A Rigorous Link between Deep Ensembles and (Variational) Bayesian Methods

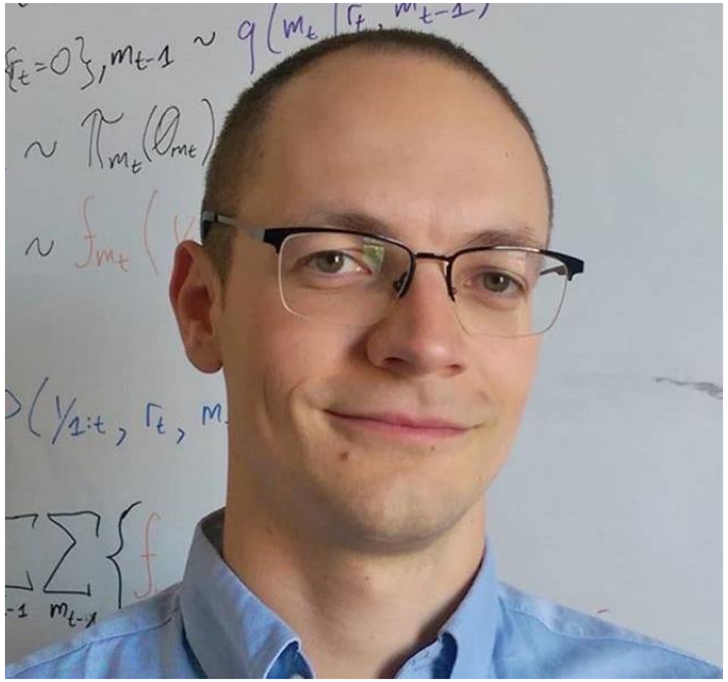
Veit D. Wild

UNIVERSITY OF OXFORD

Department of Statistics

Sahra Ghalebikesabi

Dino Sejdinovic



Jeremias Knoblauch

Generalised Variational Inference

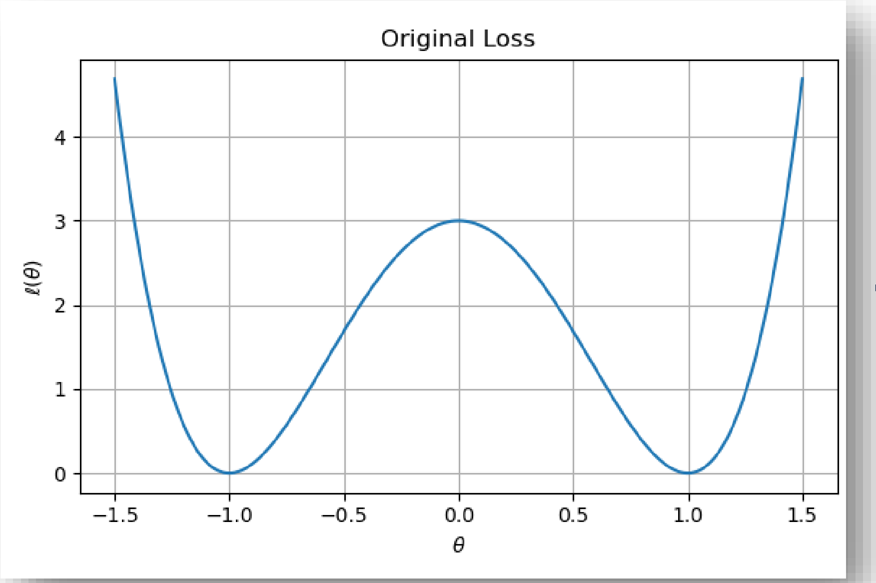
P prior/reference prob. measure
$\lambda > 0$ reg. parameter
$D(\cdot, P)$ convex regulariser

