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

Jeremias Knoblauch; Lecturer/Assistant Prof & EPSRC Fellow @ UCL Stats

22/07/23

Collaborators

Veit Wild (Oxford)

Sahra Ghalebikesabi
(Oxford)

Dino Sejdinovic
(Adelaide)

Work available as preprint:

https://arxiv.org/abs/2305.15027

Today's key take-away

$$\min_{\theta\in\Theta}\ell(\theta) \qquad \min_{Q\in\mathcal{P}(\mathbb{R}^J)}\int \ell(\theta)dQ(\theta) \qquad \min_{Q\in\mathcal{P}(\mathbb{R}^J)}\left\{\int \ell(\theta)dQ(\theta)+\lambda D(Q,P)\right\}$$

Step 1: probabilistic lifting **Step 2:** convexification through regularisation

1. **Proposal**: Non-convex, finite-dimensional (FD) => convex, infinite-dimensional (ID)

Today's key take-away

$$\min_{\theta\in\Theta} \ell(\theta) \xrightarrow{\textbf{Step 1: } \text{probabilistic lifting}} \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \int \ell(\theta)\,dQ(\theta) \xrightarrow{\textbf{Step 2: } \text{convexification through regularisation}} \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta)\,dQ(\theta) + \lambda D(Q,P) \right\}$$

1. **Proposal**: Non-convex, finite-dimensional (FD) => convex, infinite-dimensional (ID)
2. **Problem**: generally unsolvable

Today's key take-away

$$\min_{\theta\in\Theta} \ell(\theta) \qquad \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \int \ell(\theta)dQ(\theta) \qquad \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta)dQ(\theta) + \lambda D(Q,P) \right\}$$

Step 1: probabilistic lifting **Step 2:** convexification through regularisation

1. **Proposal**: Non-convex, finite-dimensional (FD) => convex, infinite-dimensional (ID)
2. **Problem**: generally unsolvable
3. **Solution**: build ID gradient descent (GD) algorithm?
   (tells us about interplay of Bayes & Deep ensembles)

Outline

1. **Motivation**: convexity > finite-dimensionality
2. **Connections**: other ID problems over measures & algorithms
3. **Proposal**: gradient descent schemes in infinite dimensions
4. **Lessons & Experiments**

Motivation: loss-minimisation & convexity

Classical problem: find θ^*

$$\boxed{\theta^* = \text{argmin}_{\theta \in \Theta}\, \ell(\theta) \text{ for some loss } \ell : \Theta \to \mathbb{R}}$$

Motivation: loss-minimisation & convexity

Classical problem: find θ^*

Easy to do via GD if ℓ (strictly) convex

$$\boxed{\theta^* = \mathrm{argmin}_{\theta \in \Theta}\, \ell(\theta) \text{ for some loss } \ell : \Theta \to \mathbb{R}}$$

Motivation: loss-minimisation & convexity

Classical problem: find θ^*

Easy to do via GD if ℓ (strictly) convex

$$\boxed{\theta^* = \mathrm{argmin}_{\theta \in \Theta}\, \ell(\theta) \text{ for some loss } \ell : \Theta \to \mathbb{R}}$$

An incomplete list of advantages:

- Unique minima
- GD guaranteed to converge to it
- Rates of convergence
- …

Motivation: loss-minimisation & convexity

Classical problem: find θ^*

Easy to do via GD if ℓ (strictly) convex

$$\boxed{\theta^* = \mathrm{argmin}_{\theta \in \Theta}\, \ell(\theta) \text{ for some loss } \ell : \Theta \to \mathbb{R}}$$

An incomplete list of advantages:

- Unique minima
- GD guaranteed to converge to it
- Rates of convergence
- …

But: many losses we care about NOT convex

e.g., $\ell(\theta) = \sum_{i=1}^{n} (y_i - \mathrm{NN}_\theta(x_i))^2$ for NN_θ a NN parameterised by θ

Motivation: loss-minimisation & convexity

Classical problem: find θ^*

Easy to do via GD if ℓ (strictly) convex

$$\boxed{\theta^* = \mathrm{argmin}_{\theta \in \Theta}\, \ell(\theta) \text{ for some loss } \ell : \Theta \to \mathbb{R}}$$

Key question: in the absence of a convex loss, can we force convexity?

=> Standard answer: you can often make a loss ‘more convex’ via regularisation

Motivation: loss-minimisation & convexity

Classical problem: find θ^*

Easy to do via GD if ℓ (strictly) convex

$$\boxed{\theta^* = \mathrm{argmin}_{\theta \in \Theta}\, \ell(\theta) \text{ for some loss } \ell : \Theta \to \mathbb{R}}$$

Key question: in the absence of a convex loss, can we force convexity?

=> Standard answer: you can often make a loss 'more convex' via regularisation

e.g., Tikhonov regularisation: $\ell(\theta) = \|A\theta - b\|_2^2$ not strictly convex if underdetermined system;

but $\ell(\theta) = \|A\theta - b\|_2^2 + \|\Gamma\theta\|_2^2$ is strictly convex

Motivation: loss-minimisation & convexity

Classical problem: find θ^*

Easy to do via GD if ℓ (strictly) convex

$$\boxed{\theta^* = \mathrm{argmin}_{\theta \in \Theta}\, \ell(\theta) \text{ for some loss } \ell : \Theta \to \mathbb{R}}$$

Key question: in the absence of a convex loss, can we force convexity?

=> Standard answer: you can often make a loss ‘more convex’ via regularisation

e.g., Tikhonov regularisation: $\ell(\theta) = \|A\theta - b\|_2^2$ not strictly convex if underdetermined system;

but $\ell(\theta) = \|A\theta - b\|_2^2 + \|\Gamma\theta\|_2^2$ is strictly convex

Problem: most ‘convexification’ strategies use some structure in ℓ ; can we do without that?

Motivation: loss-minimisation & convexity

$$\min_{\theta \in \Theta} \ell(\theta) \xrightarrow{\textbf{Step 1: } \text{probabilistic lifting}} \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \int \ell(\theta) dQ(\theta)$$

Motivation: loss-minimisation & convexity

$$\min_{\theta \in \Theta} \ell(\theta) \xrightarrow{\textbf{Step 1: } \text{probabilistic lifting}} \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \int \ell(\theta) dQ(\theta)$$

$L(Q) = \int \ell(\theta) dQ(\theta)$ convex in Q

Motivation: loss-minimisation & convexity

$$\min_{\theta \in \Theta} \ell(\theta) \xrightarrow{\textbf{Step 1: } \text{probabilistic lifting}} \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \int \ell(\theta) dQ(\theta)$$

$L(Q) = \int \ell(\theta) dQ(\theta)$ convex in Q $\longrightarrow$ $\text{argmin}_Q L(Q) = \{\delta_{\theta*}$ for all $\theta* \in \text{argmin}_{\theta \in \Theta} \ell(\theta)\}$

Motivation: loss-minimisation & convexity

$$\min_{\theta\in\Theta} \ell(\theta) \qquad \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \int \ell(\theta)dQ(\theta)$$

Step 1: probabilistic lifting

$L(Q)$ has no unique minimum unless $\ell(\theta)$ does!

$L(Q) = \int \ell(\theta)dQ(\theta)$ convex in Q $\longrightarrow$ $\text{argmin}_Q L(Q) = \{\delta_{\theta*}$ for all $\theta* \in \text{argmin}_{\theta\in\Theta} \ell(\theta)\}$

Motivation: loss-minimisation & convexity

$$\min_{\theta \in \Theta} \ell(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \int \ell(\theta) dQ(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta) dQ(\theta) + \lambda D(Q, P) \right\}$$

Step 1: probabilistic lifting **Step 2:** convexification through regularisation

Motivation: loss-minimisation & convexity

$\lambda \in \mathbb{R}_+$, $D : \mathscr{P}(\mathbb{R}^J) \times \mathscr{P}(\mathbb{R}^J) \mapsto \mathbb{R}_+$ a divergence [i.e., $D(Q, P) = 0 \iff Q = P$]

$$\min_{\theta \in \Theta} \ell(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \int \ell(\theta) dQ(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta) dQ(\theta) + \lambda D(Q, P) \right\}$$

Step 1: probabilistic lifting **Step 2:** convexification through regularisation

Motivation: loss-minimisation & convexity

$\lambda \in \mathbb{R}_+$, $D : \mathscr{P}(\mathbb{R}^J) \times \mathscr{P}(\mathbb{R}^J) \mapsto \mathbb{R}_+$ a divergence [i.e., $D(Q, P) = 0 \iff Q = P$]

$$\min_{\theta \in \Theta} \ell(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \int \ell(\theta) dQ(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta) dQ(\theta) + \lambda D(Q, P) \right\}$$

Step 1: probabilistic lifting **Step 2:** convexification through regularisation

P is a reference meausure/like a Bayesian prior

Motivation: loss-minimisation & convexity

$$\min_{\theta \in \Theta} \ell(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \int \ell(\theta) dQ(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta) dQ(\theta) + \lambda D(Q, P) \right\}$$

Step 1: probabilistic lifting **Step 2:** convexification through regularisation

$L(Q) = \int \ell(\theta) dQ(\theta) + \lambda D(Q, P)$ strictly convex in Q if D is strictly convex in Q

Motivation: loss-minimisation & convexity

$$\min_{\theta \in \Theta} \ell(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \int \ell(\theta) dQ(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta) dQ(\theta) + \lambda D(Q, P) \right\}$$

Step 1: probabilistic lifting **Step 2:** convexification through regularisation

$L(Q) = \int \ell(\theta) dQ(\theta) + \lambda D(Q, P)$ strictly convex in Q if D is strictly convex in Q

$\Longrightarrow$ intuitively: strict convexity should guarantee a unique minimiser (and we prove this formally)

Motivation: loss-minimisation & convexity

$$\min_{\theta\in\Theta} \ell(\theta) \xrightarrow{\textbf{Step 1: } \text{probabilistic lifting}} \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \int \ell(\theta)dQ(\theta) \xrightarrow{\textbf{Step 2: } \text{convexification through regularisation}} \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta)dQ(\theta) + \lambda D(Q,P) \right\}$$

(1) Instead of computing $\theta^* = \text{argmin}_{\theta\in\Theta}\, \ell(\theta)$, we compute

$$Q^* = \text{argmin}_{Q\in\mathscr{P}(\Theta)} \left\{ \int \ell(\theta)dQ(\theta) + \lambda D(Q,P) \right\}$$

Motivation: loss-minimisation & convexity

$$\min_{\theta\in\Theta} \ell(\theta) \xrightarrow{\textbf{Step 1:}\ \text{probabilistic lifting}} \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \int \ell(\theta)\,dQ(\theta) \xrightarrow{\textbf{Step 2:}\ \text{convexification through regularisation}} \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta)\,dQ(\theta) + \lambda D(Q, P) \right\}$$

(1) Instead of computing $\theta^* = \mathrm{argmin}_{\theta\in\Theta}\,\ell(\theta),$ we compute

$$Q^* = \mathrm{argmin}_{Q\in\mathscr{P}(\Theta)} \left\{ \int \ell(\theta)dQ(\theta) + \lambda D(Q, P) \right\}$$

(2) And then perform prediction/downstream tasks with $\theta^* \sim Q^*$

Motivation: loss-minimisation & convexity

$$\min_{\theta\in\Theta} \ell(\theta) \qquad \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \int \ell(\theta)dQ(\theta) \qquad \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta)dQ(\theta) + \lambda D(Q,P) \right\}$$

Step 1: probabilistic lifting **Step 2:** convexification through regularisation

- \+ Universally applicable for any loss ℓ e.g., NN-based losses
- \+ strictly convex (we can get theoretical guarantees!)
- \+ These types of objectives have already been studied in generalised Bayes & PAC-Bayes!

