A Existence and uniqueness of global minimiser

In this section, we discuss assumptions under which the global minimiser of the optimisation problem

$$L(Q) = \int \ell(\theta)\, dQ(\theta) + \lambda D(Q, P) \tag{8}$$

over $\mathcal{P}(\mathbb{R}^J)$ exists and is unique. We assume throughout that the optimisation problem is not pathological, in the sense that there exists a measure $\widehat{Q} \in \mathcal{P}(\mathbb{R}^J)$ such that $L(\widehat{Q}) < \infty$. This is in applications often trivial to verify. A good candidate for $\widehat{Q}$ is typically the reference measure P.

Loss assumptions Let $\ell : \mathbb{R}^J \to \mathbb{R}$ be a loss satisfying the following assumptions:

(L1) The loss ℓ is bounded from below which means that

$$c := \inf \left\{ \ell(\theta) : \theta \in \mathbb{R}^J \right\} > -\infty. \tag{9}$$

(L2) The loss is norm-coercive which means that

$$\ell(\theta) \to \infty \tag{10}$$

if $\|\theta\| \to \infty$.

(L3) The loss ℓ is lower semi-continuous which means that

$$\liminf_{\theta \to \theta_0} \ell(\theta) \geq \ell(\theta_0) \tag{11}$$

for all $\theta_0 \in \mathbb{R}^J$.

Regulariser assumptions Let $D : \mathcal{P}(\mathbb{R}^J) \times \mathcal{P}(\mathbb{R}^J) \to [0, \infty]$ be a regulariser and $P \in \mathcal{P}(\mathbb{R}^J)$ a reference measure. We define $D_P(\cdot) := D(\cdot, P)$ for notational convenience. We assume the following for D_P:

(D1) The function D_P is lower semi-continuous w.r.t. to the topology of weak-convergence, i.e. for all sequences $\big(Q_n\big)_{n \in \mathbb{N}} \subset \mathcal{P}(\mathbb{R}^J)$ and all Q with $D_P(Q) < \infty$, it holds that $Q_n \xrightarrow{\mathcal{D}} Q$ implies

$$\liminf_{n \to \infty} D_P(Q_n) \geq D_P(Q). \tag{12}$$

Here, $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

(D2) D_P is strictly convex, i.e. for all $Q_1 \neq Q_2 \in \mathcal{P}(\mathbb{R}^J)$ with $D_P(Q_1) < \infty$ and $D_P(Q_2) < \infty$, it holds that

$$D_P\big(\alpha Q_1 + (1 - \alpha) Q_2\big) < \alpha D_P(Q_1) + (1 - \alpha) D_P(Q_2) \tag{13}$$

with $\alpha \in (0, 1)$.

The next theorem provides an existence result for the optimisation problem (8). The result is similar in spirit to Lemma 2.1 in Knoblauch (2021) with the important difference that our assumptions are easier to verify, since they are formulated in terms of ℓ and D_P.

Theorem 3 (Existence of global minimiser)**.** *Under the assumptions (L1)-(L3) and (D1) there exists a probability measure $Q^* \in \mathcal{P}(\mathbb{R}^J)$ with*

$$L(Q^*) = \inf \big\{ L(Q) \,:\, Q \in \mathcal{P}(\mathbb{R}^J) \big\}. \tag{14}$$

Proof. Let $c > -\infty$ be the lower bound for ℓ. It follows immediately that $L(Q) \geq c$ for all $Q \in \mathcal{P}(\mathbb{R}^J)$ since $D(P, Q) \geq 0$. As a consequence we know that

$$\infty > L^* := \inf \big\{ L(Q) \,:\, Q \in \mathcal{P}(\mathbb{R}^J) \big\} \geq c > -\infty. \tag{15}$$

By definition of the infimum we can construct a sequence $l_n = L(Q_n) \in \mathbb{R}$ in the image of L such

$$l_n \to L^* \tag{16}$$

for $n \to \infty$. We now show by contradiction that the corresponding sequence $\big(Q_n\big) \subset \mathcal{P}(\mathbb{R}^J)$ is *tight*[3]. Assume that $\big(Q_n\big)$ is not tight. By definition we can then find an $\epsilon > 0$ such that for each $k \in \mathbb{N}$ there exists $n = n_k \in \mathbb{N}$ with $Q_{n_k}([-k,k]^J) \le 1-\epsilon$. We set $A_k := [-k,k]^J \subset \mathbb{R}^J$ and obtain

$$
\begin{aligned}
l_{n_k} &= L(Q_{n_k}) && (17)\\
&= \int_{A_k} \ell(\theta)\, dQ_{n_k}(\theta) + \int_{\mathbb{R}^J\setminus A_k} \ell(\theta)\, dQ_{n_k}(\theta) + \lambda D(Q,P) && (18)\\
&\ge \int_{A_k} \ell(\theta)\, dQ_{n_k}(\theta) + \int_{\mathbb{R}^J\setminus A_k} \ell(\theta)\, dQ_{n_k}(\theta) && (19)\\
&\ge cQ_{n_k}(A_k) + \inf\big\{\ell(\theta) \,:\, \theta \in\, \mathbb{R}^J\setminus A_k\big\} Q_{n_k}(\mathbb{R}^J\setminus A_k) && (20)\\
&\ge cQ_{n_k}(A_k) + \epsilon \inf\big\{\ell(\theta) \,:\, \theta \in\, \mathbb{R}^J\setminus A_k\big\}. && (21)
\end{aligned}
$$

Due to the coerciveness of ℓ, we know that $\inf\big\{\ell(\theta) \,:\, \theta \in\, \mathbb{R}^J\setminus A_k\big\} \to \infty$ for $k \to \infty$ and therefore $l_{n_k} \to \infty$ for $k \to \infty$. However, this is a contradiction: The sequence (l_n) is convergent and therefore in particular bounded. As a consequence, it cannot contain the unbounded sub-sequence (l_{n_k}). It follows that the sequence (Q_n) is tight. By Prokhorov's theorem we can now extract a sub sequence (Q_{n_k}) of (Q_n) and a measure $Q^* \in \mathcal{P}(\mathbb{R}^J)$ such that

$$Q_{n_k} \xrightarrow{\mathcal{D}} Q^* \tag{22}$$

for $k \to \infty$. Due to Lemma 5.1.7 in Ambrosio et al. (2005) the lower semi-continuity of ℓ implies that $Q \mapsto \int \ell(\theta)\, dQ(\theta)$ is lower semi-continuous. This combined with the lower semi-continuity of D_P gives

$$\liminf_{k\to\infty} L(Q_{n_k}) \ge L(Q^*). \tag{23}$$

From this it immediately follows that

$$L(Q^*) \le \liminf_{k\to\infty} L(Q_{n_k}) = L^*, \tag{24}$$

but by definition L^* is the global minimum of L which implies $L^* \le L(Q^*)$. We therefore conclude that $L(Q^*) = L^*$. □

Theorem 3 only shows the existence of a global minimiser. In order to show uniqueness we use the convexity assumption (D2). The proof is the same as in finite dimensions and only included for completeness.

Theorem 4 (Uniqueness of global minimiser)**.** *Assume that (D2) holds. Then, the global minimiser of L is unique (whenever it exists).*

Proof. Assume there exits two probability measures $Q_1, Q_2 \in \mathcal{P}(\mathbb{R}^J)$ such that

$$L(Q_1) = L^* = L(Q_2). \tag{25}$$

where $\infty > L^* := \inf\big\{L(Q) \,:\, Q \in \mathcal{P}(\mathbb{R}^J)\big\} > -\infty$. We define the probability measure $Q_3 := \frac{1}{2}Q_1 + \frac{1}{2}Q_3$. By strict convexity we obtain

$$L(Q_3) < \frac{1}{2}L(Q_1) + \frac{1}{2}L(Q_3) = L^*, \tag{26}$$

which is a contradiction to Q_1 and Q_2 being global minimisers. □

Note that in the literature on GVI (Knoblauch et al., 2022) it is common to assume that the regulariser is definite, i.e.

$$D(P,Q) = 0 \Longleftrightarrow P = Q \tag{27}$$

for all $P, Q \in \mathcal{P}(\mathbb{R}^J)$. We did not use this assumption in neither Theorem 3 nor Theorem 4. However, the next lemma shows that it is basically implied by strict convexity.

[3]A sequence of probability measures (Q_n) is called tight if and only if for every $\epsilon > 0$ there exists a compact set $K \in \mathbb{R}^J$ such that for all $n \in \mathbb{N}$ holds: $Q_n(K) > 1-\epsilon$.

Lemma 1. *Let* $D_P : \mathcal{P}(\mathbb{R}^J) \to [0,\infty]$ *be strictly convex and assume further* $D(Q,Q) = 0$ *for all* $Q \in \mathcal{P}(\mathbb{R}^J)$. *Then it follows that* $D(Q,P) = 0$ *implies* $P = Q$.

