_{k+1}^i|^2 \,|\, \mathscr{F}_k] + 2\gamma_{k+1}\tau(1+\alpha)d \\ &\le 4\gamma_{k+1}^2\, \mathbb{E}[|b(q_k^i, \widehat{\mu}_k)|^2 \,|\, \mathscr{F}_k] + 4\gamma_{k+1}^2\, \mathbb{E}[|\varepsilon_{k+1}^i|^2 \,|\, \mathscr{F}_k] + 2\gamma_{k+1}\tau(1+\alpha)d \end{aligned}$$

Combining the last two inequalities, we have

$$\begin{aligned}\mathbb{E}[|\varepsilon^i_{k+1}|^2 \,|\, \mathscr{F}_k] &\le 3\Big(4\gamma^2_{k+1}\,\mathbb{E}[|b(q^i_k,\widehat{\mu}_k)|^2 \,|\, \mathscr{F}_k] + 4\gamma^2_{k+1}\,\mathbb{E}[|\varepsilon^i_{k+1}|^2 \,|\, \mathscr{F}_k] + 2\gamma_{k+1}\tau(1+\alpha)d\Big)\\ &\quad + 3\gamma^2_{k+1}\,\mathbb{E}[|b(q^i_k,\widehat{\mu}_k)|^2 \,|\, \mathscr{F}_k] + 6\gamma_{k+1}\tau(1+\alpha)d\\ &= 12\gamma^2_{k+1}\,\mathbb{E}[|\varepsilon^i_{k+1}|^2 \,|\, \mathscr{F}_k] + 15\gamma^2_{k+1}\,\mathbb{E}[|b(q^i_k,\widehat{\mu}_k)|^2 \,|\, \mathscr{F}_k] + 12\gamma_{k+1}\tau(1+\alpha)d\end{aligned}$$

Since $\gamma_{k+1} \to 0$, we can assume that $12\gamma^2_{k+1} \le 1/2$, which implies

$$\mathbb{E}[|\varepsilon^i_{k+1}|^2 \,|\, \mathscr{F}_k] \le 30\gamma^2_{k+1}\,\mathbb{E}[|b(q^i_k,\widehat{\mu}_k)|^2 \,|\, \mathscr{F}_k] + 24\gamma_{k+1}\tau(1+\alpha)d,$$

which is exactly what we are after. ∎

C.4 Kinetic Equations

Proof of Corollary 4. **Smoothness of the drift:** Let $x, y \in \mathbb{R}^d$ and $\mu, \nu \in \mathscr{P}_2(\mathbb{R}^d)$ and set π be an optimal coupling between μ and ν (in W_1 sense). Then

$$|b(x,\mu) - b(y,\nu)| \le |\nabla V(x) - \nabla V(y)| + \left|\int \nabla W(x-z)\,\mu(dz) - \int \nabla W(y-z)\,\nu(dz)\right|.$$

By L-Lipschitzness of ∇V, the first term is bounded by $L|x-y|$. For the second term, using the coupling, we can write it as

$$\begin{aligned}\left|\iint \nabla W(x-z_1) - \nabla W(y-z_2)\,\pi(dz_1,dz_2)\right| &\le \iint |\nabla W(x-z_1) - \nabla W(y-z_2)|\,\pi(dz_1,dz_2)\\ &\le L\iint |x-y+z_2-z_1|\,\pi(dz_1,dz_2)\\ &\le L\iint |x-y| + |z_2-z_1|\,\pi(dz_1,dz_2)\\ &\le L|x-y| + LW_1(\mu,\nu)\\ &\le L|x-y| + LW_2(\mu,\nu).\end{aligned}$$

Putting these together we get

$$|b(x,\mu) - b(y,\nu)| \le 2L(|x-y| + \mathcal{W}_2(\mu,\nu)).$$

Average dissipativity of the drift: First we show that for $x \in \mathbb{R}^d$ and a probability measure μ, we have

$$\iint \langle x, \nabla W(x-y)\rangle\,\mu(dx)\mu(dy) \ge -M_W/2. \tag{C.1}$$

This holds, since

$$\begin{aligned}&\iint \langle x, \nabla W(x-y)\rangle\,\mu(dx)\mu(dy)\\ &\quad= \iint \langle x-y+y, \nabla W(x-y)\rangle\,\mu(dx)\mu(dy)\\ &\quad= \iint \langle x-y, \nabla W(x-y)\rangle\,\mu(dx)\mu(dy) + \iint \langle y, \nabla W(x-y)\rangle\,\mu(dx)\mu(dy)\\ &\quad\ge -M_W + \iint \langle y, \nabla W(x-y)\rangle\,\mu(dx)\mu(dy)\\ &\quad\ge -M_W - \iint \langle x, \nabla W(x-y)\rangle\,\mu(dx)\mu(dy),\end{aligned}$$

where in the penultimate inequality we used the assumption (which implies $\langle \nabla W(x), x\rangle \ge -M_W$), and in the last one, we used the that W is symmetric (which implies $\nabla W(-z) = -\nabla W(z)$), and used Fubini's theorem to exchange integrals. Bringing the last term to the left and dividing by 2 shows (C.1).