Motivation: loss-minimisation & convexity

$$\min_{\theta \in \Theta} \ell(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \int \ell(\theta) dQ(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta) dQ(\theta) + \lambda D(Q, P) \right\}$$

Step 1: probabilistic lifting **Step 2:** convexification through regularisation

+ Universally applicable for any loss ℓ e.g., NN-based losses
+ strictly convex (we can get theoretical guarantees!)
+ These types of objectives have already been studied in generalised Bayes & PAC-Bayes!

– The problem is infinite-dimensional...

$\Longrightarrow$ exact minimisers unknown for most regularisers

$\Longrightarrow$ BUT: all existing approximation algorithms rely on knowing the exact minimiser...

Motivation: loss-minimisation & convexity

$$\min_{\theta \in \Theta} \ell(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \int \ell(\theta) dQ(\theta) \qquad \min_{Q \in \mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta) dQ(\theta) + \lambda D(Q, P) \right\}$$

Step 1: probabilistic lifting **Step 2:** convexification through regularisation

+ Universally applicable for any loss ℓ e.g., NN-based losses
+ strictly convex (we can get theoretical guarantees!)
+ These types of objectives have already been studied in generalised Bayes & PAC-Bayes!

Research problem we solve!

+ The problem is infinite-dimensional...
  - $\Longrightarrow$ exact minimisers unknown for most regularisers
  - $\Longrightarrow$ BUT: all existing approximation algorithms rely on knowing the exact minimiser...

Connections: one objective, many meanings

Bayes posterior (for $\lambda = 1, D = \mathrm{KL}, \ell(\theta) = -\log p(x_{1:n} \mid \theta)$)

$$\frac{p(x_{1:n} \mid \theta) dP(\theta)}{\int p(x_{1:n} \mid \theta) dP(\theta)} = \operatorname{argmin}_{Q \in \mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim Q} \left[-\log p(x_{1:n} \mid \theta) \right] + \mathrm{KL}(Q \| P) \right\}$$

Connections: one objective, many meanings

Bayes posterior (for $\lambda = 1, D = \mathrm{KL}, \ell(\theta) = -\log p(x_{1:n} \mid \theta)$)

$$\frac{p(x_{1:n} \mid \theta) dP(\theta)}{\int p(x_{1:n} \mid \theta) dP(\theta)} = \operatorname{argmin}_{Q \in \mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim Q} \left[-\log p(x_{1:n} \mid \theta) \right] + \mathrm{KL}(Q \| P) \right\}$$

Generalised Bayes posterior (for any loss $\ell, \lambda > 0$, and $D = \mathrm{KL}$)

$$\frac{\exp\{-\frac{1}{\lambda}\ell(\theta)\} dP(\theta)}{\int \exp\{-\frac{1}{\lambda}\ell(\theta)\} dP(\theta)} = \operatorname{argmin}_{Q \in \mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim Q} \left[\ell(\theta) \right] + \lambda \mathrm{KL}(Q \| P) \right\}$$

An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference, **Knoblauch, J.**, Jewson, J., & Damoulas, T., JMLR (2022).
A general framework for updating belief distributions, Bissiri, P.G., Holmes, C., & Walker, S.G. (2016)

Connections: one objective, many meanings

Generalised Bayes posterior (for any loss $\ell, \lambda > 0$, and $D = \mathrm{D}_f$ an f-divergence)

$$\nabla f^* \left(Z - \frac{\ell(\theta)}{\lambda} \right) dP(\theta) = \mathrm{argmin}_{Q \in \mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim Q} \left[\ell(\theta) \right] + \lambda \mathrm{D}_f(Q \| P) \right\}$$

An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference, **Knoblauch, J.**, Jewson, J., & Damoulas, T., JMLR (2022).
Non-exponentially weighted aggregation: regret bounds for unbounded loss functions, Alquier, P. (2021)

Connections: one objective, many meanings

Generalised Bayes posterior (for any loss $\ell, \lambda > 0$, and $D = \mathrm{D}_f$ an f-divergence)

$$\nabla f^* \left(Z - \frac{\ell(\theta)}{\lambda} \right) dP(\theta) = \operatorname{argmin}_{Q \in \mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim Q} \left[\ell(\theta) \right] + \lambda \mathrm{D}_f(Q \| P) \right\}$$

f^* is the Fenchel conjugate of f

An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference, **Knoblauch, J.**, Jewson, J., & Damoulas, T., JMLR (2022).
Non-exponentially weighted aggregation: regret bounds for unbounded loss functions, Alquier, P. (2021)

Connections: one objective, many meanings

Generalised Bayes posterior (for any loss $\ell, \lambda > 0$, and $D = \mathrm{D}_f$ an f-divergence)

$$\nabla f^* \left(Z - \frac{\ell(\theta)}{\lambda} \right) dP(\theta) = \mathrm{argmin}_{Q \in \mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim Q} \left[\ell(\theta) \right] + \lambda \mathrm{D}_f(Q \| P) \right\}$$

Z is the normaliser/unique constant defined via $\int \nabla f^* \left(Z - \frac{\ell(\theta)}{\lambda} \right) dP(\theta) = 1$

f^* is the Fenchel conjugate of f

An Optimization-centric View on Bayes' Rule: Reviewing and Generalizing Variational Inference, **Knoblauch, J.**, Jewson, J., & Damoulas, T., JMLR (2022).
Non-exponentially weighted aggregation: regret bounds for unbounded loss functions, Alquier, P. (2021)

Connections: one objective, many meanings

PAC-Bayesian bound; holds with high probability under (usually strict) assumptions like $x_i \overset{iid}{\sim} \mathbb{P}, a \leq \ell(\theta) \leq b,$ and λ depending on moment conditions.

$$\mathbb{E}_{X\sim\mathbb{P}}[\ell(\theta, X)] \leq \min_{Q\in\mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta\sim Q}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\theta, x_i)\right] + \lambda \mathrm{D}(Q\|P) \right\} + \mathcal{O}(n^{-1})$$

Connections: one objective, many meanings

PAC-Bayesian bound; holds with high probability under (usually strict) assumptions like $x_i \overset{iid}{\sim} \mathbb{P}, a \leq \ell(\theta) \leq b,$ and λ depending on moment conditions.

$$\mathbb{E}_{X\sim\mathbb{P}}[\ell(\theta, X)] \leq \min_{Q\in\mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta\sim Q}\left[\frac{1}{n}\sum_{i=1}^{n}\ell(\theta, x_i)\right] + \lambda \mathrm{D}(Q\|P) \right\} + \mathcal{O}(n^{-1})$$

Most PAC-Bayes bounds rely on $D = \mathrm{KL}$, but not all!

User-friendly introduction to PAC-Bayes, Alquier, P. arXiv:2110.11216 (2022)
Simpler PAC-Bayesian bounds for hostile data, Alquier, P. & Guedj, B., Machine Learning (2018).
PAC-Bayesian bounds based on the Renyi divergence, Begin, L., Germain, P., Laviolette, F., & Roy, J.-F. , AISTATS (2016).
Wasserstein PAC-Bayes learning: a bridge between generalisation and optimisation, Haddouche, M. & Guedj, B. arXiv:2304.07048 (2023)

Existing algorithms for computation

$$\frac{\exp\{-\frac{1}{\lambda}\ell(\theta)\}dP(\theta)}{\int \exp\{-\frac{1}{\lambda}\ell(\theta)\}dP(\theta)} = \operatorname{argmin}_{Q \in \mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim Q}\left[\ell(\theta)\right] + \lambda \mathrm{KL}(Q \| P) \right\}$$

Existing algorithms for computation

$$\frac{\exp\{-\frac{1}{\lambda}\ell(\theta)\}dP(\theta)}{\int \exp\{-\frac{1}{\lambda}\ell(\theta)\}dP(\theta)} = \operatorname{argmin}_{Q \in \mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim Q}\left[\ell(\theta)\right] + \lambda \mathrm{KL}(Q \| P) \right\}$$

$$\Bigg\downarrow$$

Approximation via Variational Inference (VI)
(Variational family = $\mathcal{Q} \subset \mathscr{P}(\Theta)$)

$$Q^*_{\mathrm{VI}} = \operatorname{argmin}_{Q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim Q}\left[\ell(\theta)\right] + \lambda \mathrm{KL}(Q \| P) \right\}$$

Existing algorithms for computation

$$\frac{\exp\{-\frac{1}{\lambda}\ell(\theta)\}dP(\theta)}{\int \exp\{-\frac{1}{\lambda}\ell(\theta)\}dP(\theta)} = \operatorname{argmin}_{Q\in\mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta\sim Q}\left[\ell(\theta)\right] + \lambda \mathrm{KL}(Q\|P) \right\}$$

Approximation via Variational Inference (VI)
(Variational family = $\mathcal{Q} \subset \mathscr{P}(\Theta)$)

$$Q^*_{\mathrm{VI}} = \operatorname{argmin}_{Q\in\mathcal{Q}} \left\{ \mathbb{E}_{\theta\sim Q}\left[\ell(\theta)\right] + \lambda \mathrm{KL}(Q\|P) \right\}$$

Q^*_{VI} ill-defined/may not exist or be unique
(parameterisation of $\mathcal{Q}$ breaks convexity!)

Existing algorithms for computation

$$\frac{\exp\{-\frac{1}{\lambda}\ell(\theta)\}dP(\theta)}{\int \exp\{-\frac{1}{\lambda}\ell(\theta)\}dP(\theta)} = \operatorname{argmin}_{Q \in \mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta \sim Q}\left[\ell(\theta)\right] + \lambda \mathrm{KL}(Q \| P) \right\}$$

$\downarrow$

Methods relying on posterior having analytical form

(1) Sampling-based (e.g., MCMC methods)

(2) Optimisation-based (e.g., Wasserstein gradient flows minimising $Q \mapsto D(Q, Q^*)$ for a divergence D)

Existing algorithms for computation

$$\frac{\exp\{-\frac{1}{\lambda}\ell(\theta)\}dP(\theta)}{\int \exp\{-\frac{1}{\lambda}\ell(\theta)\}dP(\theta)} = \operatorname{argmin}_{Q\in\mathscr{P}(\Theta)} \left\{ \mathbb{E}_{\theta\sim Q}\left[\ell(\theta)\right] + \lambda \mathrm{KL}(Q\|P) \right\}$$

Methods relying on posterior having analytical form

(1) Sampling-based (e.g., MCMC methods)

(2) Optimisation-based (e.g., Wasserstein gradient flows minimising $Q \mapsto D(Q, Q^*)$ for a divergence D)

Algorithms rely on analytical forms of Q^*!

A new proposal

Key idea: algorithms that use strict convexity of $Q \mapsto L(Q)$ via GD

+ these algorithms would NOT rely on analytical forms!

A new proposal

Key idea: algorithms that use strict convexity of $Q \mapsto L(Q)$ via GD

➕ these algorithms would NOT rely on analytical forms!

Recap GD (with constant step size & in Euclidean spaces):

Initialisation: $\theta_0 \in \Theta$

$$\theta_{k+1} = \operatorname{argmin}_{\theta \in \Theta} \left\{ \ell(\theta) + \frac{1}{2\eta} \|\theta - \theta_k\|_2^2 \right\} = \theta_k - \eta \cdot \nabla \ell(\theta_k); \quad k \in \mathbb{N}, \eta > 0$$

A new proposal

Key idea: algorithms that use strict convexity of $Q \mapsto L(Q)$ via GD

➕ these algorithms would NOT rely on analytical forms!

Recap GD (with constant step size & in Euclidean spaces):

Initialisation: $\theta_0 \in \Theta$

$$\theta_{k+1} = \operatorname{argmin}_{\theta \in \Theta} \left\{ \ell(\theta) + \frac{1}{2\eta} \|\theta - \theta_k\|_2^2 \right\} = \theta_k - \eta \cdot \nabla \ell(\theta_k); \quad k \in \mathbb{N}, \eta > 0$$

extends to infinite dimensions

does NOT extend to infinite dimensions

A new proposal

Key idea: algorithms that use strict convexity of $Q \mapsto L(Q)$ via GD

➕ these algorithms would NOT rely on analytical forms!