Proof. We prove the claim by contradiction. Assume that there exists $P \neq Q$ such that $D(P,Q) = 0$. The strict convexity and $D(P,P) = 0$ imply combined that

$$D(\frac{1}{2}P + \frac{1}{2}Q, P) < \frac{1}{2}D(P,P) + \frac{1}{2}D(Q,P) \tag{28}$$

$$= 0. \tag{29}$$

However, we know that $D(\frac{1}{2}P + \frac{1}{2}Q, P) \geq 0$ by assumption. This is a contradiction. $\square$

Discussion on loss assumptions The assumptions on the loss ℓ in (L1) and (L3) are rather weak. Typically loss functions in machine learning are bounded from below and continuous (and therefore in particular lower semi-continuous). However, norm-coercivity can be violated. Consider for example the squared loss

$$\ell(\theta) := \sum_{n=1}^{N} \big(y_n - f_\theta(x_n)\big)^2, \tag{30}$$

where f_θ is the parametrisation of a neural network with one hidden layer, i.e. $\theta = (w, A)$ and

$$f_\theta(x) = w^T \sigma(Ax), \tag{31}$$

where $\sigma : \mathbb{R} \to \mathbb{R}$ is an activation function which is applied pointwise to the vector Ax and has the property that $\sigma(0) = 0$. It is now possible to find a sequence of parameters $(\theta_k)_{k\in\mathbb{N}} \subset \mathbb{R}^J$ with $\|\theta_k\| \to \infty$ such that $\ell(\theta_k)$ does not converge to infinity. Define $w_k := k\big(1\dots1\big)$, $A_k := 0$ and $\theta_k = (w_k\ A_k)$ for $k \in \mathbb{N}$. Then we obviously have that

$$\|\theta_k\| = \|w_k\| \to \infty \tag{32}$$

for $k \to \infty$ but

$$\ell(\theta_k) = \sum_{n=1}^{N} \big(y_n - f_{\theta_k}(x_n)\big)^2 \tag{33}$$

$$= \sum_{n=1}^{N} \big(y_n - w^T\sigma(0)\big)^2 \tag{34}$$

$$= \sum_{n=1}^{N} y_n^2, \tag{35}$$

which is constant and therefore does not converge to ∞. A similar, but notationally more involved, construction can be made for neural networks with more than one hidden layer. However, this is an issue that can be easily resolved by adding what is known as weight decay to the loss. For example, consider for $\gamma > 0$ the loss

$$\ell(\theta) := \sum_{n=1}^{N} \big(y_n - f_\theta(x_n)\big)^2 + \gamma\|\theta\|^2 \tag{36}$$

with weight decay. This loss is by construction norm-coercive and therefore the previous existence proof applies.

Discussion on regulariser assumptions The assumptions (D1) and (D2) are quite weak. The KL-divergence for example is known to be lower semi-continuous (Polyanskiy and Wu, 2014, Theorem 3.7) and strictly convex (Polyanskiy and Wu, 2014, Theorem 4.1). This immediately implies lower semi-continuity and convexity of $\mathrm{KL}(\cdot, P)$ for any fixed P. The MMD is also known to be strictly convex (Arbel et al., 2019, Lemma 25), whenever it is well-defined, which can be guaranteed under weak assumptions on κ (Muandet et al., 2017, Lemma 3.1). The lower semi-continuity properties also depend on the kernel κ. However, for bounded kernels it is trivial to verify. We include the proof for completeness, but assume this has been shown before elsewhere.

Lemma 2. *Let the kernel* $\kappa : \mathbb{R}^J \times \mathbb{R}^J$ *be continuous and bounded:* $\|\kappa\|_\infty := \sup_{\theta,\theta' \in \mathbb{R}^J} |k(\theta, \theta')| < \infty$ *and* P *be fixed. Then* $\mathrm{MMD}(\cdot, P)$ *is continuous and therefore, in particular, lower semi-continuous.*

Proof. Let $(Q_n)_{n\in\mathbb{N}}$ and Q^* be such that

$$Q_n \xrightarrow{\mathcal{D}} Q^* \tag{37}$$

for $n \to \infty$. This immediately implies that

$$Q_n \otimes Q_n \xrightarrow{\mathcal{D}} Q^* \otimes Q^* \tag{38}$$

for $n \to \infty$, where $Q^* \otimes Q^*$ denotes the product measure of Q^* with itself. Further, note that the kernel mean embedding μ_P is continuous as integral with respect to the second component of a continuous function and bounded since

$$|\mu_P(\theta)| = |\int \kappa(\theta, \theta')\, dP(\theta')| \tag{39}$$

$$\leq \int |\kappa(\theta, \theta')| dP(\theta') \tag{40}$$

$$\leq \|\kappa\|_\infty. \tag{41}$$

By the definition of weak convergence for measures, we therefore have

$$\iint \kappa(\theta, \theta')\, d(Q_n \otimes Q_n)(\theta, \theta') \longrightarrow \iint \kappa(\theta, \theta')\, d(Q^* \otimes Q^*)(\theta, \theta') \tag{42}$$

$$\int \mu_P(\theta)\, dQ_n(\theta) \longrightarrow \int \mu_P(\theta)\, dQ^*(\theta) \tag{43}$$

for $n \to \infty$. This immediately implies continuity of $\mathrm{MMD}(\cdot, P)$ with respect to the topology of weak convergence. □

Notice that most kernels common in machine learning, such as the squared exponential or the Matérn kernel, are continuous and bounded and therefore Lemma 2 applies.

Remark 1. The astute reader may have noticed that our existence proof only guarantees the existence of measure $Q^* \in \mathcal{P}(\mathbb{R}^J)$. However, the Wasserstein gradient flow is by definition only formulated in the space of probability measures with finite second moment, denoted $\mathcal{P}_2(\mathbb{R}^J)$. Assumptions which guarantee that $Q^* \in \mathcal{P}_2(\mathbb{R}^J)$ are easy to formulate. For example, we can require that there exists $C > 0$ and $R > 0$ such that the loss ℓ satisfies

$$|\ell(\theta)| > C\|\theta\|^2 \tag{44}$$

for all $\|\theta\| > R$. This immediately implies that $Q^* \in \mathcal{P}_2(\mathbb{R}^J)$ since otherwise

$$\int |\ell(\theta)|\, dQ^*(\theta) = \infty \tag{45}$$

gives a contradiction to the finiteness of $L(Q^*)$. However, even if (44) is violated, the reference measure P may still guarantee that $Q^* \in \mathcal{P}_2(\mathbb{R}^J)$. For example, if $P \in \mathcal{P}_2(\mathbb{R}^J)$, then $D_P(Q^*)$ will typically be large if $Q^* \notin \mathcal{P}_2(\mathbb{R}^J)$ and the global minimiser is therefore in a sense *unlikely* to have fat tails. We therefore assume $Q^* \in \mathcal{P}_2(\mathbb{R}^J)$ throughout the paper and consider it to be a minor practical concern.

B Realising the Wasserstein gradient flow

In this section, we identify a suitable stochastic process that allows us to follow the WGF.

Let $L^{\text{fe}} : \mathcal{P}(\mathbb{R}^J) \to (-\infty, \infty]$ be the free energy discussed in Section 3.2 given as

$$L^{\text{fe}}(Q) := \int V(\theta)\, dQ(\theta) + \frac{\lambda_1}{2} \int \kappa(\theta, \theta')\, dQ(\theta) dQ(\theta') + \lambda_2 \int \log\big(q(\theta)\big) q(\theta)\, d\theta, \tag{46}$$

where $\lambda_1, \lambda_2 \geq 0$ are constants, $V : \mathbb{R}^J \to \mathbb{R}$ is the potential, $\kappa : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}$ is symmetric. We will write L for L^{fe} from now on to simplify notation. The Wasserstein gradient of L is given as (cf. Chapter 9.1 [Villani, 2003, Equation 9.4)

$$\nabla_W L[Q](\theta) = \nabla V(\theta) + \lambda_1 (\nabla_1 \kappa * Q)(\theta) + \lambda_2 \nabla \log \big(q(\theta)\big), \tag{47}$$

where $\nabla_1 \kappa : \mathbb{R}^J \times \mathbb{R}^J \to \mathbb{R}^J$ is the (vector-valued) derivative of κ with respect to the first component, ∇ denotes the euclidean gradient with respect to θ and $(\nabla_1 \kappa * Q)(\theta) := \int \nabla_1 \kappa(\theta, \theta')\, dQ(\theta')$ for $\theta \in \mathbb{R}^J$. The corresponding Wasserstein gradient flow is therefore given as (cf. Chapter 9.1 Villani, 2003, Equation 9.3)

$$\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta) \big(\nabla V(\theta) + \lambda_1 (\nabla_1 \kappa * Q)(\theta) + \lambda_2 \nabla \log \big(q_t(\theta) \big) \big) \Big). \tag{48}$$