To show average dissipativity, it suffices to observe

$$-\int \langle x, b(x,\mu)\rangle\,\mu(dx) = \int \langle x, \nabla V(x)\rangle\,\mu(dx) + \iint \langle x, \nabla W(x-y)\rangle\,\mu(dy)\mu(dx)$$
$$\geq \alpha \int |x|^2\,\mu(dx) - \beta - M_W/2.$$

Proximal algorithm fits Assumption 4: Note that this implicit algorithm corresponds to the following proximal step

$$x^i_{k+1} = \arg\min_x \left\{ V(x) + \frac{1}{N}\sum_{j=1}^N W(x - x^j_k) + \frac{1}{2\gamma_{k+1}}\left|x - (x^i_k + \sqrt{2\gamma_{k+1}}\,\xi^i_{k+1})\right|^2 \right\}.$$

By defining the perturbation as

$$P^i_{k+1} = \varepsilon^i_{k+1} = \nabla V(x^i_{k+1}) - \nabla V(x^i_k) + \frac{1}{N}\sum_{j=1}^N (\nabla W(x^i_{k+1} - x^j_k) - \nabla W(x^i_k - x^j_k)),$$

we see that the algorithm (Kin-Prox) fits the template (SAA). For brevity, let us write $\mathscr{F}_k := \mathscr{F}_{\tau_k}$. We only have to show that

$$\mathbb{E}[|\varepsilon_{k+1}|^2 \mid \mathscr{F}_k] = \sum_{i=1}^N \mathbb{E}[|\varepsilon^i_{k+1}|^2 \mid \mathscr{F}_k] = \mathcal{O}(\gamma^2_{k+1}|\boldsymbol{b}(x_k)|^2 + \gamma_{k+1}).$$

We have

$$\begin{aligned}
\left|\varepsilon^i_{k+1}\right|^2 &= \left|\nabla V(x^i_{k+1}) - \nabla V(x^i_k) + \frac{1}{N}\sum_{j=1}^N (\nabla W(x^i_{k+1} - x^j_k) - \nabla W(x^i_k - x^j_k))\right|^2 \\
&\leq 2\left|\nabla V(x^i_{k+1}) - \nabla V(x^i_k)\right|^2 + \frac{2}{N}\sum_{j=1}^N \left|\nabla W(x^i_{k+1} - x^j_k) - \nabla W(x^i_k - x^j_k)\right|^2 \\
&\leq 2L^2\left|x^i_{k+1} - x^i_k\right|^2 + \frac{2L^2}{N}\sum_{j=1}^N \left|x^i_{k+1} - x^i_k\right|^2 \\
&= 4L^2\left|x^i_{k+1} - x^i_k\right|^2.
\end{aligned}$$

For brevity, let

$$f(x) = \nabla V(x) + \frac{1}{N}\sum_{j=1}^N \nabla W(x - x^j_k),$$

noticing that $\varepsilon^i_{k+1} = f(x^i_{k+1}) - f(x^i_k)$. By the update rule (Kin-Prox)

$$\mathbb{E}[|x^i_{k+1} - x^i_k|^2 \mid \mathscr{F}_k] \leq 2\gamma^2_{k+1}\,\mathbb{E}[|f(x^i_{k+1})|^2 \mid \mathscr{F}_k] + 4\gamma_{k+1}d.$$

Moreover, we have that $|f(x^i_{k+1})|^2 \leq 2|f(x^i_{k+1}) - f(x^i_k)|^2 + 2|f(x^i_k)|^2$. Since $\gamma_{k+1} \to 0$, we can assume that $16L^2\gamma^2_{k+1} < \frac{1}{2}$. All in all, this gives

$$\begin{aligned}
\mathbb{E}[|\varepsilon^i_{k+1}|^2 \mid \mathscr{F}_k] &\leq 4L^2\,\mathbb{E}[|x^i_{k+1} - x^i_k|^2 \mid \mathscr{F}_k] \\
&\leq 8L^2\gamma^2_{k+1}\,\mathbb{E}[|f(x^i_{k+1})|^2 \mid \mathscr{F}_k] + 16L^2\gamma_{k+1}d \\
&\leq 16L^2\gamma^2_{k+1}\,\mathbb{E}[|f(x^i_{k+1}) - f(x^i_k)|^2 \mid \mathscr{F}_k] + 16L^2\gamma^2_{k+1}|f(x^i_k)|^2 + 16L^2\gamma_{k+1}d \\
&\leq 16L^2\gamma^2_{k+1}\,\mathbb{E}[|\varepsilon^i_{k+1}|^2 \mid \mathscr{F}_k] + 16L^2\gamma^2_{k+1}|f(x^i_k)|^2 + 16L^2\gamma_{k+1}d \\
&\leq \frac{1}{2}\,\mathbb{E}[|\varepsilon^i_{k+1}|^2 \mid \mathscr{F}_k] + 16L^2\gamma^2_{k+1}|f(x^i_k)|^2 + 16L^2\gamma_{k+1}d.
\end{aligned}$$

This implies that

$$\mathbb{E}[|\varepsilon^i_{k+1}|^2 \mid \mathscr{F}_k] \leq 32L^2\gamma^2_{k+1}|f(x^i_k)|^2 + 32L^2\gamma_{k+1}d.$$

Summing over i and observing that $\sum|f(x^i_k)|^2 = |\boldsymbol{b}(x_k)|^2$ concludes the proof. ∎