Recap GD (with constant step size & in Euclidean spaces):

Initialisation: $\theta_0 \in \Theta$

$$\theta_{k+1} = \mathrm{argmin}_{\theta \in \Theta} \left\{ \ell(\theta) + \frac{1}{2\eta} \|\theta - \theta_k\|_2^2 \right\} = \theta_k - \eta \cdot \nabla \ell(\theta_k); \quad k \in \mathbb{N}, \eta > 0$$

Recap Gradient Flow (GF) [$\approx$ continuous version of GD]:

Continuous interpolation: $\theta^\eta : [0,\infty) \to \Theta$ s.t. $\theta^\eta(t) = \theta_{t/\eta}$ for $t \in \{0,\eta,2\eta,\ldots\}$

A new proposal

Key idea: algorithms that use strict convexity of $Q \mapsto L(Q)$ via GD

➕ these algorithms would NOT rely on analytical forms!

Recap GD (with constant step size & in Euclidean spaces):

Initialisation: $\theta_0 \in \Theta$

$$\theta_{k+1} = \mathrm{argmin}_{\theta \in \Theta} \left\{ \ell(\theta) + \frac{1}{2\eta} \|\theta - \theta_k\|_2^2 \right\} = \theta_k - \eta \cdot \nabla \ell(\theta_k); \quad k \in \mathbb{N}, \eta > 0$$

Recap Gradient Flow (GF) [$\approx$ continuous version of GD]:

Continuous interpolation: $\theta^\eta : [0,\infty) \to \Theta$ s.t. $\theta^\eta(t) = \theta_{t/\eta}$ for $t \in \{0, \eta, 2\eta, \ldots\}$

$\theta^\eta(t) \longrightarrow \theta_*(t)$ as $\eta \to 0$, with $\theta_0 = \theta_*(0)$

A new proposal

Key idea: algorithms that use strict convexity of $Q \mapsto L(Q)$ via GD

+ these algorithms would NOT rely on analytical forms!

Recap GD (with constant step size & in Euclidean spaces):

Initialisation: $\theta_0 \in \Theta$

$$\theta_{k+1} = \text{argmin}_{\theta \in \Theta} \left\{ \ell(\theta) + \frac{1}{2\eta} \|\theta - \theta_k\|_2^2 \right\} = \theta_k - \eta \cdot \nabla \ell(\theta_k); \quad k \in \mathbb{N}, \eta > 0$$

Recap Gradient Flow (GF) [$\approx$ continuous version of GD]:

Continuous interpolation: $\theta^\eta : [0,\infty) \to \Theta$ s.t. $\theta^\eta(t) = \theta_{t/\eta}$ for $t \in \{0, \eta, 2\eta, \dots\}$

$\theta^\eta(t) \longrightarrow \theta_*(t)$ as $\eta \to 0$, with $\theta_0 = \theta_*(0)$

$\theta_*'(t) = -\nabla \ell(\theta_*(t))$, with $\theta_*(0) = \theta_0$ (GF solves this ODE)

A new proposal

Intuitive construction of GD on probability measures (with 2nd moment):

Initialisation: $Q_0 \in \mathscr{P}_2(\Theta)$

$$Q_{k+1} = \operatorname{argmin}_{Q \in \mathscr{P}_2(\Theta)} \left\{ L(Q) + \frac{1}{2\eta} W_2(Q, Q_k)^2 \right\}; \quad k \in \mathbb{N}, \eta > 0$$

Squared 2-Wasserstein distance [$\approx$ natural analogy of squared Euclidean distance for measures]

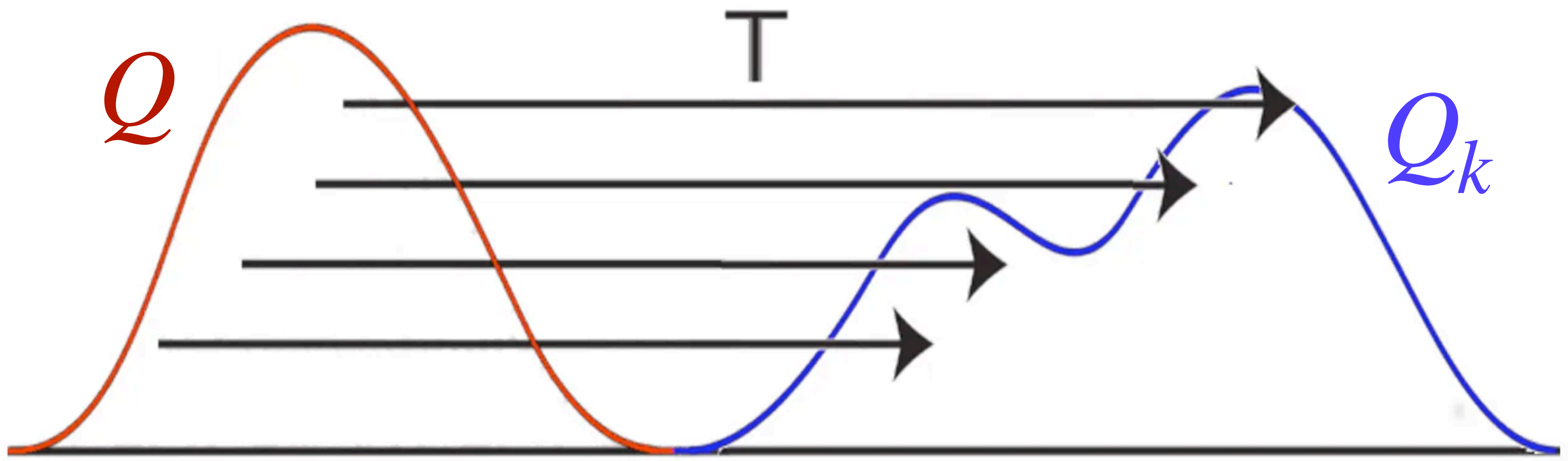
A new proposal

Intuitive construction of GD on probability measures (with 2nd moment):

Initialisation: $Q_0 \in \mathscr{P}_2(\Theta)$

$$Q_{k+1} = \operatorname{argmin}_{Q \in \mathscr{P}_2(\Theta)} \left\{ L(Q) + \frac{1}{2\eta} W_2(Q, Q_k)^2 \right\}; \quad k \in \mathbb{N}, \eta > 0$$

Squared 2-Wasserstein distance [$\approx$ natural analogy of squared Euclidean distance for measures]



A new proposal

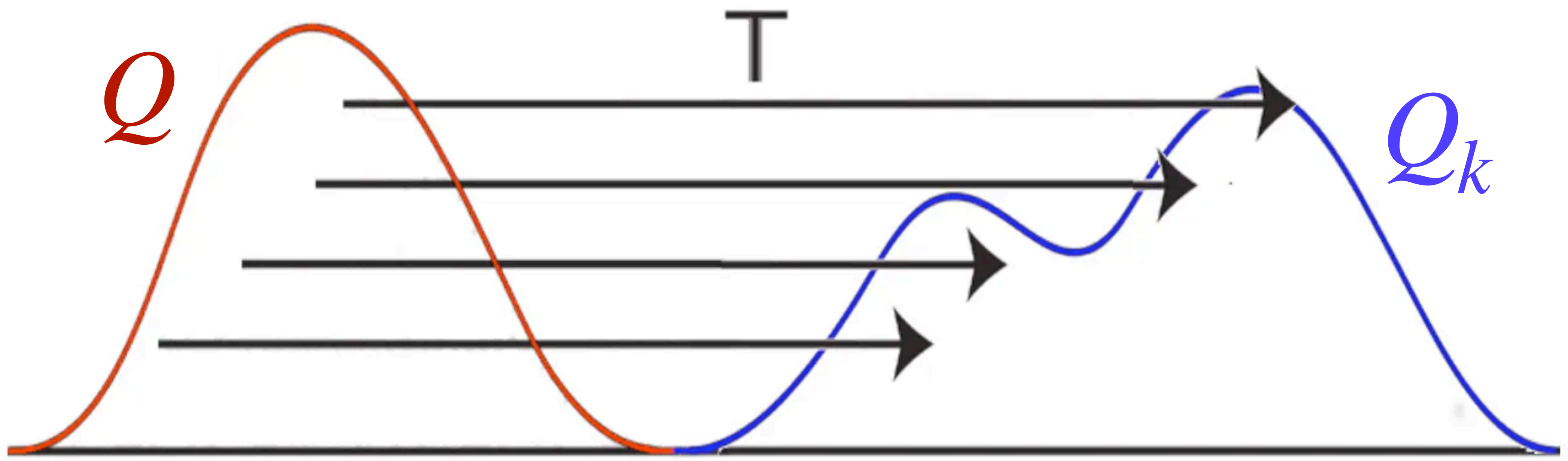
Intuitive construction of GD on probability measures (with 2nd moment):

Initialisation: $Q_0 \in \mathscr{P}_2(\Theta)$

$$Q_{k+1} = \text{argmin}_{Q \in \mathscr{P}_2(\Theta)} \left\{ L(Q) + \frac{1}{2\eta} W_2(Q, Q_k)^2 \right\}; \quad k \in \mathbb{N}, \eta > 0$$

Squared 2-Wasserstein distance [≈ natural analogy of squared Euclidean distance for measures]

$$W_2(Q, Q_k)^2 = \inf_{\substack{C \in \mathscr{P}(\Theta \times \Theta) \text{ s.t.} \\ \int C(d\theta, x) = Q(x), \\ \int C(x, d\theta) = Q_k(x)}} \left\{ \int \|\theta - \theta'\|_2^2 \, C(d(\theta, \theta')) \right\}$$



A new proposal

Intuitive construction of GD on probability measures (with 2nd moment):

Initialisation: $Q_0 \in \mathscr{P}_2(\Theta)$

$$Q_{k+1} = \operatorname{argmin}_{Q \in \mathscr{P}_2(\Theta)} \left\{ L(Q) + \frac{1}{2\eta} W_2(Q, Q_k)^2 \right\}; \quad k \in \mathbb{N}, \eta > 0$$

Wasserstein Gradient Flow (GF) [$\approx$ continuous version of GD construction]:

As $\eta \to 0$, we get continuously indexed collection of measures $\{Q_t\}_{t \geq 0}$

A new proposal

Intuitive construction of GD on probability measures (with 2nd moment):

Initialisation: $Q_0 \in \mathscr{P}_2(\Theta)$

$$Q_{k+1} = \text{argmin}_{Q \in \mathscr{P}_2(\Theta)} \left\{ L(Q) + \frac{1}{2\eta} W_2(Q, Q_k)^2 \right\}; \quad k \in \mathbb{N}, \eta > 0$$

Wasserstein Gradient Flow (GF) [$\approx$ continuous version of GD construction]:

As $\eta \to 0$, we get continuously indexed collection of measures $\{Q_t\}_{t \geq 0}$

Hope/idea 1: $\text{argmin}_{Q \in \mathscr{P}_2(\Theta)} L(Q) = Q^* = \lim_{t \to \infty} Q_t$

A new proposal

Intuitive construction of GD on probability measures (with 2nd moment):

Initialisation: $Q_0 \in \mathscr{P}_2(\Theta)$

$$Q_{k+1} = \text{argmin}_{Q \in \mathscr{P}_2(\Theta)} \left\{ L(Q) + \frac{1}{2\eta} W_2(Q, Q_k)^2 \right\}; \quad k \in \mathbb{N}, \eta > 0$$

Wasserstein Gradient Flow (GF) [$\approx$ continuous version of GD construction]:

As $\eta \to 0$, we get continuously indexed collection of measures $\{Q_t\}_{t \geq 0}$

Hope/idea 1: $\text{argmin}_{Q \in \mathscr{P}_2(\Theta)} L(Q) = Q^* = \lim_{t \to \infty} Q_t$

Hope/idea 2: rewriting things in terms of densities!