In general the probability density evolution of a stochastic process is—via the Fokker-Planck equation—associated with the adjoint of the (infinitesimal) generator of the stochastic process. We will therefore try to identify the generator associated to the density evolution in (48). To this end let $h \in C_c^2(\mathbb{R}^J, \mathbb{R})$ where $C_c^2(\mathbb{R}^J, \mathbb{R})$ denotes the space of twice continuously differentiable functions with compact support. We multiply both sides of (48) with h, integrate, and apply the partial integration rule to obtain

$$\frac{d}{dt} \int h(\theta) q(t,\theta)\, d\theta = -\int \nabla_W L\big[Q(t)\big](\theta) \cdot \nabla h(\theta)\, q(t,\theta)\, d\theta. \tag{49}$$

$$= -\int \big(\nabla V(\theta) + \lambda_1 (\nabla_1 \kappa * Q_t)(\theta) \big) \cdot \nabla h(\theta)\, dQ_t(\theta) \tag{50}$$

$$- \lambda_2 \int \nabla \log \big(q_t(\theta) \big) \cdot \nabla h(\theta)\, dQ_t(\theta). \tag{51}$$

By chain-rule and partial integration, (51) can be rewritten as

$$-\lambda_2 \int \nabla \log \big(q_t(\theta) \big) \cdot \nabla h(\theta)\, dQ_t(\theta) = -\lambda_2 \int \nabla q_t(\theta) \cdot \nabla h(\theta)\, d\theta \tag{52}$$

$$= \lambda_2 \int \Delta h(\theta)\, dQ_t(\theta). \tag{53}$$

Putting everything together, we obtain

$$\frac{d}{dt} \int h(\theta) q(t,\theta)\, d\theta = \int \big(A[Q(t)]h \big)(\theta)\, dQ_t(\theta), \tag{54}$$

where $\big\{A[Q]\big\}_{Q \in \mathcal{P}(\mathbb{R}^J)}$ is a family of operators defined as

$$\big(A[Q]h \big)(\theta) := -\Big(\nabla V(\theta) + \lambda_1 (\nabla_1 \kappa * Q)(\theta) \Big) \cdot \nabla h(\theta) + \lambda_2 \Delta h. \tag{55}$$

for $h \in C_c^2(\mathbb{R}^J, \mathbb{R})$. The reader may recognize this operator family as the generator of a so called *nonlinear Markov processes* (Kolokoltsov, 2010, Chapter 1.4). The nonlinearity in this case refers to the dependency on the measure Q. Linear Markov processes have no measure-dependency. This family of generators corresponds to a McKean-Vlasov process of the form

$$d\theta(t) = -\Big(\nabla V(\theta(t)) + \lambda_1 (\nabla_1 \kappa * Q_t)(\theta(t)) \Big) dt + \sqrt{2\lambda_2}\, dB(t), \tag{56}$$

where $\big(B(t)\big)_{t>0}$ is a Brownian motion and Q_t the law of $\theta(t)$. In other words: The solution to (56) has the time marginals $Q(t)$ such that (54) holds for every $h \in C_c^2(\mathbb{R}^J, \mathbb{R})$. Furthermore, the corresponding pdfs $\big(q(t)\big)$ satisfy the nonlinear Fokker-Planck equation given as

$$\partial_t q_t = A^*[Q_t] q_t, \tag{57}$$

where $A^*[Q]$ denotes the L^2-adjoint of the operator $A[Q]$ and is given as

$$\big(A^*[Q]h \big)(\theta) = \nabla \cdot \Big(h(\theta) \big(\nabla V(\theta) + \lambda_1 (\nabla_1 \kappa * Q)(\theta) + \lambda_2 \nabla \log \big(h(\theta) \big) \big) \Big) \tag{58}$$

for $h \in C_c^2(\mathbb{R}^J, \mathbb{R})$(Barbu and Röckner, 2020, cf. equation (1.1)-(1.4)). Note that (57) corresponds exactly to the Wasserstein gradient flow equation in (48). We can therefore follow the WGF by simulating solutions to (56).

The standard approach to simulate solutions to (56) (Veretennikov, 2006) is to use an ensemble of interacting particles. Formally, we replace $Q(t)$ by $\frac{1}{N_E}\sum_{n=1}^{N_E}\delta_{\theta_n(t)}$ and obtain

$$d\theta_n(t) = -\Big(\nabla V\big(\theta_n(t)\big) + \frac{\lambda_1}{N_E}\sum_{j=1}^{N_E}(\nabla_1\kappa)\big(\theta_n(t), \theta_j(t)\big)\Big)dt + \sqrt{2\lambda_2}dB_n(t) \tag{59}$$

for $n = 1, \dots, N_E$ where $N_E \in \mathbb{N}$ denotes the number of particles. The Euler-Maruyama approximation of (59) leads to the final algorithm:

Step 1: Initialise $N_E \in \mathbb{N}$ particles $\theta_{1,0}, \dots, \theta_{N_E,0}$ from a use chosen initial distribution Q_0.

Step 2: Evolve the particles forward in time according to

$$\theta_{n,k+1} = \theta_{n,k} - \eta\Big(\nabla V\big(\theta_{n,k}\big) + \frac{\lambda_1}{N_E}\sum_{j=1}^{N_E}(\nabla_1\kappa)\big(\theta_{n,k}, \theta_{j,k}\big)\Big) + \sqrt{2\eta\lambda_2}Z_{n,k} \tag{60}$$

for $n = 1, \dots, N_E$, $k = 0, \dots, T-1$ with $Z_{n,k} \sim \mathcal{N}(0, I_{J\times J})$.

Note that $\theta_{n,k}$ is thought of as approximation of $\theta_n(t)$ at position $t = k\eta$. Furthermore, as discussed in Section 4, various choices of V, λ_1 and λ_2 allow us to implement the WGF for different regularised optimisation problems in the space of probability measures. This is summarised below:

- Deep ensembles: $V(\theta) = \ell(\theta)$, $\lambda_1 = 0$, $\lambda_2 = 0$
- Deep Langevin ensembles: $V(\theta) = \ell(\theta) - \lambda \log p(\theta)$, $\lambda_1 := 0$, $\lambda := \lambda_2$
- Deep repulsive Langevin ensembles: $V(\theta) = \ell(\theta) - \lambda_1 \log p(\theta) - \lambda_2 \mu_P(\theta)$

C Asymptotic distribution of particles: unregularised objective

In this section, we investigate the asymptotic distribution of the WGF for the objective

$$L(Q) := \int \ell(\theta)\, dQ(\theta) \tag{61}$$

for $Q \in \mathcal{P}(\mathbb{R}^J)$. The associated particle method is:

- Sample $\theta_1(0), \dots, \theta_{N_E}(0)$ independently from Q_0.
- Simulate (deterministically) $\theta_n'(t) = -\nabla\ell\big(\theta_n(t)\big)$ for $n = 1, \dots, N_E$.

We start by introducing some notation for the deterministic gradient system. Let $\phi^t(\theta_0)$ denote the solution to the ordinary differential equation (ODE)

$$\theta(0) = \theta_0 \in \mathbb{R}^J \tag{62}$$
$$\theta'(t) = -\nabla\ell\big(\theta(t)\big) \tag{63}$$

at time $t > 0$. In a first step, we show the following lemma, which is a simple application of the famous Lojasiewicz theorem (Colding and Minicozzi II, 2014), and the fact that Lebesgue almost every initialisation leads to a local minimum (Lee et al., 2016).

Lemma 3. *Assume $\ell : \mathbb{R}^J \to \mathbb{R}$ is norm-coercive and satisfies the Lojasiewicz inequality, i.e. for every $\theta \in \mathbb{R}^J$ exists an environment U of θ and constants $0 < \gamma < 1$ and $C > 0$ such that*

$$|\ell(\theta) - \ell(\bar\theta)|^\gamma < C|\nabla\ell(\theta)|. \tag{64}$$

for all $\bar\theta \in U$. Then we know that $\phi^t(\theta_0)$ converges for $t \to \infty$ to a local minimum of ℓ for Lebesgue almost every $\theta_0 \in \mathbb{R}^J$.

Proof. First we show that $t \mapsto \phi^t(\theta_0)$ is bounded. We proof this by contradiction. Assume that $\phi^t(\theta_0)$ is unbounded. Then there exists a subsequence $(t_n)_{n\in\mathbb{N}} \subset [0,\infty)$ with $t_n \to \infty$ for $n \to \infty$ such that

$$|\phi^{t_n}(\theta_0)| \to \infty \tag{65}$$

for $n \to \infty$. The norm-coercivity immediately implies that

$$\ell\big(\phi^{t_n}(\theta_0)\big) \to \infty \tag{66}$$

for $n \to \infty$. However, this contradicts

$$\ell\big(\phi^{t}(\theta_0)\big) \leq \ell\big(\phi^{0}(\theta_0)\big) = \ell(\theta_0) < \infty, \tag{67}$$

where the first inequality follows from the fact that $t \mapsto \ell\big(\phi^t(\theta_0)\big)$ is decreasing, which is a consequence of

$$\frac{d}{dt}\ell\big(\phi^t(\theta_0)\big) = \nabla\ell\big(\phi^t(\theta_0)\big)\frac{d}{dt}\phi^t(\theta_0) \tag{68}$$

$$= -|\nabla\ell\big(\phi^t(\theta_0)\big)|^2 \leq 0. \tag{69}$$

Hence $t \mapsto \phi^t(\theta_0)$ is bounded. By the Bolzano-Weierstrass theorem we can find a sequence $(t_n)_{n\in\mathbb{N}} \subset [0,\infty)$ with $t_n \to \infty$ and a point $\theta_\infty \in \mathbb{R}^J$ such that

$$\phi^{t_n}(\theta_0) \to \theta_\infty \tag{70}$$

for $n \to \infty$. Hence $\big(\phi^t(\theta_0)\big)_{t>0}$ has the accumulation point θ_∞. The Lojasiewicz theorem (Colding and Minicozzi II, 2014) allows us to deduce that

$$\phi^{t}(\theta_0) \to \theta_\infty \tag{71}$$

for $t \to \infty$, and that θ_∞ satisfies $\nabla\ell(\theta_\infty) = 0$.