$$\theta^* \in \arg\min_{\theta \in \mathbb{R}^J} \ell(\theta)$$

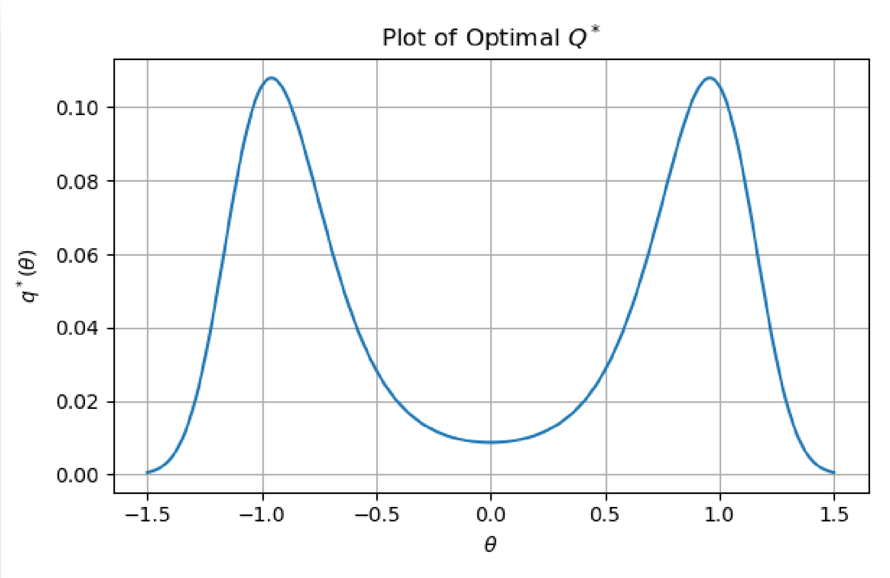
prob. lifting + convexify

$$L(Q) = \int \ell(\theta)\, dQ(\theta) \; + \lambda D(Q, P)$$

$$Q^* = \arg\min_{Q \in \mathcal{P}(\mathbb{R}^J)} L(Q)$$



choose λ, D, P



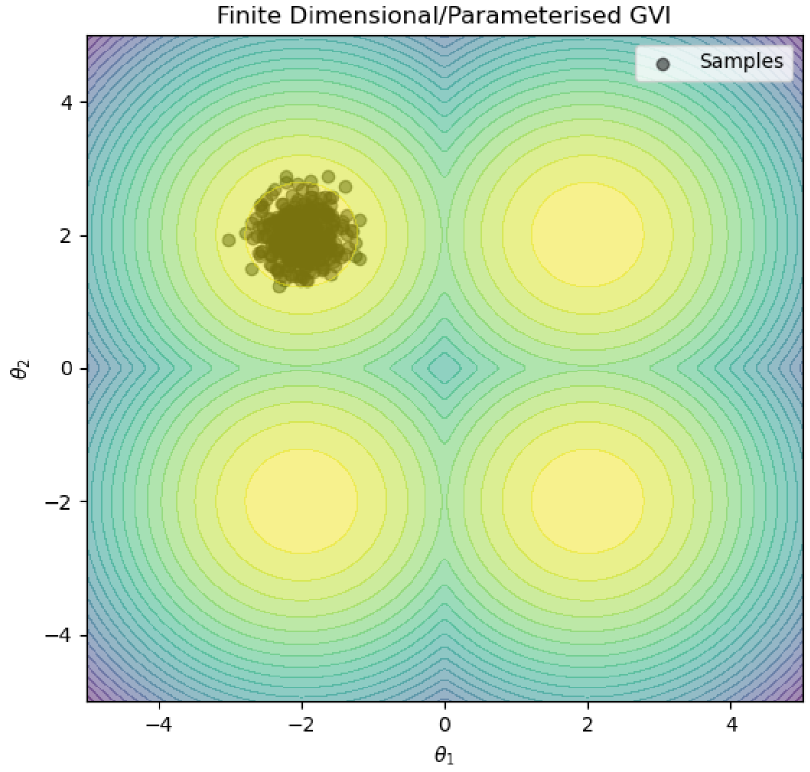
How to find the global minimiser?

Parameterised/Finite Dim GVI:

$$\mathcal{Q} := \{Q_\nu \, : \nu \in \Gamma\} \subset \mathcal{P}(\mathbb{R}^J)$$

e.g. : $Q_\nu = \mathcal{N}(\mu, \Sigma), \nu = (\mu, \Sigma)$

$$\widetilde{L}(\nu) := L(Q_\nu) \longrightarrow \arg\min_{\nu\in\Gamma} \widetilde{L}(\nu)$$



Alternative:

→ Gradient descent in infinite dimensions

Gradient Descent in Infinite Dimensions

Finite Dimensions	Infinite Dimensions
$\theta_0 \in \mathbb{R}^J$	$Q_0 \in \mathcal{P}_2(\mathbb{R}^J)$
$\theta_{k+1} = \underset{\theta \in \mathbb{R}^J}{\arg\min} \left\{ \ell(\theta) + \frac{1}{2\eta} \lVert \theta - \theta_k \rVert_2^2 \right\}$	$Q_{k+1} := \underset{Q \in \mathcal{P}_2(\mathbb{R}^J)}{\arg\min} \left\{ L(Q) + \frac{1}{2\eta} \underbrace{W_2(Q, Q_k)^2}_{\text{Wasserstein metric}} \right\}$
$\downarrow \ \eta \to 0$	$\eta \to 0 \ \downarrow$
$\theta'(t) = -\nabla \ell(\theta(t))$	$\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta) \underbrace{\nabla_W L[Q(t)](\theta)}_{\text{Wasserstein gradient}} \Big)$