A new proposal

Intuitive construction of GD on probability measures (with 2nd moment):

Initialisation: $Q_0 \in \mathscr{P}_2(\Theta)$

$$Q_{k+1} = \text{argmin}_{Q \in \mathscr{P}_2(\Theta)} \left\{ L(Q) + \frac{1}{2\eta} W_2(Q, Q_k)^2 \right\}; \quad k \in \mathbb{N}, \eta > 0$$

Wasserstein Gradient Flow (GF) [$\approx$ continuous version of GD construction]:

As $\eta \to 0$, we get continuously indexed collection of measures $\{Q_t\}_{t \geq 0}$

Hope/idea 1: $\text{argmin}_{Q \in \mathscr{P}_2(\Theta)} L(Q) = Q^* = \lim_{t \to \infty} Q_t$

Hope/idea 2: rewriting things in terms of densities!

If L is sufficiently smooth and Q_0 has density $q_0 = q(0, \cdot)$,

$\implies q(t, \cdot)$ density of Q_t, $\quad \partial_t q(t, \theta) = \nabla \cdot \Big(q(t, \theta) \nabla_W L(Q_t)(\theta) \Big)$

A new proposal

Intuitive construction of GD on probability measures (with 2nd moment):

Initialisation: $Q_0 \in \mathscr{P}_2(\Theta)$

$$Q_{k+1} = \text{argmin}_{Q \in \mathscr{P}_2(\Theta)} \left\{ L(Q) + \frac{1}{2\eta} W_2(Q, Q_k)^2 \right\}; \quad k \in \mathbb{N}, \eta > 0$$

Wasserstein Gradient Flow (GF) [$\approx$ continuous version of GD construction]:

As $\eta \to 0$, we get continuously indexed collection of measures $\{Q_t\}_{t \geq 0}$

Hope/idea 1: $\text{argmin}_{Q \in \mathscr{P}_2(\Theta)} L(Q) = Q^* = \lim_{t \to \infty} Q_t$

A density evolution equation & PDE
(let's unpack this...)

Hope/idea 2: rewriting things in terms of densities!

If L is sufficiently smooth and Q_0 has density $q_0 = q(0, \cdot)$,

$\implies q(t, \cdot\,)$ density of Q_t, $\quad \partial_t q(t, \theta) = \nabla \cdot \big(q(t, \theta) \, \nabla_W L(Q_t)(\theta) \big)$

A new proposal

$$\boxed{\theta_*{}'(t) = -\nabla \ell(\theta_*(t)), \text{ with } \theta_*(0) = \theta_0}$$

gradient flow

infitesimal change (in time)

initial condition

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \big(q(t,\theta)\,\nabla_W L(Q_t)(\theta)\big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

A new proposal

$$\boxed{\theta_*{}'(t) = -\nabla \ell(\theta_*(t)), \ \text{with } \theta_*(0) = \theta_0}$$

gradient flow

infitesimal change (in time)

ODE

initial condition

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta)\, \nabla_W L(Q_t)(\theta) \Big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

PDE

A new proposal

gradient

$$\boxed{\theta_*{}'(t) = -\nabla \ell(\theta_*(t)), \text{ with } \theta_*(0) = \theta_0}$$

gradient flow

infitesimal change (in time)

ODE

initial condition

divergence operator: $\nabla \cdot f = \sum \partial_j f_j$

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \big(q(t,\theta) \nabla_W L(Q_t)(\theta) \big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

PDE

A new proposal

gradient

$$\boxed{\theta_*{}'(t) = -\nabla \ell(\theta_*(t)),\ \text{with } \theta_*(0) = \theta_0}$$

gradient flow

infitesimal change (in time)

ODE

initial condition

divergence operator: $\nabla \cdot f = \sum \partial_j f_j$

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \big(q(t,\theta)\, \nabla_W L(Q_t)(\theta)\big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

PDE

Wasserstein gradient

(Note 1: gradient of $L(Q)$ w.r.t. W_2)

(Note 2: here = gradient of first variation of L at Q)

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \big(q(t,\theta)\,\nabla_W L(Q_t)(\theta)\big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

How to translate this into computational algorithm?

- PDE solvers? $\implies$ computationally infeasible for high/medium high dimension of Θ

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta)\, \nabla_W L(Q_t)(\theta) \Big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

How to translate this into computational algorithm?

- PDE solvers? $\Longrightarrow$ computationally infeasible for high/medium high dimension of Θ
- Evolving N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$? $\Longrightarrow$ feasible, even in high dimensions

$\Longrightarrow$ Question: which choices of $L(Q)$ lead to such a particle evolution framework?

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta)\, \nabla_W L(Q_t)(\theta) \Big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

Want: evolve N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$

$\Longrightarrow$ Question: which choices of $L(Q)$ lead to such a particle evolution framework?

$\Longrightarrow$ Note 1: particle methods often come from thermodynamics

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \left(q(t,\theta)\, \nabla_W L(Q_t)(\theta) \right), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

Want: evolve N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$

$\Longrightarrow$ Question: which choices of $L(Q)$ lead to such a particle evolution framework?

$\Longrightarrow$ Note 1: particle methods often come from thermodynamics

$\Longrightarrow$ Note 2: recall classical free energy

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

energy of particles sampled from Q

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta)\, \nabla_W L(Q_t)(\theta)\Big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

Want: evolve N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$

$\Longrightarrow$ Question: which choices of $L(Q)$ lead to such a particle evolution framework?

$\Longrightarrow$ Note 1: particle methods often come from thermodynamics

$\Longrightarrow$ Note 2: recall classical free energy

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

energy of particles sampled from Q

external potential (acts on particles individually)

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \left(q(t,\theta)\, \nabla_W L(Q_t)(\theta) \right), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

Want: evolve N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$

$\implies$ Question: which choices of $L(Q)$ lead to such a particle evolution framework?

$\implies$ Note 1: particle methods often come from thermodynamics

$\implies$ Note 2: recall classical free energy

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

energy of particles sampled from Q

external potential (acts on particles individually)

pairwise interaction potential

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \left(q(t,\theta)\,\nabla_W L(Q_t)(\theta)\right), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

Want: evolve N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$

$\implies$ Question: which choices of $L(Q)$ lead to such a particle evolution framework?

$\implies$ Note 1: particle methods often come from thermodynamics

$\implies$ Note 2: recall classical free energy

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

energy of particles sampled from Q

external potential (acts on particles individually)

pairwise interaction potential

system's overall entropy

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta)\, \nabla_W L(Q_t)(\theta) \Big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

Want: evolve N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$

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

$\implies$ For this, we know how to build an interacting particle system:

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta)\,\nabla_W L(Q_t)(\theta)\Big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

Want: evolve N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$

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

$\Longrightarrow$ For this, we know how to build an interacting particle system:

Step 1: Sample $\theta_n(0) \sim Q_0$, $n = 1,2,\ldots N$

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \left(q(t,\theta)\, \nabla_W L(Q_t)(\theta)\right), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

Want: evolve N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$

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

$\Longrightarrow$ For this, we know how to build an interacting particle system:

Step 1: Sample $\theta_n(0) \sim Q_0$, $n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

Brownian motion s.t. $\{B_n(t)\}_{t\geq 0}$ stochastically independent

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

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta)\, \nabla_W L(Q_t)(\theta) \Big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

Want: evolve N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$

$$L^{\text{fe}}(Q) \overset{!}{=} L(Q) = \int \ell(\theta)dQ(\theta) + D(Q, P)$$

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta)\, \nabla_W L(Q_t)(\theta) \Big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

Want: evolve N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$

$$L^{\text{fe}}(Q) \overset{!}{=} L(Q) = \int \ell(\theta) dQ(\theta) + D(Q,P) \quad \text{for } D(Q,P) = \frac{\lambda_1}{2}\text{MMD}^2(Q,P) + \lambda_2 \text{KL}(Q\|P)$$

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta)\,\nabla_W L(Q_t)(\theta)\Big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

Want: evolve N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$

$$L^{\text{fe}}(Q) \stackrel{!}{=} L(Q) = \int \ell(\theta) dQ(\theta) + D(Q,P) \quad \text{for } D(Q,P) = \frac{\lambda_1}{2}\text{MMD}^2(Q,P) + \lambda_2 \text{KL}(Q\|P)$$

$$\text{MMD}^2(Q,P) = \|\mu_P - \mu_Q\|_\kappa^2$$

KME defined via κ

RKHS defined via κ

A new proposal

$$\boxed{\partial_t q(t,\theta) = \nabla \cdot \Big(q(t,\theta)\, \nabla_W L(Q_t)(\theta) \Big), \quad \text{with } q(0,\cdot) = Q_0}$$

Wasserstein gradient flow

Want: evolve N particles $\theta_n(t)$ s.t. $\theta_n(t) \sim Q_t$ for all $n = 1,2,\ldots N$

$$L^{\text{fe}}(Q) \stackrel{!}{=} L(Q) = \int \ell(\theta) dQ(\theta) + D(Q,P) \quad \text{for } D(Q,P) = \frac{\lambda_1}{2}\text{MMD}^2(Q,P) + \lambda_2\,\text{KL}(Q \| P)$$

For $V(\theta) = \ell(\theta) - \lambda_1 \mu_P(\theta) - \lambda_2 \log p(\theta)$ and

$$\text{MMD}^2(Q,P) = \|\mu_P - \mu_Q\|_\kappa^2$$

KME defined via κ

RKHS defined via κ

Summary: a new proposal

$$Q^* = \mathrm{argmin}_{Q \in \mathscr{P}(\Theta)} L(Q)$$

$$L(Q) = \int \ell(\theta) dQ(\theta) + \frac{\lambda_1}{2} \mathrm{MMD}^2(Q, P) + \lambda_2 \mathrm{KL}(Q \| P)$$

Step 1: Sample $\theta_n(0) \sim Q_0$, $n = 1,2,\ldots N$

Brownian motion s.t. $\{B_n(t)\}_{t \geq 0}$ stochastically independent

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) - \lambda_1 \nabla \mu_P(\theta_n(t)) - \lambda_2 \nabla \log p(\theta_n(t)) + \frac{\lambda_1}{N} \sum_{i=1}^{N} \nabla_1 \kappa(\theta_n(t), \theta_i(t)) \right) dt + \sqrt{2\lambda_2 dB_n(t)}$$

$\Longrightarrow$ Hope: $\frac{1}{N} \sum_{n=1}^{N} \theta_n(T) \approx Q^*$ for large T, N.

Summary: a new proposal

$$Q^* = \mathrm{argmin}_{Q \in \mathscr{P}(\Theta)} L(Q)$$

$$L(Q) = \int \ell(\theta) dQ(\theta) + \frac{\lambda_1}{2} \mathrm{MMD}^2(Q, P) + \lambda_2 \mathrm{KL}(Q \| P)$$

Implements WGF/ID GD!

Step 1: Sample $\theta_n(0) \sim Q_0$, $n = 1,2,\ldots N$

Brownian motion s.t. $\{B_n(t)\}_{t \geq 0}$ stochastically independent

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) - \lambda_1 \nabla \mu_P(\theta_n(t)) - \lambda_2 \nabla \log p(\theta_n(t)) + \frac{\lambda_1}{N} \sum_{i=1}^{N} \nabla_1 \kappa(\theta_n(t), \theta_i(t)) \right) dt + \sqrt{2\lambda_2 dB_n(t)}$$

$\Longrightarrow$ Hope: $\frac{1}{N} \sum_{n=1}^{N} \theta_n(T) \approx Q^*$ for large T, N.