It remains to show that θ_∞ is not a saddle point for Lebesgue almost every initial value θ_0. However, this is very similar to the proof in Lee et al. (2016). The only difference is that one would need to use a continuous-time version of the stable manifold theorem, which is readily available, for example in Bressan (2003). □

Let $\{m_i\}_{i\in\mathbb{N}}$ denote the local minima of ℓ which are by assumption countable. Denote further by

$$\Theta_i := \big\{\theta_0 \in \mathbb{R}^J \, : \, \lim_{t\to\infty} \phi^t(\theta_0) \to m_i\big\} \tag{72}$$

the domain of attraction for the minimum m_i. The next theorem is then an easy consequence of Lemma 3.

Theorem 5. *Assume that the loss function ℓ only has countably many local minima, is norm coercive, and satisfies the Lojasiewicz inequality. Let further $\theta_0 \sim Q_0$ for some $Q_0 \in \mathcal{P}(\mathbb{R}^J)$ such that $\sum_{i=1}^{\infty} Q_0(\Theta_i) = 1$. Then,*

$$\phi^t(\theta_0) \xrightarrow{\mathcal{D}} \sum_{i=1}^{\infty} Q_0(\Theta_i)\, \delta_{m_i} =: Q_\infty \tag{73}$$

for $t \to \infty$. Here $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution.

Proof. Let $\theta_0 \in \mathbb{R}^J$ be fixed. Due to Lemma 3, we know that

$$\phi^t(\theta_0) \to \sum_{i=1}^{\infty} m_i \mathbb{1}\{\theta_0 \in \Theta_i\} \tag{74}$$

for Lebesgue almost every θ_0 for $t \to \infty$. Here, $\mathbb{1}\{\cdot\}$ denotes the indicator function. Let Y now be a random variable with law Q_0. By assumption, we know that $Y \in \Theta_i$ for some $i \in \mathbb{N}$ with probability 1. Hence,

$$\phi^t(Y) \to \sum_{i=1}^{\infty} m_i \mathbb{1}\{Y \in \Theta_i\} \tag{75}$$

almost surely for $t \to \infty$. Since almost sure convergence implies convergence in distribution, we conclude that

$$\phi^t(Y) \xrightarrow{\mathcal{D}} \mathcal{L}\big(\sum_{i=1}^{\infty} m_i \mathbb{1}\{Y \in \Theta_i\}\big), \tag{76}$$

where $\mathcal{L}(\cdot)$ denotes the law of a random variable. However, the law of the RHS is easily recognised as

$$\mathcal{L}\big(\sum_{i=1}^{\infty} m_i \mathbb{1}\{Y \in \Theta_i\}\big) = \sum_{i=1}^{\infty} Q_0(\Theta_i)\delta_{m_i}, \tag{77}$$

which concludes the proof. $\square$

Remark 2. Note that the condition

$$\sum_{i=1}^{\infty} Q_0(\Theta_i) = 1 \tag{78}$$

in Theorem 5 is easy to satisfy. According to Lemma 3 the set

$$\mathbb{R}^J \setminus \bigcup_{i=1}^{n} \Theta_i \tag{79}$$

has Lebesgue measure zero. Therefore, any Q_0 which has a density w.r.t. the Lebesgue measure will satisfy (78).

D Asymptotic distribution for deep Langevin ensembles

In this section, we analyse the objective

$$L(Q) := \int \ell(\theta)\, dQ(\theta) + \lambda\, \mathrm{KL}(Q, P) \tag{80}$$

for $Q \in \mathcal{P}(\mathbb{R}^J)$. The corresponding particle method is given as:

- Sample $\theta_1(0), \ldots, \theta_{N_E}(0)$ independently from Q_0.
- Simulate the SDE $d\theta_n(t) = -\nabla V\big(\theta_n(t)\big)dt + \sqrt{2\lambda}dB_n(t)$ for each $n = 1, \ldots, N_E$.

Recall that $V(\theta) = \ell(\theta) - \lambda \log p(\theta)$. This case is well-studied in the literature and known as Langevin diffusion. Under mild assumptions (Chiang et al., 1987; Roberts and Tweedie, 1996),

$$\theta_n(t) \xrightarrow{\mathcal{D}} Q_\infty \tag{81}$$

for $t \to \infty$ and each particle $n = 1, \ldots, N_E$ independently. The probability measure Q_∞ has the density

$$q_\infty(\theta) = \frac{1}{Z} \exp\big(-\frac{V(\theta)}{\lambda}\big) \tag{82}$$

$$= \frac{1}{Z} \exp\big(-\frac{\ell(\theta)}{\lambda}\big)p(\theta), \tag{83}$$

where $Z > 0$ is the normalising constant. As a consequence, the WGF asymptotically produces samples from Q_∞. However, it is a priori unclear that Q_∞ is in fact the same as the global minimiser Q^* of L.

We investigate this question by relating invariant measures to stationary points of the Wasserstein gradient.

Definition 1. (Liggett, 2010, Thm. 3.3.7) A measure Q is called an invariant measure (for a given Feller-process) if

$$\int Ah(\theta)\, dQ(\theta) = 0 \tag{84}$$

for all $h \in C_c^2(\mathbb{R}^J)$. Here A is the infinitesimal generator of the corresponding Feller-process.

Recall that the infinitesimal generator of the Langevin diffusion for $h \in C_c^2(\mathbb{R}^J)$ is given as

$$Ah = -\nabla V \cdot \nabla h + \lambda \Delta h. \tag{85}$$

Definition 2. A measure $Q \in \mathcal{P}_2(\mathbb{R}^J)$ is called a stationary point of the Wasserstein gradient if

$$\nabla_W L[Q](\theta) = 0 \tag{86}$$

for $Q-$almost every $\theta \in \mathbb{R}^J$.

In finite dimensions, it is well-known that a local minimiser is a stationary point of the gradient. This carries over to the infinite-dimensional case, with a similar proof. Since we could not find this result anywhere in the literature we included it for completeness.

Lemma 4. *Let $\widehat{Q}$ be a local minimiser of L, i.e. there exits and $\epsilon > 0$ such that*

$$L(\widehat{Q}) \leq L(Q) \tag{87}$$

for all Q with $W_2(\widehat{Q}, Q) \leq \epsilon$. Then $\widehat{Q}$ is a stationary point of the Wasserstein gradient in the sense of Definition 2.

Proof. Let $h \in C_c^2(\mathbb{R}^J)$ be arbitrary and $\widehat{Q} \in \mathcal{P}_2(\mathbb{R}^J)$ be a local minimum of L. Further, let $\phi^t(\theta_0)$ be the solution to the initial value problem

$$\theta(0) = \theta_0 \tag{88}$$

$$\theta'(t) = \nabla h(\theta(t)) \tag{89}$$

for $t \in (-\epsilon, \epsilon)$ for some $\epsilon > 0$. We now define $Q(t) := \phi^t \# \widehat{Q}$ for $t \in (-\epsilon, \epsilon)$ where $f\#\mu$ denotes the push-forward of the measures μ through the function f. In the Riemannian interpretation of the Wasserstein space, $\big(Q(t)\big)_{t\in(-\epsilon,\epsilon)}$ is a curve in $\mathcal{P}_2(\mathbb{R}^J)$ with tangent vector h at point $\widehat{Q}$ (Ambrosio et al., 2005, Chapter 8). We, further, define $f : (-\epsilon, \epsilon) \to \mathbb{R}$ as $f(t) := L(Q(t))$. Application of the chain-rule (Ambrosio et al., 2005, p. 233) gives

$$f'(0) = \frac{d}{dt} L(Q(t))\big|_{t=0} \tag{90}$$

$$= \langle \nabla_W L[Q(0)], \nabla h \rangle_{L^2(Q(0))} \tag{91}$$

$$= \int \nabla_W L[\widehat{Q}](\theta) \cdot \nabla h(\theta)\, d\widehat{Q}(\theta). \tag{92}$$

We know that f has a local minimum at $t = 0$ and, therefore, $f'(0) = 0$ which gives

$$0 = \int \nabla_W L[\widehat{Q}](\theta) \cdot \nabla h(\theta)\, d\widehat{Q}(\theta). \tag{93}$$

Since (93) holds for arbitrary test functions $h \in C_c^2(\mathbb{R}^J)$ and as $C_c^2(\mathbb{R}^J)$ is dense in $L^2(\widehat{Q})$, we obtain that $\nabla_W L[\widehat{Q}](\theta) = 0$ for $\widehat{Q}$-a.e $\theta \in \mathbb{R}^J$. □

The next lemma relates invariant measures and stationary points of the Wasserstein gradient for infinitesimal generators of the form (85). It will prove extremely useful to translate between the Langevin diffusion literature and our optimisation perspective.