Hope: $q(t,\cdot) \to q^*$ for $t \to \infty$

Follow the WGF

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

$$\nabla_W L^{\text{fe}}[Q](\theta) = \nabla V(\theta) + \lambda_1 \int (\nabla_1 \kappa)(\theta, \theta')\, dQ(\theta') + \lambda_2 \nabla \log q(\theta),$$

Step 1: Sample $N_E \in \mathbb{N}$ particles $\theta_1(0), \ldots, \theta_{N_E}(0)$ independently from $Q_0 \in \mathcal{P}_2(\mathbb{R}^J)$.

Step 2: Evolve the particle θ_n by following the stochastic differential equation (SDE)

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

for $n = 1, \ldots, N_E$, and $\{B_n(t)\}_{t>0}$ stochastically independent Brownian motions.

Theoretical Analysis

Step 1: Initialise $N_E \in \mathbb{N}$ 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

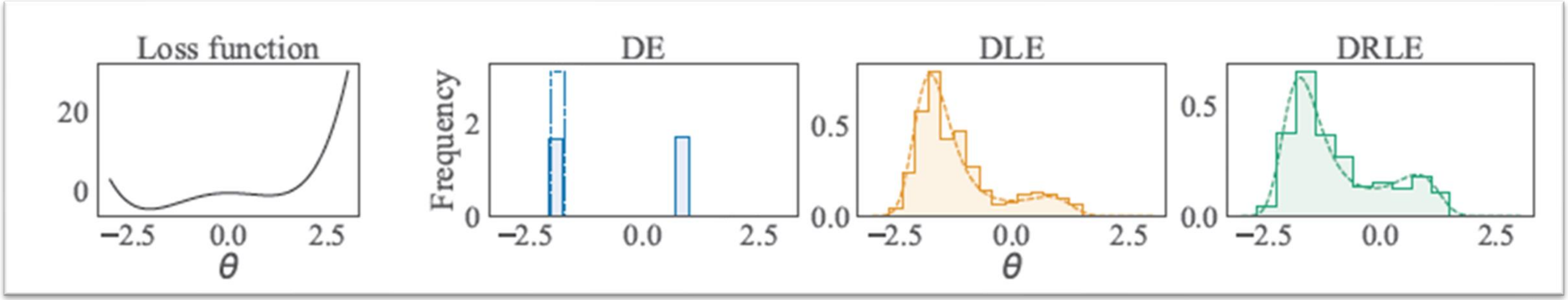
$$\theta_{n,k+1} = \theta_{n,k} - \eta\Big(\nabla V(\theta_{n,k}) + \frac{\lambda_1}{N_E}\sum_{j=1}^{N_E}(\nabla_1\kappa)(\theta_{n,k}, \theta_{j,k})\Big) + \sqrt{2\eta\lambda_2}Z_{n,k}$$

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

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

$D(Q,P)$	$V(\theta)$	λ_1	λ_2	Method	Convergence
0	$\ell(\theta)$	0	0	DE	🚫
$\mathrm{KL}(Q,P)$	$\ell(\theta) - \lambda \log p(\theta)$	0	λ	DLE	✓
$\lambda\,\mathrm{MMD}^2(Q,P) + \lambda'\,\mathrm{KL}(Q,P)$	$\ell(\theta) - 2\lambda\underbrace{\mu_P(\theta)}_{\text{kernel mean embedding of P}} - \lambda' \log p(\theta)$	2λ	λ'	DRLE	✓