Special case: No regulariser

Base case: NO regularisation $(D = 0;\ \lambda_1 = \lambda_2 = 0)$

Special case: No regulariser

NO convexity!

$\Longrightarrow$ NO uniqueness!

Base case: NO regularisation $(D = 0;\ \lambda_1 = \lambda_2 = 0)$

$$Q^* \in \text{argmin}_{Q \in \mathscr{P}(\Theta)} \int \ell(\theta) dQ(\theta)$$

Special case: No regulariser

NO convexity!

$\Longrightarrow$ NO uniqueness!

Base case: NO regularisation $(D = 0;\ \lambda_1 = \lambda_2 = 0)$

$$Q^* \in \operatorname{argmin}_{Q \in \mathscr{P}(\Theta)} \int \ell(\theta) dQ(\theta)$$

Step 1: Sample $\theta_n(0) \sim Q_0$, $n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) - \lambda_1 \nabla \mu_P(\theta_n(t)) - \lambda_2 \nabla \log p(\theta_n(t)) + \frac{\lambda_1}{N} \sum_{i=1}^{N} \nabla_1 \kappa(\theta_n(t), \theta_i(t)) \right) dt + \sqrt{2\lambda_2 dB_n(t)}$$

Special case: No regulariser

NO convexity!

$\implies$ NO uniqueness!

Base case: NO regularisation $(D = 0;\ \lambda_1 = \lambda_2 = 0)$

$$Q^* \in \text{argmin}_{Q \in \mathcal{P}(\Theta)} \int \ell(\theta) dQ(\theta)$$

Step 1: Sample $\theta_n(0) \sim Q_0$, $n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) \cancel{- \lambda_1 \nabla \mu_P(\theta_n(t)) - \lambda_2 \nabla \log p(\theta_n(t))} \cancel{+ \frac{\lambda_1}{N} \sum_{i=1}^{N} \nabla_1 \kappa(\theta_n(t), \theta_i(t))} \right) dt \cancel{+ \sqrt{2\lambda_2} dB_n(t)}$$

NO prior influence

NO repulsion

NO noise injection

Special case: No regulariser

NO convexity!

$\Longrightarrow$ NO uniqueness!

Base case: NO regularisation $(D = 0;\ \lambda_1 = \lambda_2 = 0)$

$$Q^* \in \text{argmin}_{Q \in \mathcal{P}(\Theta)} \int \ell(\theta) dQ(\theta)$$

Step 1: Sample $\theta_n(0) \sim Q_0$, $n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) \cancel{- \lambda_1 \nabla \mu_P(\theta_n(t)) - \lambda_2 \nabla \log p(\theta_n(t))} \cancel{+ \frac{\lambda_1}{N} \sum_{i=1}^{N} \nabla_1 \kappa(\theta_n(t), \theta_i(t))} \right) dt + \cancel{\sqrt{2\lambda_2} dB_n(t)}$$

NO prior influence

NO repulsion

NO noise injection

$\Longrightarrow$ This implements a (deep) ensemble! ($\Longrightarrow$ DEs implement WGF!)

Special case: No regulariser

NO convexity!
$\implies$ NO uniqueness!

Base case: NO regularisation $(D = 0;\ \lambda_1 = \lambda_2 = 0)$

$$Q^* \in \text{argmin}_{Q \in \mathcal{P}(\Theta)} \int \ell(\theta) dQ(\theta)$$

Step 1: Sample $\theta_n(0) \sim Q_0, n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) - \cancel{\lambda_1 \nabla \mu_P(\theta_n(t)) - \lambda_2 \nabla \log p(\theta_n(t))} + \cancel{\frac{\lambda_1}{N} \sum_{i=1}^{N} \nabla_1 \kappa(\theta_n(t), \theta_i(t))} \right) dt + \cancel{\sqrt{2\lambda_2} dB_n(t)}$$

NO prior influence

NO repulsion

NO noise injection

$\implies$ This implements a (deep) ensemble! ($\implies$ DEs implement WGF!)

$\implies$ Question: $\frac{1}{N} \sum_{n=1}^{N} \theta_n(T) \approx Q^*$ for large T, N?

Special case: No regulariser

Base case: NO regularisation $(D = 0;\ \lambda_1 = \lambda_2 = 0)$ $\quad Q^* \in \text{argmin}_{Q \in \mathscr{P}(\Theta)} \int \ell(\theta) dQ(\theta)$

$$\implies \text{Question: } \frac{1}{N} \sum_{n=1}^{N} \theta_n(T) \approx Q^* \text{ for large } T, N?$$

Special case: No regulariser

Base case: NO regularisation $(D = 0;\ \lambda_1 = \lambda_2 = 0)$ $\quad Q^* \in \mathrm{argmin}_{Q \in \mathscr{P}(\Theta)} \int \ell(\theta) dQ(\theta)$

$\implies$ Question: $\frac{1}{N}\sum_{n=1}^{N} \theta_n(T) \approx Q^*$ for large T, N?

$Q^* \in \left\{ Q \in \mathscr{P}(\Theta) : Q(\Theta_{\min}) = 1, \text{ for } \Theta_{\min} = \mathrm{argmin}_{\theta \in \Theta} \ell(\theta) \right\}$

Special case: No regulariser

Base case: NO regularisation $(D = 0;\ \lambda_1 = \lambda_2 = 0)$ $\qquad Q^* \in \mathrm{argmin}_{Q \in \mathcal{P}(\Theta)} \int \ell(\theta) dQ(\theta)$

$\implies$ Question: $\frac{1}{N}\sum_{n=1}^{N} \theta_n(T) \approx Q^*$ for large T, N?

$Q^* \in \left\{ Q \in \mathcal{P}(\Theta) : Q(\Theta_{\min}) = 1, \text{ for } \Theta_{\min} = \mathrm{argmin}_{\theta \in \Theta} \ell(\theta) \right\}$

We show: $\xrightarrow{D} \sum_{i=1}^{\infty} Q_0(\Theta_i)\delta_{m_i} := Q_\infty$

Here, m_i is i-th local minimum of ℓ,
and Θ_i the region of attraction for m_i.

Special case: No regulariser

Base case: NO regularisation $(D = 0;\ \lambda_1 = \lambda_2 = 0)$ $\qquad Q^* \in \text{argmin}_{Q \in \mathscr{P}(\Theta)} \int \ell(\theta) dQ(\theta)$

!!!

$\implies$ Question: $\frac{1}{N}\sum_{n=1}^{N} \theta_n(T) \cancel{\approx} Q^*$ for large T, N?

$Q^* \in \{Q \in \mathscr{P}(\Theta) : Q(\Theta_{\min}) = 1$, for $\Theta_{\min} = \text{argmin}_{\theta \in \Theta} \ell(\theta)\}$

We show: $\xrightarrow{D} \sum_{i=1}^{\infty} Q_0(\Theta_i)\delta_{m_i} := Q_\infty$

Q_∞ NOT in this set!
(Unless all limiting local minimisers of ℓ are global)

Here, m_i is i-th local minimum of ℓ,
and Θ_i the region of attraction for m_i.

Special case: No regulariser

Base case: NO regularisation $(D=0;\ \lambda_1=\lambda_2=0)$ $\qquad Q^* \in \operatorname{argmin}_{Q\in\mathscr{P}(\Theta)} \int \ell(\theta)dQ(\theta)$

!!!

$\Longrightarrow$ Question: $\frac{1}{N}\sum_{n=1}^{N}\theta_n(T) \cancel{\approx} Q^*$ for large T, N?

$Q^* \in \left\{Q \in \mathscr{P}(\Theta) : Q(\Theta_{\min}) = 1, \text{ for } \Theta_{\min} = \operatorname{argmin}_{\theta\in\Theta}\ell(\theta)\right\}$

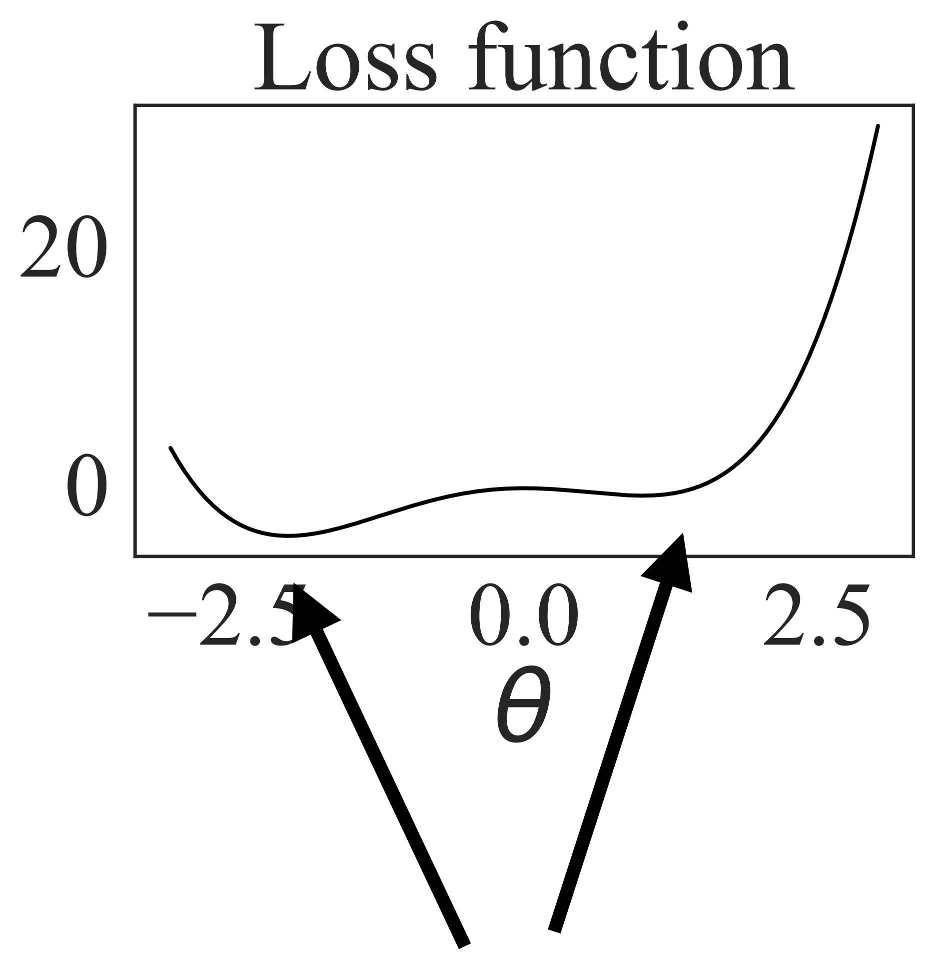
We show: $\xrightarrow{D} \sum_{i=1}^{\infty} Q_0(\Theta_i)\delta_{m_i} := Q_\infty$

Q_∞ NOT in this set!
(Unless all limiting local minimisers of ℓ are global)

Here, m_i is i-th local minimum of ℓ,
and Θ_i the region of attraction for m_i.