Lemma 5. *Let $Q \in \mathcal{P}_2(\mathbb{R}^J)$ be such that Q has a density q with respect to the Lebesgue measure. Then, the following two statements are equivalent:*

- *Q is a stationary point of the Wasserstein gradient.*
- *Q is an invariant measure.*

Proof. Let Q be a measure with density q. Recall that the generator of the Langevin diffusion is for $h \in C_c^2(\mathbb{R}^J)$ given as

$$Ah = -\nabla V \cdot \nabla h + \lambda \Delta h. \tag{94}$$

By partial integration, it is easy to verify that the L^2- adjoint (w.r.t the Lebesgue measure) is given as

$$A^* h = \nabla \cdot (h \cdot \nabla V) + \lambda \Delta h. \tag{95}$$

We, therefore, conclude that

$$\begin{aligned} \int Ah(\theta)\, dQ(\theta) &= \int Ah(\theta) q(\theta)\, d\theta && (96) \\ &= \int h(\theta) A^* q(\theta)\, d\theta && (97) \\ &= \int h(\theta) \Big(\nabla \cdot (q(\theta) \cdot \nabla V(\theta)) + \lambda \Delta q(\theta) \Big)\, d\theta. && (98) \end{aligned}$$

Furthermore, we have $\nabla_W L[Q] = \nabla V + \lambda \nabla \log q$, and therefore

$$\begin{aligned} \int \nabla_W L[Q](\theta) \cdot \nabla h(\theta)\, dQ(\theta) &= \int \nabla_W L[Q](\theta) \cdot \nabla h(\theta) q(\theta)\, d\theta && (99) \\ &= \int \Big(\nabla V(\theta) q(\theta) + \lambda \nabla q(\theta) \Big) \cdot \nabla h(\theta)\, d\theta && (100) \\ &= -\int h(\theta) \Big(\nabla \cdot (q(\theta) \nabla V(\theta)) + \lambda \Delta q(\theta) \Big)\, d\theta, && (101) \end{aligned}$$

where the last line follows from applying partial integration. This allows us to conclude that

$$\int Ah(\theta)\, dQ(\theta) = -\int \nabla_W L[Q](\theta) \cdot \nabla h(\theta)\, dQ(\theta) \tag{102}$$

whenever Q has a density. As a consequence we have that Q is invariant if and only if it is a stationary point of the Wasserstein gradient. $\square$

Lemma 5 allows us to move between the optimisation and stochastic differential equation perspective. In Appendix A, we discussed the existence and uniqueness of a global minimiser Q^* of L. We know that Q^* has a density since the Kullback-Leibler divergence would be infinite otherwise (assuming P has a Lebesgue-density which we assume throughout the paper). Lemma 4 guarantees that Q^* is a stationary point of the Wasserstein gradient. Due to Lemma 5, we can infer that Q^* must be an invariant measure. However, due to the uniqueness of the invariant measure under the previously mentioned mild assumptions (Chiang et al., 1987; Roberts and Tweedie, 1996), we can conclude that $Q^* = Q_\infty$.

E Asymptotic distribution of deep repulsive Langevin ensembles

In this section, we consider

$$\begin{aligned} L(Q) &= \int \ell(\theta)\, dQ(\theta) + \frac{\lambda_1}{2} \operatorname{MMD}(Q, P)^2 + \lambda_2 \operatorname{KL}(Q, P) && (103) \\ &= \int V(\theta)\, dQ(\theta) + \frac{\lambda_1}{2} \int \kappa(\theta, \theta')\, dQ(\theta) dQ(\theta') - \lambda_2 H(Q) + const, && (104) \end{aligned}$$

as optimisation objective. Here, $H(Q) = -\int \log q(\theta) q(\theta)\, d\theta$ denotes the differential entropy.

Recall that in this case $V(\theta) = \ell(\theta) - \lambda_1 \mu_P(\theta) - \lambda_2 \log p(\theta)$. We already discussed in Appendix B that the McKean-Vlasov process of the form

$$\begin{aligned} &\theta(0) \sim Q_0 && (105) \\ &d\theta(t) = -\Big(\nabla V(\theta(t)) + \lambda_1 (\nabla_1 \kappa * Q_t)(\theta(t)) \Big) dt + \sqrt{2\lambda_2} dB(t), && (106) \end{aligned}$$

with $\big(B(t)\big)_{t \geq 0}$ being a Brownian motion achieves the desired density evolution. Furthermore, the particle approximation of (105) is given as

$$d\theta_n(t) = -\Big(\nabla V\big(\theta_n(t)\big) + \frac{\lambda_1}{N_E} \sum_{j=1}^{N_E} (\nabla_1 \kappa)\big(\theta_n(t), \theta_j(t)\big) \Big) dt + \sqrt{2\lambda_2} dB_n(t) \tag{107}$$

for $n = 1, \dots, N_E$ where $N_E \in \mathbb{N}$ denotes the number of particles.

The approach follows the same procedure as in Appendix D. We show the notions of invariant measures and stationary points of the Wasserstein gradient are the same for measures with Lebesgue density. We start by introducing the concept of an invariant measure for a nonlinear Markov process (Ahmed and Ding, 1993, Definition 1).

Definition 3. A measure Q is called an invariant measure for a nonlinear Markov process with the family of infinitesimal generators $\{A[Q]\}_{Q\in\mathcal{P}(\mathbb{R}^J)}$ if

$$\int A[Q]h(\theta)\, dQ(\theta) = 0 \tag{108}$$

for all $h \in C_c^2(\mathbb{R}^J)$.

Recall that the family of infinitesimal generators in our case is given as

$$\big(A[Q]h\big)(\theta) := -\Big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta)\Big)\cdot\nabla h(\theta) + \lambda_2\Delta h. \tag{109}$$

for $h \in C_c^2(\mathbb{R}^J, \mathbb{R})$. In analogy to Lemma 5, we obtain the following result.

Lemma 6. *Let $Q \in \mathcal{P}_2(\mathbb{R}^J)$ be such that Q has a density q with respect to the Lebesgue measure. Then, the following two statements are equivalent:*

- *Q is a stationary point of the Wasserstein gradient for L in (103) in the sense of Def. 2.*
- *Q is an invariant measure for the McKean-Vlasov process with infinitesimal generator defined in (109)*

Proof. First, we notice that

$$\int A[Q]h(\theta)dQ(\theta) = \int A[Q]h(\theta)q(\theta)\, d\theta \tag{110}$$

$$= \int h(\theta)\big(A^*[Q]q\big)(\theta)\, d\theta. \tag{111}$$

Recall, that $A^*[Q]$ denotes the L^2-adjoint of the operator $A[Q]$ and that it is given as

$$\big(A^*[Q]h\big)(\theta) = \nabla\cdot\Big(h(\theta)\big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta) + \lambda_2\nabla\log\big(h(\theta)\big)\big)\Big) \tag{112}$$

for $h \in C^2(\mathbb{R}^J, \mathbb{R})$ with compact support. This implies

$$\big(A^*[Q]q\big)(\theta) = \nabla\cdot\Big(q(\theta)\big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta)\big)\Big) + \lambda_2\Delta q(\theta). \tag{113}$$

We plug this into (111) to obtain

$$\int A[Q]h(\theta)\, dQ(\theta) \tag{114}$$

$$= \int h(\theta)\,\nabla\cdot\Big(q(\theta)\big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta)\big)\Big)\, d\theta + \int \lambda_2 h(\theta)\Delta q(\theta)\, d\theta. \tag{115}$$

On the other hand, we have that

$$\nabla_W L[Q](\theta) = \nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta) + \lambda_2\nabla\log q(\theta), \tag{116}$$

and therefore

$$\int \nabla L[Q](\theta)\cdot\nabla h(\theta)\, dQ(\theta) \tag{117}$$

$$= \int \Big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta) + \lambda_2\nabla\log q(\theta)\Big)\cdot\nabla h(\theta)\, dQ(\theta) \tag{118}$$

$$= \int q(\theta)\big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta)\big)\cdot\nabla h(\theta)\, d\theta + \lambda_2\int \nabla q(\theta)\cdot\nabla h(\theta)\, d\theta \tag{119}$$

$$= -\int \nabla\cdot\Big(q(\theta)\big(\nabla V(\theta) + \lambda_1(\nabla_1\kappa * Q)(\theta)\big)\Big)h(\theta)\, d\theta - \lambda_2\int q(\theta)\Delta h(\theta)\, d\theta, \tag{120}$$

where the last line follows from partial integration. Comparing (115) to (120) gives

$$\int A[Q]h(\theta)\,dQ(\theta) = -\int \nabla L[Q](\theta)\cdot\nabla h(\theta)\,dQ(\theta) \tag{121}$$

for all $h \in C_c^2(\mathbb{R}^J)$ whenever Q has a density. This immediately implies that Q is invariant iff it is a stationary point. □

Again, we leverage this correspondence between stationary point and invariant measures. There is a rich literature on ergodicity of nonlinear Markov processes. For example, Theorem 2 of Veretennikov (2006) specifies conditions on κ and V such that

$$Q^{n,N_E}(t) \xrightarrow{\mathcal{D}} Q_\infty \tag{122}$$

for $N_E, t \to \infty$. Here $Q^{n,N_E}(t)$ denotes the law of a fixed particle $\theta_n(t)$, $n = 1, \ldots, N_E$, whose distribution is characterised by the SDE (59). The measure Q_∞ is the unique invariant measure of the nonlinear Markov process. By Lemma 6 every invariant measure is a stationary point of the Wasserstein gradient and vice versa. Hence, existence and uniqueness of the stationary point of the Wasserstein gradient is immediately implied. However, since the global minimiser Q^* is a stationary point of the Wasserstein gradient (cf. Lemma 4), we conclude by uniqueness that $Q_\infty = Q^*$.