Visualisation



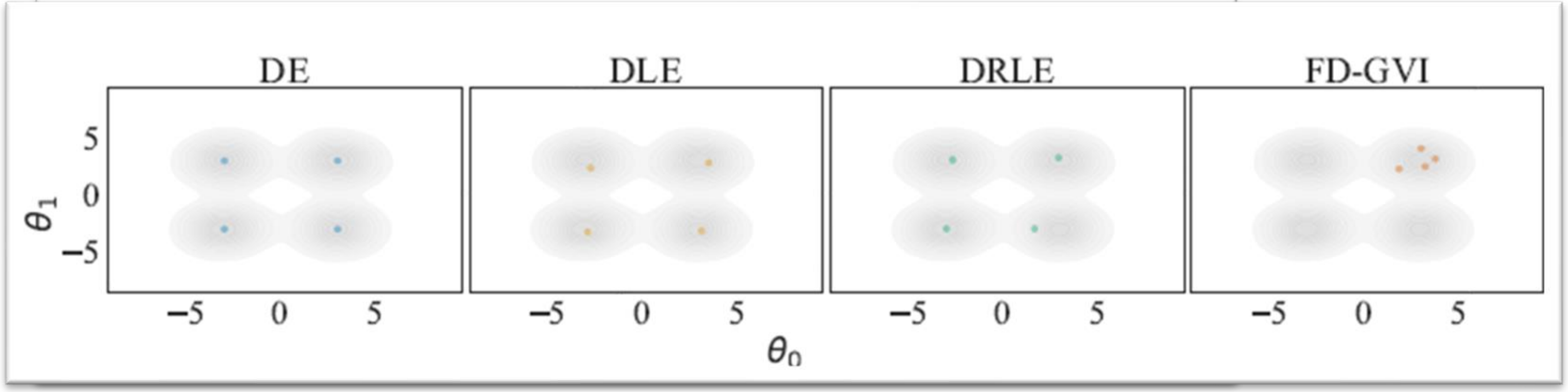
Theory meets Practice

	KIN8NM	CONCRETE	ENERGY	NAVAL	POWER	PROTEIN	WINE	YACHT
DE	$\mathbf{0.33_{\pm 0.1}}$	$6.10_{\pm 0.3}$	$2.83_{\pm 0.2}$	$-0.40_{\pm 0.3}$	$\mathbf{13.70_{\pm 2.6}}$	$\mathbf{11.22_{\pm 2.2}}$	$14.65_{\pm 1.9}$	$2.20_{\pm 0.4}$
DLE	$13.25_{\pm 4.3}$	$\mathbf{5.11_{\pm 0.2}}$	$\mathbf{2.43_{\pm 0.1}}$	$3.46_{\pm 2.4}$	$13.87_{\pm 2.3}$	$43.20_{\pm 12.5}$	$13.73_{\pm 1.4}$	$\mathbf{1.64_{\pm 0.1}}$
DRLE	$0.46_{\pm 0.1}$	$8.30_{\pm 0.6}$	$4.01_{\pm 0.3}$	$\mathbf{-3.04_{\pm 0.2}}$	$23.21_{\pm 2.0}$	$48.80_{\pm 2.1}$	$\mathbf{7.13_{\pm 0.6}}$	$7.80_{\pm 2.7}$

No difference, but why?

- Finite time?
- Discretisation error?
- Batch estimation?

nr. sample size << nr. modes



Thank you for your attention!

Deep Ensembles

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

$V(\theta)$	λ_1	λ_2
$\ell(\theta)$	0	0

$$L(Q) = \int \ell(\theta)\, dQ(\theta) + \lambda D(Q,P) \xrightarrow{\text{choose}} D(Q,P) = 0$$

Theorem 1. *If ℓ has countably many local minima $\{m_i : i \in \mathbb{N}\}$, then it holds independently for each $n = 1, \ldots, N_E$ that*

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

for $t \to \infty$. Here $\xrightarrow{\mathcal{D}}$ denotes convergence in distribution and $\Theta_i = \{\theta \in \mathbb{R}^J : \lim_{t\to\infty} \theta_*(t) = m_i$ and $\theta_*(0) = \theta\}$ denotes the domain of attraction for m_i with respect to the gradient flow θ_*.

Deep Langevin Ensembles

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

$V(\theta)$	λ_1	λ_2
$\ell(\theta) - \lambda \log p(\theta)$	0	λ

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

$$\xrightarrow{\text{choose}} \quad D(Q, P) = \mathrm{KL}(Q, P)$$

Theorem 2. *For each $n = 1, \ldots, N_E$ independently*

$$\theta_n(t) \xrightarrow{\mathcal{D}} Q^*$$

with $q^*(\theta) \propto \exp\big(-\frac{1}{\lambda}\ell(\theta)\big)p(\theta)$.

Deep Rep. Langevin Ensembles

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

$V(\theta)$	λ_1	λ_2
$\ell(\theta) - 2\lambda\mu_P(\theta) - \lambda' \log p(\theta)$	2λ	λ'

$$L(Q) = \int \ell(\theta)\, dQ(\theta) + D(Q,P)$$

$$\xrightarrow{\text{choose}} \quad D(Q,P) = \lambda\,\mathrm{MMD}^2(Q,P) + \lambda'\,\mathrm{KL}(Q,P)$$

Theorem 3. *Let $Q^{n,N_E}(t)$ be the distribution of $\theta_n(t)$, $n = 1,\ldots,N_E$, generated via the WGF. Then*

$$Q^{n,N_E}(t) \xrightarrow{\mathcal{D}} Q^*$$

for each $n = 1,\ldots,N_E$ and as $N_E \to \infty$, $t \to \infty$.