Assumes countable local minima;
same message for uncountably many

Special case: No regulariser

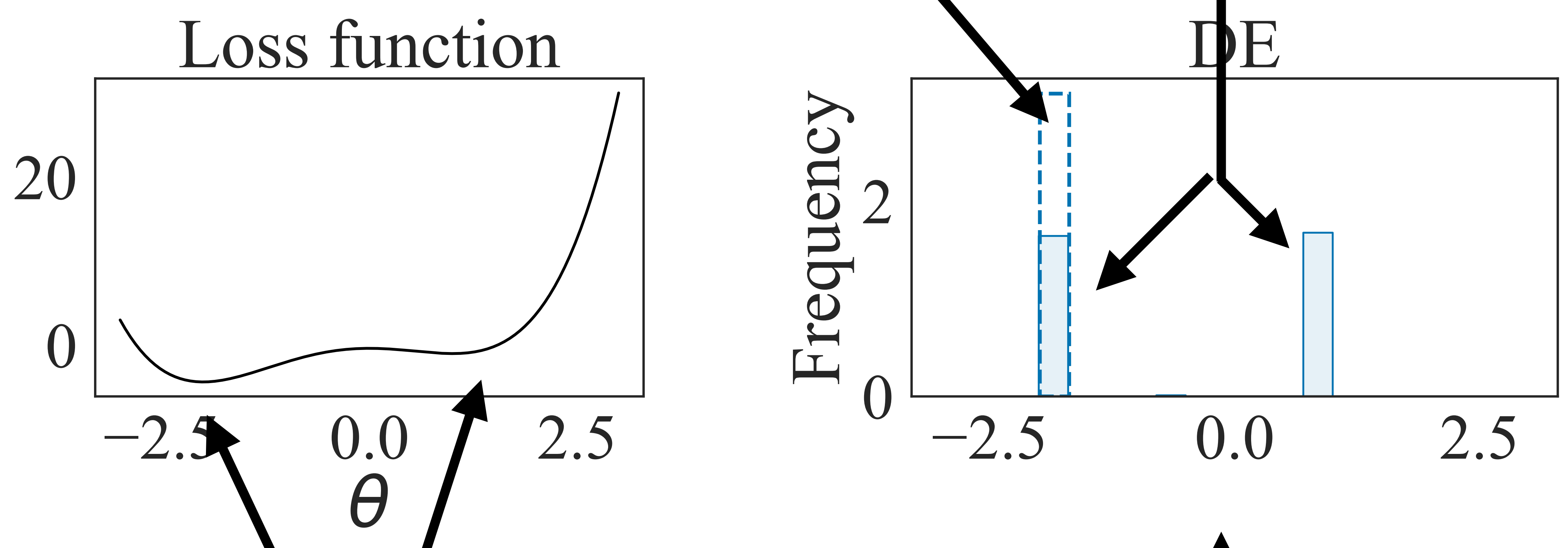


Two local minima

Special case: No regulariser

Solid histogram: Q_∞

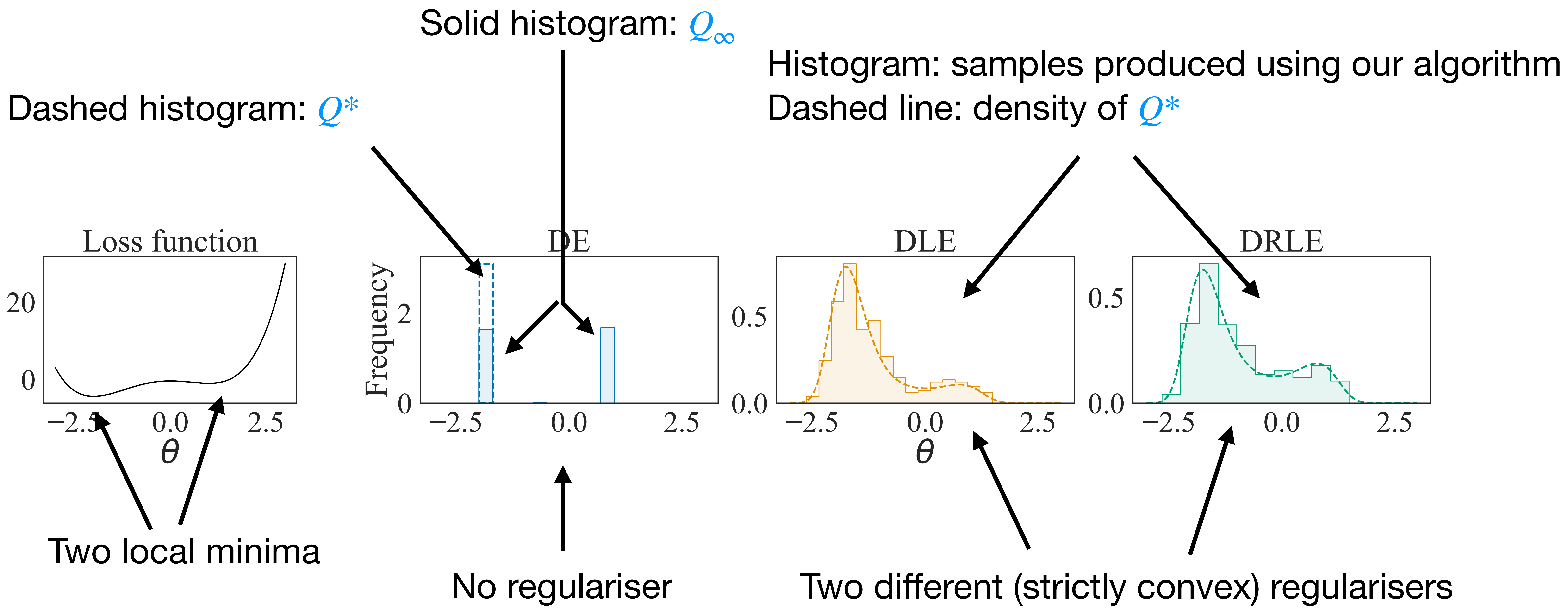
Dashed histogram: Q^*



Two local minima

No regulariser

Special case: No regulariser



Special case: No regulariser

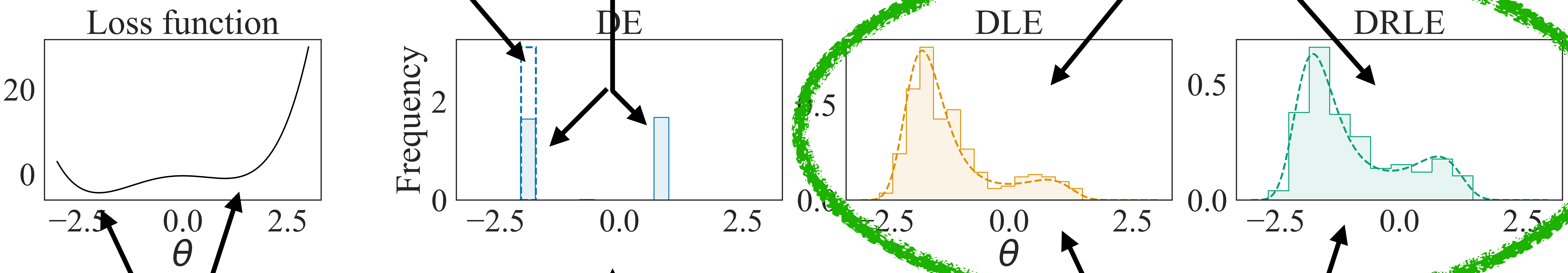
$$\frac{1}{N}\sum_{n=1}^{N}\theta_n(T) \approx Q^* \text{ for large } T, N?$$

Solid histogram: Q_∞

Histogram: samples produced using our algorithm

Dashed line: density of Q^*

Dashed histogram: Q^*



Two local minima

No regulariser

Two different (strictly convex) regularisers

Special case: Only KL-regulariser

Step 1: Sample $\theta_n(0) \sim Q_0, n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) - \lambda_1 \nabla \mu_P(\theta_n(t)) - \lambda_2 \nabla \log p(\theta_n(t)) + \frac{\lambda_1}{N}\sum_{i=1}^{N} \nabla_1 \kappa(\theta_n(t), \theta_i(t))\right) dt + \sqrt{2\lambda_2 dB_n(t)}$$

Special case: Only KL-regulariser

Step 1: Sample $\theta_n(0) \sim Q_0, n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) - \cancel{\lambda_1 \nabla \mu_P(\theta_n(t))} - \lambda_2 \nabla \log p(\theta_n(t)) + \cancel{\frac{\lambda_1}{N} \sum_{i=1}^{N} \nabla_1 k(\theta_n(t), \theta_i(t))} \right) dt + \sqrt{2\lambda_2 dB_n(t)}$$

NO influence of prior KME

NO repulsion

Special case: Only KL-regulariser

Step 1: Sample $\theta_n(0) \sim Q_0, n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) - \cancel{\lambda_1 \nabla \mu_P(\theta_n(t))} - \lambda_2 \nabla \log p(\theta_n(t)) + \cancel{\frac{\lambda_1}{N} \sum_{i=1}^{N} \nabla_1 \kappa(\theta_n(t), \theta_i(t))} \right) dt + \sqrt{2\lambda_2 dB_n(t)}$$

NO influence of prior KME

NO repulsion

$\Longrightarrow$ This implements sampling from a (generalised) Bayes posterior

Special case: Only KL-regulariser

Step 1: Sample $\theta_n(0) \sim Q_0, n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) - \cancel{\lambda_1 \nabla \mu_P(\theta_n(t))} - \lambda_2 \nabla \log p(\theta_n(t)) + \cancel{\frac{\lambda_1}{N} \sum_{i=1}^{N} \nabla_1 k(\theta_n(t), \theta_i(t))} \right) dt + \sqrt{2\lambda_2 dB_n(t)}$$

NO influence of prior KME

NO repulsion

$\implies$ This implements sampling from a (generalised) Bayes posterior

$\implies$ This is similar to unadjusted Langevin Sampling; can show $\frac{1}{N}\sum_{n=1}^{N} \theta_n(T) \xrightarrow{D} Q^*$ for large T, N

Special case: KL + MMD regularisers

Step 1: Sample $\theta_n(0) \sim Q_0, n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) - \lambda_1 \nabla \mu_P(\theta_n(t)) - \lambda_2 \nabla \log p(\theta_n(t)) + \frac{\lambda_1}{N} \sum_{i=1}^{N} \nabla_1 \kappa(\theta_n(t), \theta_i(t))\right) dt + \sqrt{2\lambda_2 dB_n(t)}$$

Special case: KL + MMD regularisers

Step 1: Sample $\theta_n(0) \sim Q_0, n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) - \lambda_1 \nabla \mu_P(\theta_n(t)) - \lambda_2 \nabla \log p(\theta_n(t)) + \frac{\lambda_1}{N} \sum_{i=1}^{N} \nabla_1 \kappa(\theta_n(t), \theta_i(t)) \right) dt + \sqrt{2\lambda_2 dB_n(t)}$$

$\implies$ This implements sampling from an (unknown!!!) optimal measure/generalised posterior

Special case: KL + MMD regularisers

Step 1: Sample $\theta_n(0) \sim Q_0, n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) - \lambda_1 \nabla \mu_P(\theta_n(t)) - \lambda_2 \nabla \log p(\theta_n(t)) + \frac{\lambda_1}{N}\sum_{i=1}^{N} \nabla_1 \kappa(\theta_n(t), \theta_i(t))\right) dt + \sqrt{2\lambda_2 dB_n(t)}$$

$\Longrightarrow$ This implements sampling from an (unknown!!!) optimal measure/generalised posterior

$\Longrightarrow$ We can show $\frac{1}{N}\sum_{n=1}^{N} \theta_n(T) \xrightarrow{D} Q^*$ for large T, N