F Asymptotic analysis of deep repulsive ensembles

In this section, we consider the objective

$$L(Q) := \int \ell(\theta)\,dQ(\theta) + \lambda\,\mathrm{MMD}(Q,P) \tag{123}$$

for $Q \in \mathcal{P}(\mathbb{R}^J)$. The corresponding McKean-Vlasov process is of the form

$$d\theta(t) = -\Big(\nabla V(\theta(t)) + \lambda(\nabla_1\kappa * Q_t)(\theta(t))\Big)dt, \tag{124}$$

where Q_t denotes the distribution of $\theta(t)$ and $V(\theta) = \ell(\theta) - \mu_P(\theta)$ with $\mu_P(\theta) = \int \kappa(\theta,\theta')\,dP(\theta)$ the kernel mean-embedding of P. We call the particle method in this case **deep repulsive ensembles (DRE)**.

The existence of the global minimiser Q^* is still guaranteed under the assumptions in Appendix A. Lemma 4 guarantees that Q^* is a stationary point of the Wasserstein gradient, i.e.

$$\nabla V(\theta) + \lambda(\nabla_1\kappa * Q^*)(\theta) = 0 \tag{125}$$

for Q^*-a.e. $\theta \in \mathbb{R}^J$. Recall that the infinitesimal generator in this case is given as

$$\big(A[Q]h\big)(\theta) := -\Big(\nabla V(\theta) + \lambda(\nabla_1\kappa * Q)(\theta)\Big)\cdot\nabla h(\theta) \tag{126}$$

for $Q \in \mathcal{P}(\mathbb{R}^J)$, $h \in C_c^2(\mathbb{R}^J)$. It immediately follows from the definition that

$$\big(A[Q]h\big)(\theta) = -\nabla_W L[Q](\theta)\cdot\nabla h(\theta) \tag{127}$$

for all $h \in C_c^2(\mathbb{R}^J)$, $\theta \in \mathbb{R}^J$. As in Lemma 5 & 6, this implies that each stationary point of the Wasserstein gradient is an invariant measure of the McKean-Vlasov process and vice versa. In Appendix D & E, we cite relevant literature that guarantees uniqueness of the invariant measure, which is a necessary (but not sufficient) condition for convergence to the invariant measure. The next theorem shows that uniqueness will in general not hold without the presence of the diffusion term.

Theorem 6. *The invariant measure for the McKean-Vlasov process with the family of generators* $\big(A[Q]\big)_{Q\in\mathcal{P}(\mathbb{R}^J)}$ *defined in (127) is (in general) not unique.*

Proof. Let $N_E \in \mathbb{N}$ and define $\widetilde{L} : \big(\mathbb{R}^J\big)^{N_E} \to \mathbb{R}$ as

$$\widetilde{L}(\theta_1,\ldots,\theta_{N_E}) := \sum_{i=1}^{N_E} V(\theta_i) + \frac{\lambda}{2N_E}\sum_{i,j=1}^{N_E}\kappa(\theta_i,\theta_j). \tag{128}$$

Assume that V is bounded from below and norm-coercive. Then $\widetilde{L}$ is bounded from below and norm-coercive and therefore we can find a global minimiser $\theta^* := (\theta_1^*, \ldots, \theta_{N_E}^*) \in (\mathbb{R}^J)^{N_E}$ of $\widetilde{L}$. Since $\widetilde{L}$ is differentiable, we know that θ^* is a stationary point of the gradient which implies

$$\nabla V(\theta_i^*) + \frac{\lambda}{N_E} \sum_{j=1}^{N_E} (\nabla_1 \kappa)(\theta_i^*, \theta_j^*) = 0 \tag{129}$$

for all $i = 1, \ldots, N_E$. Here, we assume that the kernel κ is symmetric, which is standard in the MMD literature. Note that (129) is equivalent to

$$\nabla V(\theta) + \lambda(\nabla_1 \kappa * \widehat{Q})(\theta) = 0 \tag{130}$$

for $\widehat{Q}$-a.e. $\theta \in \mathbb{R}^J$ where

$$\widehat{Q}(d\theta) := \frac{1}{N_E} \sum_{j=1}^{N_E} \delta_{\theta_j^*}(d\theta). \tag{131}$$

This means that $\widehat{Q})$ is a stationary point of the Wasserstein gradient, and therefore an invariant measure for the McKean-Vlasov process. Since $N_E \in \mathbb{N}$ was arbitrary, we have constructed countably many invariant measures and therefore uniqueness can't hold in general. □

The reason that non-uniqueness of the invariant measure is an immediate contradiction to convergence is the following: If we initialise with any of the invariant measures constructed in the proof of Theorem 6, then the particle distribution of the McKean-Vlasov process will remain unchanged over time. Convergence to the global minimiser can therefore surely not hold for arbitrary initialisation Q_0. It may be possible to construct conditions on Q_0 under which convergence still holds. For example, for Stein variational gradient descent a similar issue occurs. However, in this case one can guarantee convergence (Lu et al., 2019, Theorem 2.8) if Q_0 has a Lebesgue-density (and if the kernel satisfies further restrictive assumptions). The existence of conditions that guarantee convergence for DRE remains an open problem.

G Implementation details

In Appendix A, we derived the following algorithm:

Step 1: Simulate $N_E \in \mathbb{N}$ particles $\theta_{1,0}, \ldots, \theta_{N_E,0}$ from a use chosen initial distribution Q_0.

Step 2: Evolve the particles forward in time according to

$$\theta_{n,k+1} = \theta_{n,k} - \eta\Big(\nabla V\big(\theta_{n,k}\big) + \frac{\lambda_1}{N_E} \sum_{j=1}^{N_E} (\nabla_1 \kappa)\big(\theta_{n,k}, \theta_{j,k}\big)\Big) + \sqrt{2\eta\lambda_2} Z_{n,k} \tag{132}$$

for $n = 1, \ldots, N_E$, $k = 0, \ldots, K-1$ with $Z_{n,k} \sim \mathcal{N}(0, I_{J \times J})$.

We can generate samples from DE, DLE and DRLE by setting the potential and regularisation parameters as described below:

- Deep ensembles: $V(\theta) = \ell(\theta)$, $\lambda_1 = 0$, $\lambda_2 = 0$
- Deep Langevin ensembles: $V(\theta) = \ell(\theta) - \lambda \log p(\theta)$, $\lambda_1 = 0$, $\lambda = \lambda_2$
- Deep repulsive Langevin ensembles: $V(\theta) = \ell(\theta) - \lambda_1 \log p(\theta) - \lambda_2 \mu_P(\theta)$

Due to Appendix D & E, we can think of $\theta_{1,K}, \ldots, \theta_{N_E,K}$ as approximately sampled from the global minimiser Q^* for DLE and DRLE if K is large enough. All experiments use the SE kernel given as

$$\kappa(\theta, \theta') = \exp\big(- \frac{\|\theta - \theta'\|^2}{2\sigma_\kappa^2}\big) \tag{133}$$

with lengthscale parameter $\sigma_\kappa > 0$. The kernel mean embedding μ_P can easily be approximated as

$$\mu_P(\theta) = \frac{1}{M} \sum_{i=1}^{M} \kappa(\theta, \theta_i),\ \theta \in \mathbb{R}^J, \tag{134}$$

where $\theta_1, \ldots, \theta_M \sim P$ independently. We chose $M = 20$.

G.1 Toy example: global minimiser

We describe details regarding the experiments conducted to produce Figure 2 below.

We generate $N_E = 300$ particles and make the following choices:

- Loss: $\ell(\theta) := \frac{3}{2}(\frac{1}{4}\theta^4 + \frac{1}{3}\theta^3 - \theta^2) - \frac{3}{8}$
- Prior: $P \sim \mathcal{N}(0, 1)$ and therefore $\log p(\theta) = -\frac{1}{2}\theta^2$
- Initialisation: $Q_0 = P$
- Reg. parameter: $\lambda_{DLE} = 1$, $\lambda_{DRLE} = 1$, $\lambda'_{DRLE} = 1$
- Step size: $\eta = 10^{-4}$, Iterations: $K = 100,000$
- Kernel lengthscale, σ_κ, is chosen according to the median heuristic (Garreau et al., 2017) based on samples from the prior P

The loss is constructed such that we have a global minimum at $\theta = -2$, a turning point at $\theta = 0$, and a local minimum at $\theta = 1$.