Special case: KL + MMD regularisers

Step 1: Sample $\theta_n(0) \sim Q_0, n = 1,2,\ldots N$

Step 2: Evolve via SDE given as

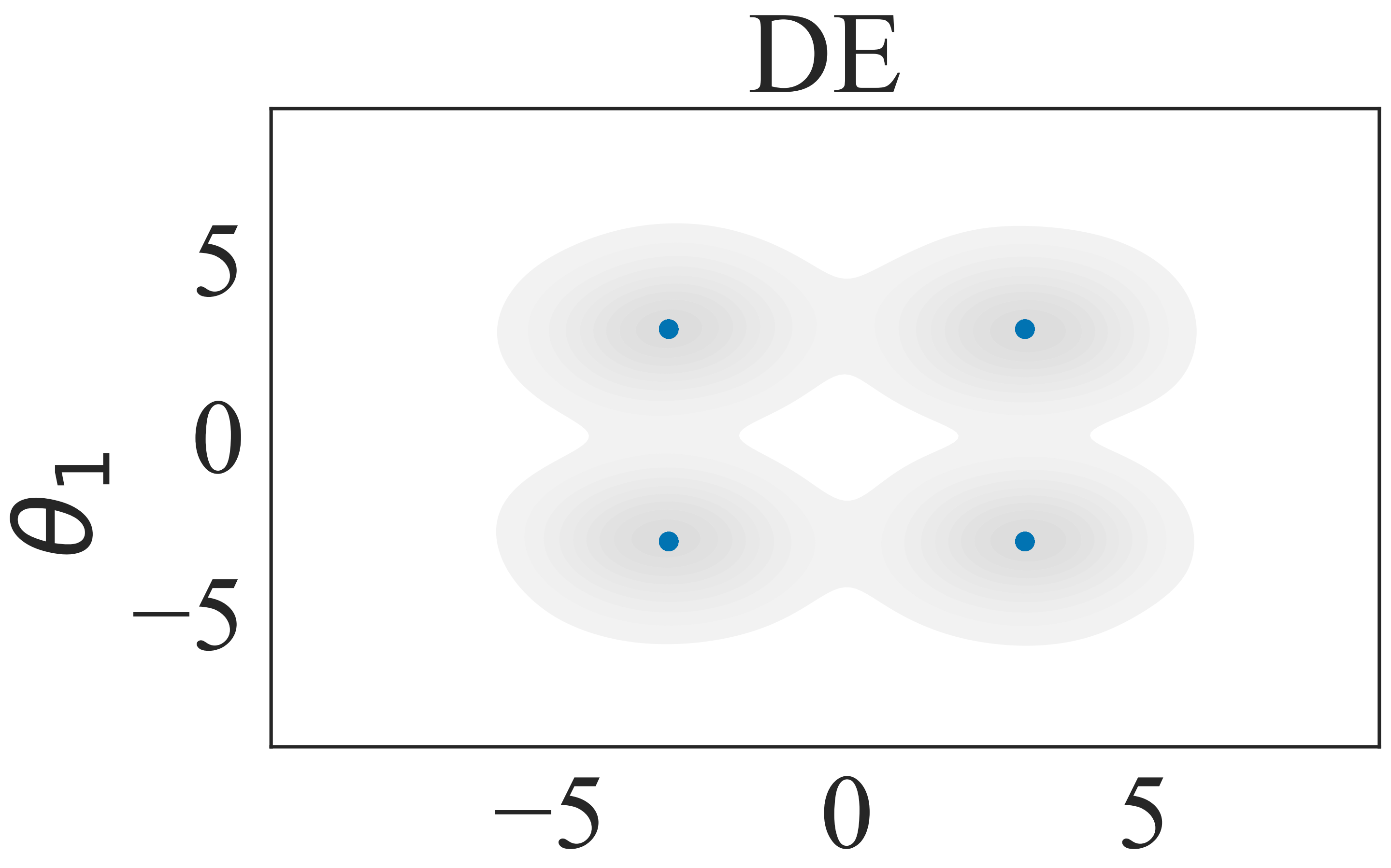
$$d\theta_n(t) = -\left(\nabla \ell(\theta_n(t)) - \lambda_1 \nabla \mu_P(\theta_n(t)) - \lambda_2 \nabla \log p(\theta_n(t)) + \frac{\lambda_1}{N}\sum_{i=1}^{N} \nabla_1 \kappa(\theta_n(t), \theta_i(t))\right) dt + \sqrt{2\lambda_2 dB_n(t)}$$

$\implies$ This implements sampling from an (unknown!!!) optimal measure/generalised posterior

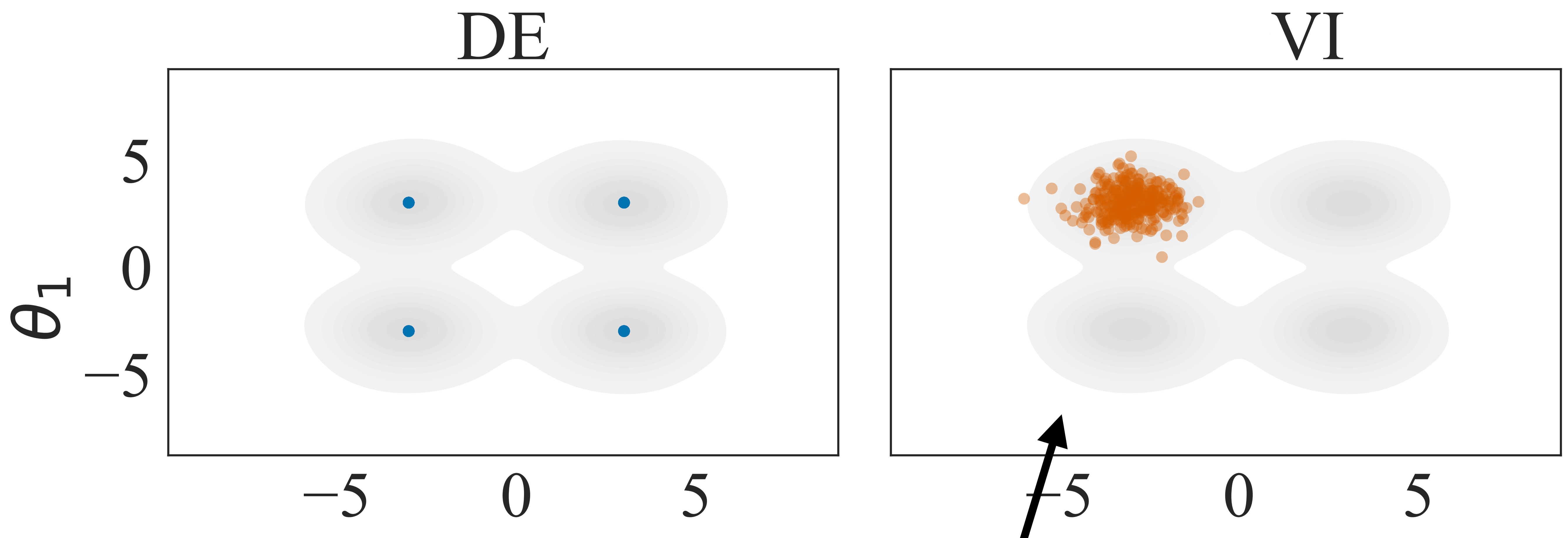
$\implies$ We can show $\frac{1}{N}\sum_{n=1}^{N} \theta_n(T) \xrightarrow{D} Q^*$ for large T, N

Without further conditions, only if $\lambda_2 > 0$ [i.e., KL used]!
(Technical problem: Q could be discrete)

Experiment 1: simple comparison



Experiment 1: simple comparison



$$Q^*_{\mathrm{VI}} = \mathrm{argmin}_{Q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim Q} \left[\ell(\theta) \right] + \lambda \mathrm{KL}(Q \| P) \right\}$$

Approximation via Variational Inference (VI)
(Variational family = $\mathcal{Q} \subset \mathcal{P}(\Theta)$)

Experiment 1: simple comparison

$$Q^*_{\mathrm{VI}} = \mathrm{argmin}_{Q \in \mathcal{Q}} \left\{ \mathbb{E}_{\theta \sim Q} \left[\ell(\theta) \right] + \lambda \mathrm{KL}(Q \| P) \right\}$$

Approximation via Variational Inference (VI)

(Variational family = $\mathcal{Q} \subset \mathscr{P}(\Theta)$)

Experiment 1: simple comparison

$\Longrightarrow$ Using ANY infinite-dimensional WGF procedure gives better results than VI

Experiment 2: ‘DEs are Bayesian inference’

We clarify that the recent deep ensembles (Lakshminarayanan et al., 2017) are not a competing approach to Bayesian inference, but can be viewed as a compelling mechanism for Bayesian marginalization. Indeed, we empirically demonstrate that deep ensembles can provide a better approximation to the Bayesian predictive distribution than standard Bayesian approaches.

A.G. Wilson, P. Izmailov. *Bayesian Deep Learning and a Probabilistic Perspective of Generalization*. Advances in Neural Information Processing Systems, 2020.
(cited > 400 times according to Google scholar)

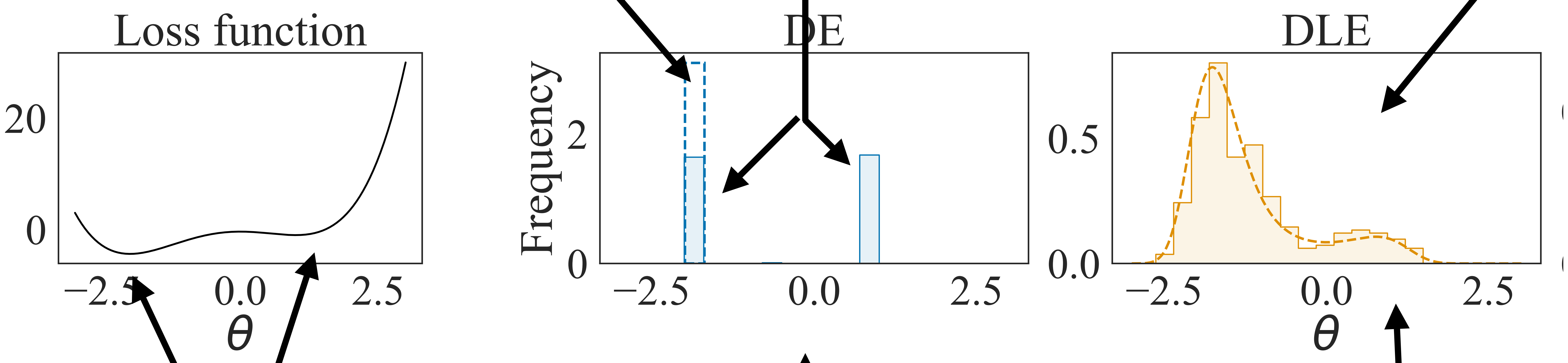
Experiment 2: ‘DEs are Bayesian inference’

Solid histogram: Q_∞

Histogram: samples produced using our algorithm

Dashed line: density of Q^*

Dashed histogram: Q^*



Two local minima

No regulariser / DE

KL -regulariser / (generalised) Bayes posterior

A.G. Wilson, P. Izmailov. *Bayesian Deep Learning and a Probabilistic Perspective of Generalization*. Advances in Neural Information Processing Systems, 2020.

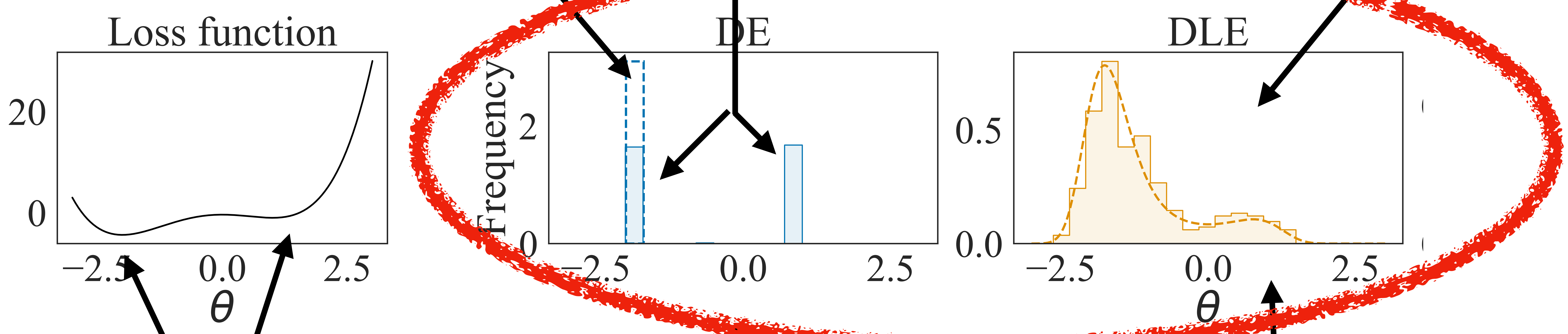
Experiment 2: ‘DEs are Bayesian inference’

Solid histogram: Q_∞

Dashed histogram: Q^*

Histogram: samples produced using our algorithm

Dashed line: density of Q^*



Two local minima

These two are not alike!