- Deep ensembles: The optimal Q^* is a Dirac measure located at the global minimiser $\theta = -2$. However, as we proved in Theorem 1, the WGF produce samples from

$$Q_\infty(d\theta) = \frac{1}{2}\delta_{-2}(d\theta) + \frac{1}{2}\delta_1(d\theta), \tag{135}$$

  as $(-\infty, 0)$ is the region of attraction for the global minimum and $(0, \infty)$ for the local minimum which both have probability 0.5 under $Q_0 = P = \mathcal{N}(0, 1)$. In particular, $Q_\infty \neq Q^*$ as expected.
- Deep Langevin ensembles: The optimal measure has the pdf

$$q^*(\theta) \propto \exp\big(-\frac{\ell(\theta)}{\lambda}\big)p(\theta) \tag{136}$$

  for $\theta \in \mathbb{R}$. As expected the WGF produces samples from Q^*.
- Deep repulsive Langevin ensembles: The optimal q^* for deep repulsive ensembles is harder to determine. From the condition that q^* is a stationary point of the Wasserstein gradient, we can derive that $u(\theta) := \log q^*(\theta)$ satisfies the integro-differential equation

$$u'(\theta) = -\frac{1}{\lambda_2}V'(\theta) - \frac{\lambda_1}{\lambda_2}\int (\nabla_1\kappa)(\theta, \theta')\exp(u(\theta'))\,d\theta' \tag{137}$$

  with some initial value $u(0) = u_0$. In principle, we could choose u_0 such that $q(\theta) := \exp(u(\theta))$ integrates to 1. However, since we do not know the appropriate initial condition a priori, we choose an arbitrary u_0 and normalise the pdf afterwards. We use an numerical solver to evaluate $u(\theta)$ on a fixed grid. As expected, the WGF produces samples from Q^* in this case.

G.2 Toy example: multimodal loss

The details below correspond to the experimental results presented in Figure 3.

DE, DLE, DRLE We generate $N_E = 300$ particles and make the following choices:

- Loss: $\ell(\theta) = -\log\sum_{i=1}^4 \frac{1}{4}\mathcal{N}(\theta; \mu_i, I_2)$, $\theta \in \mathbb{R}^2$, $\mu_i = (\pm 3, \pm 3)^T$, $i = 1, \ldots, 4$
- Prior: P flat and therefore $\log p(\theta) = 0$
- Initialisation: $Q_0 \sim \mathcal{N}(0, I_2)$
- Reg. parameter: $\lambda_{DLE} = 0.2$, $\lambda_{DRLE} = 0.2$, $\lambda'_{DRLE} = 0.6$
- Step size: $\eta = 0.1$, Iterations: $K = 10,000$
- Kernel lengthscale, σ_κ, is chosen according to the median heuristic (Garreau et al., 2017) based on samples from the prior P

Note that for a translation-invariant kernel such as the SE kernel we obtain for the flat prior P that

$$\mu_P(\theta) = \int_{-\infty}^{\infty} \kappa(\theta, \theta')\, d\theta' \tag{138}$$

$$= \int_{-\infty}^{\infty} \phi(\theta - \theta')\, d\theta' \tag{139}$$

$$= \int_{-\infty}^{\infty} \phi(\xi)\, d\xi, \tag{140}$$

where the second line follows from the fact that we can write any translation-invariant kernel as $\kappa(\theta, \theta') = \phi(\theta - \theta')$ for some function $\phi : \mathbb{R}^J \to \mathbb{R}$ and the second line is simple variable substitution. If (140) is finite, the above expression is well-defined and therefore μ_P constant. Note that in particular for the SE kernel, we have $\phi(\xi) = \exp(-\|\xi\|^2/(2\sigma_\kappa^2))$ and therefore (140) is finite. As a consequence, we have that for a flat prior P the gradient of the potential V is the same for all three methods. This means that the loss ℓ isn't adjusted and the only difference between the three methods is the presence of repulsion and noise effects.

Remark 3. The astute reader may have noticed that a flat prior P is in fact not covered by our theory in Appendix A. The problem is that $\mathrm{KL}(\cdot, \mathcal{L})$, where $\mathcal{L}$ denotes the Lebesgue measure, is not positive (and not even bounded from below). To see this, choose $Q = \mathcal{N}(0, \Sigma)$ with $\Sigma = \mathrm{diag}(\sigma_1^2,\, \sigma_2^2)$ and note that

$$\mathrm{KL}(Q, \mathcal{L}) = \int \log q(\theta)\, q(\theta)\, d\theta = -\mathrm{H}(Q), \tag{141}$$

where $\mathrm{H}(Q)$ denotes the differential entropy. For a Gaussian, it is known that

$$\mathrm{H}(Q) = \frac{1}{2} \log\big((2\pi e)^J \det(\Sigma)\big) = \frac{1}{2}\big(\log(2\pi e)^J + \log(\sigma_1^2) + \log(\sigma_2^2)\big) \tag{142}$$

and therefore if either $\sigma_1^2 \to \infty$ or $\sigma_2^2 \to \infty$ then $\mathrm{KL}(Q, \mathcal{L}) \to -\infty$. However, note that this difficulty is rather technical in nature and can easily be remedied. Instead of $\mathcal{L}$, we could have chosen the uniform prior $P \sim U(-10^{100}, 10^{100})$. In this case, the positivity of $\mathrm{KL}(\cdot, P)$ is guaranteed by Jensen's inequality. This choice of P gives—up to an additive constant—the same objective as a flat prior and up to machine precision the same kernel mean embedding μ_P. It is, therefore, algorithmically irrelevant if P is flat or uniform on a very large set.

FD-GVI We use the same prior and loss as for DE, DLE and DRLE. We parameterise the variational family as independent Gaussian, i.e.

$$\mathcal{Q} = \big\{\mathcal{N}(\mu, \Sigma) \,|\, \mu \in \mathbb{R}^2,\, \Sigma = \mathrm{diag}\big(\exp(\beta_1), \exp(\beta_2)\big), \beta := (\beta_1, \beta_2)^2 \in \mathbb{R}^2\big\}. \tag{143}$$

We learn the variational parameters $\nu := (\mu, \beta) \in \mathbb{R}^4$ by minimising

$$\widetilde{L}(\nu) = \int \ell(\theta)\, dQ_\nu(\theta) + \lambda KL(Q_\nu, P) \tag{144}$$

$$= \int \ell(\theta)\, dQ_\nu(\theta) - \lambda H\big(\mathcal{N}(\mu, \Sigma)\big) \tag{145}$$

$$\approx \frac{1}{200} \sum_{j=1}^{200} \ell(\mu + \Sigma^{0.5} Z_j) - \frac{\lambda}{2} \log\big((2\pi e)^2 \exp(\beta_1) \exp(\beta_2)\big) \tag{146}$$

$$= \frac{1}{200} \sum_{j=1}^{200} \ell(\mu + \Sigma^{0.5} Z_j) - \frac{\lambda}{2}\big(\beta_1 + \beta_2\big) + const, \tag{147}$$

where $Z_1, \ldots, Z_{200} \sim \mathcal{N}(0, I_2)$ and $H\big(\mathcal{N}(\mu, \Sigma)\big)$ denotes the differential entropy of the normal distribution. For the regularisation parameter, we chose $\lambda = 0.5$.

G.3 Toy example: more modes than particles

We generate $N_E = 20$ particles and make the following choices:

- Loss: $\ell(\theta) = -|\sin(\theta)|$, $\theta \in [-M\pi, M\pi]$, with $M = 1000$
- Prior: P flat and therefore $\log p(\theta) = 0$ and $\mu_P = \text{const.}$ (cf. Appendix G.2)
- Initialisation: $Q_0 \sim U(-M\pi, M\pi)$
- Reg. parameter: $\lambda_{DLE} = 0.001$, $\lambda_{DRLE} = 0.001$, $\lambda'_{DRLE} = 0.6$
- Step size: $\eta = 0.01$, Iterations: $K = 1.000$
- Kernel lengthscale, σ_κ, is chosen according to the median heuristic (Garreau et al., 2017) based on samples from the prior P

Note that ℓ has $2M = 2000$ local minima at locations

$$m_i := \frac{\pi}{2} + i\pi, \quad i \in \{-M, \ldots, 0, \ldots, (M-1)\}. \tag{148}$$

Due to the flat prior $\nabla V = \nabla \ell$ for all three methods. We observe that it is hard to distinguish the methods since most particles are in their local modes by themselves.

G.4 UCI Regression

The UCI data sets are licensed under Creative Commons Attribution 4.0 International license (CC BY 4.0). Following Lakshminarayanan et al. (2017), we train 5 one-hidden-layer neural networks f_θ with 50 hidden nodes for 40 epochs. We split each data set into train (81% of samples), validation (9% of samples), and test set (10% of samples). Based on the best hyperparameter runs (according to a Gaussian NLL) found via grid search on a validation data set, we make the following choices:

- Loss: $\ell(\theta) = \frac{1}{N}\sum_{n=1}^{N}(f_\theta(x_n) - y_n)^2$ where $\{x_n, y_n\}_{n=1}^{N}$ are paired observations.
- Prior: $P \sim \mathcal{N}(0, 1)$
- Initialisation: Kaiming intilisation, i.e. for each layer $l \in \{1, ...L\}$ that maps features with dimensionality n_{l-1} into dimensionality n_l, we sample $Q_{l,0} \sim \mathcal{N}(0, 2/n_l)$
- Reg. parameter: $\lambda_{DLE} = 10^{-4}$, $\lambda_{DRLE} = 10^{-4}$, $\lambda'_{DRLE} = 10^{-2}$
- Step size: $\eta = 0.1$, Iterations: $K = 10,000$
- Kernel lengthscale, σ_κ, is chosen according to the median heuristic (Garreau et al., 2017) based on samples from the prior P

G.5 Compute

While the final experimental results can be run within approximately an hour on a single GeForce RTX 3090 GPU, the complete compute needed for the final results, debugging runs, and sweeps amounts to around 9 days.

References

Ahmed, N. and Ding, X. (1993). On invariant measures of nonlinear Markov processes. *Journal of Applied Mathematics and Stochastic Analysis*, 6(4):385–406.

Alquier, P. (2021a). Non-exponentially weighted aggregation: regret bounds for unbounded loss functions. In *International Conference on Machine Learning*, pages 207–218. PMLR.

Alquier, P. (2021b). User-friendly introduction to PAC-Bayes bounds. *arXiv preprint arXiv:2110.11216*.

Alquier, P. and Guedj, B. (2018). Simpler PAC-Bayesian bounds for hostile data. *Machine Learning*, 107(5):887–902.

Altamirano, M., Briol, F.-X., and Knoblauch, J. (2023). Robust and scalable Bayesian online changepoint detection. *arXiv preprint arXiv:2302.04759*.

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

Arbel, M., Korba, A., Salim, A., and Gretton, A. (2019). Maximum mean discrepancy gradient flow. *Advances in Neural Information Processing Systems*, 32.

Barbu, V. and Röckner, M. (2020). From nonlinear Fokker–Planck equations to solutions of distribution dependent sde. *arXiv preprint arXiv:1808.10706*.

Bégin, L., Germain, P., Laviolette, F., and Roy, J.-F. (2016). PAC-Bayesian bounds based on the Rényi divergence. In *Artificial Intelligence and Statistics*, pages 435–444.

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). A general framework for updating belief distributions. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):1103–1130.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *International conference on machine learning*, pages 1613–1622. PMLR.

Bressan, A. (2003). Tutorial on the center manifold theorem. *Hyperbolic systems of balance laws*, 1911:327–344.

Chiang, T.-S., Hwang, C.-R., and Sheu, S. J. (1987). Diffusion for global optimization in R^n. *SIAM Journal on Control and Optimization*, 25(3):737–753.

Colding, T. H. and Minicozzi II, W. P. (2014). Lojasiewicz inequalities and applications. *arXiv preprint arXiv:1402.5087*.

D'Angelo, F. and Fortuin, V. (2021). Repulsive deep ensembles are Bayesian. *Advances in Neural Information Processing Systems*, 34:3451–3465.

Ermak, D. L. (1975). A computer simulation of charged particles in solution. i. technique and equilibrium properties. *The Journal of Chemical Physics*, 62(10):4189–4196.

Fort, S., Hu, H., and Lakshminarayanan, B. (2019). Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*.

Gal, Y. and Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pages 1050–1059. PMLR.

Garreau, D., Jitkrittum, W., and Kanagawa, M. (2017). Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*.

Glaser, P., Arbel, M., and Gretton, A. (2021). Kale flow: A relaxed kl gradient flow for probabilities with disjoint support. *Advances in Neural Information Processing Systems*, 34:8018–8031.

Graves, A. (2011). Practical variational inference for neural networks. *Advances in neural information processing systems*, 24.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773.

Grünwald, P. (2011). Safe learning: bridging the gap between Bayes, MDL and statistical learning theory via empirical convexity. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 397–420.

Guedj, B. and Shawe-Taylor, J. (2019). A primer on pac-Bayesian learning. In *ICML 2019-Thirty-sixth International Conference on Machine Learning*.

Haddouche, M. and Guedj, B. (2023). Wasserstein PAC-Bayes learning: A bridge between generalisation and optimisation. *arXiv preprint arXiv:2304.07048*.

Husain, H. and Knoblauch, J. (2022). Adversarial interpretation of Bayesian inference. In *International Conference on Algorithmic Learning Theory*, pages 553–572. PMLR.

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. G. (2021). What are Bayesian neural network posteriors really like? In *International conference on machine learning*, pages 4629–4640. PMLR.

Jewson, J., Smith, J., and Holmes, C. (2018). Principles of Bayesian inference using general divergence criteria. *Entropy*, 20(6):442.

Knoblauch, J. (2019). Frequentist consistency of generalized variational inference. *arXiv preprint arXiv:1912.04946*.

Knoblauch, J. (2021). *Optimization-centric generalizations of Bayesian inference*. PhD thesis, University of Warwick.

Knoblauch, J., Jewson, J., and Damoulas, T. (2018). Doubly robust Bayesian inference for non-stationary streaming data using β-divergences. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 64–75.

Knoblauch, J., Jewson, J., and Damoulas, T. (2022). An optimization-centric view on Bayes' rule: Reviewing and generalizing variational inference. *Journal of Machine Learning Research*, 23(132):1–109.

Kolokoltsov, V. N. (2010). *Nonlinear Markov processes and kinetic equations*, volume 182. Cambridge University Press.

Korba, A., Aubin-Frankowski, P.-C., Majewski, S., and Ablin, P. (2021). Kernel Stein discrepancy descent. In *International Conference on Machine Learning*, pages 5719–5730. PMLR.

Lakshminarayanan, B., Pritzel, A., and Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. (2016). Gradient descent only converges to minimizers. In *Conference on learning theory*, pages 1246–1257. PMLR.

Lichman, M. (2013). UCI machine learning repository.

Liggett, T. M. (2010). *Continuous time Markov processes: an introduction*, volume 113. American Mathematical Soc.

Liu, Q. (2017). Stein variational gradient descent as gradient flow. *Advances in neural information processing systems*, 30.

Liu, Q. and Wang, D. (2016). Stein variational gradient descent: A general purpose Bayesian inference algorithm. *Advances in neural information processing systems*, 29.

Louizos, C. and Welling, M. (2017). Multiplicative normalizing flows for variational Bayesian neural networks. In *International Conference on Machine Learning*, pages 2218–2227. PMLR.

Lu, J., Lu, Y., and Nolen, J. (2019). Scaling limit of the stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671.

Matsubara, T., Knoblauch, J., Briol, F.-X., and Oates, C. J. (2022). Robust generalised Bayesian inference for intractable likelihoods. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(3):997–1022.

McAllester, D. A. (1999a). PAC-Bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170. ACM.

McAllester, D. A. (1999b). Some PAC-Bayesian theorems. *Machine Learning*, 37(3):355–363.

Miller, J. W. and Dunson, D. B. (2019). Robust Bayesian inference via coarsening. *Journal of the American Statistical Association*, 114(527):1113–1125.

Muandet, K., Fukumizu, K., Sriperumbudur, B., Schölkopf, B., et al. (2017). Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends® in Machine Learning*, 10(1-2):1–141.

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Ovadia, Y., Fertig, E., Ren, J., Nado, Z., Sculley, D., Nowozin, S., Dillon, J., Lakshminarayanan, B., and Snoek, J. (2019). Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32.

Polyanskiy, Y. and Wu, Y. (2014). Lecture notes on information theory. *Lecture Notes for ECE563 (UIUC) and*, 6(2012-2016):7.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363.

Santambrogio, F. (2015). Optimal transport for applied mathematicians. *Birkäuser, NY*, 55.

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

Shawe-Taylor, J. and Williamson, R. C. (1997). A PAC analysis of a Bayesian estimator. In *Annual Workshop on Computational Learning Theory: Proceedings of the tenth annual conference on Computational learning theory*, volume 6, pages 2–9.

Veretennikov, A. Y. (2006). On ergodic measures for McKean-Vlasov stochastic equations. In *Monte Carlo and Quasi-Monte Carlo Methods 2004*, pages 471–486. Springer Berlin Heidelberg.

Villani, C. (2003). *Topics in optimal transportation*, volume 58. American Mathematical Soc.

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

Welling, M. and Teh, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th international conference on machine learning (ICML-11)*, pages 681–688.

Wild, V. D., Hu, R., and Sejdinovic, D. (2022). Generalized variational inference in function spaces: Gaussian measures meet Bayesian deep learning. *Advances in Neural Information Processing Systems*, 35:3716–3730.

Wilson, A. G. (2020). The case for Bayesian deep learning. *arXiv preprint arXiv:2001.10995*.

Wilson, A. G. and Izmailov, P. (2020). Bayesian deep learning and a probabilistic perspective of generalization. *Advances in neural information processing systems*, 33:4697–4708.

Wu, P.-S. and Martin, R. (2023). A comparison of learning rate selection methods in generalized Bayesian inference. *Bayesian Analysis*, 18(1):105–132.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, 67(2):301–320.