No regulariser / DE

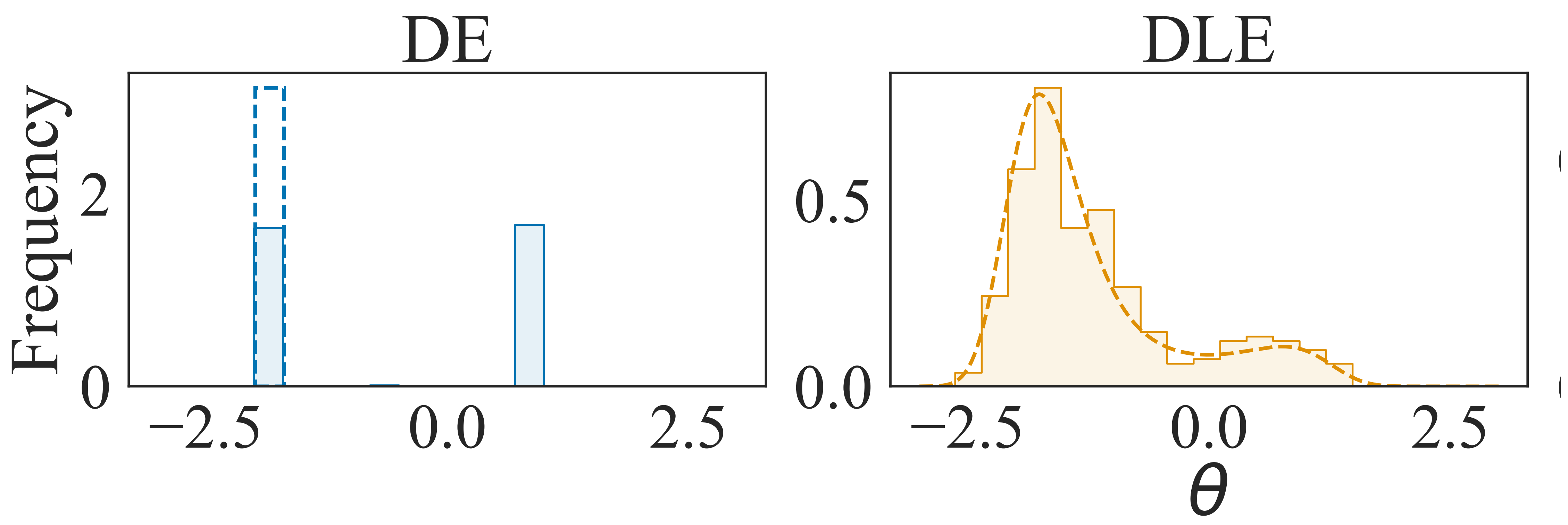
KL -regulariser / (generalised) Bayes posterior

A.G. Wilson, P. Izmailov. *Bayesian Deep Learning and a Probabilistic Perspective of Generalization*. Advances in Neural Information Processing Systems, 2020.

Experiment 2: ‘DEs are Bayesian inference’

What is going on?

Why do people claim that these distributions are the same?

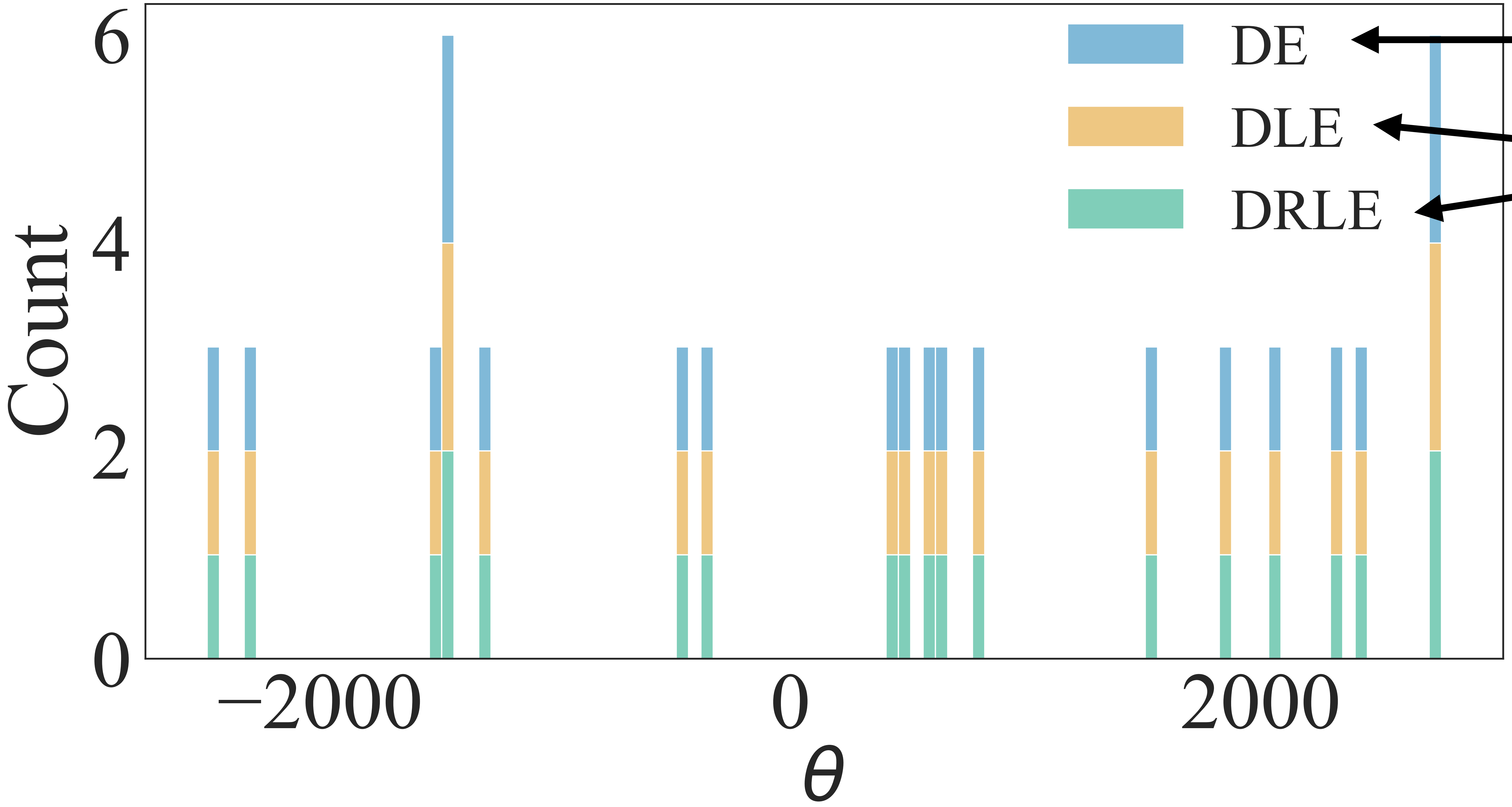


Experiment 2: ‘DEs are Bayesian inference’

$\ell(\theta) = -|\sin(\theta)|; \quad \theta \in [-1000\pi, 1000\pi]$ (2000 local minima)

Experiment 2: ‘DEs are Bayesian inference’

$\ell(\theta) = -|\sin(\theta)|; \quad \theta \in [-1000\pi, 1000\pi]$ (2000 local minima)



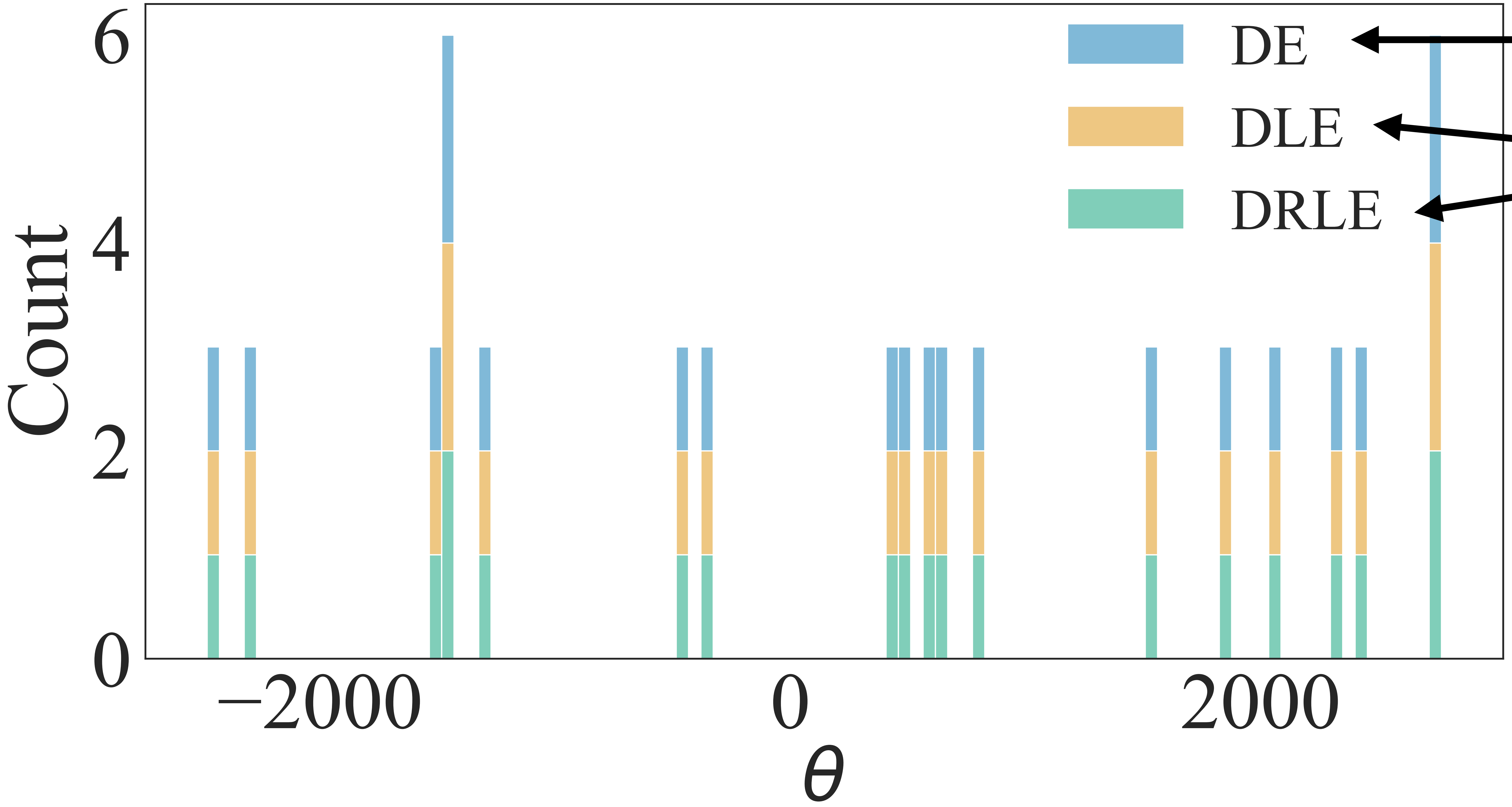
Without regulariser

With regulariser

Location of $\theta_n(T)$ for $n = 1,2,\ldots,N$ with same initialisations

Experiment 2: ‘DEs are Bayesian inference’

$$\ell(\theta) = -|\sin(\theta)|; \quad \theta \in [-1000\pi, 1000\pi]$$ (2000 local minima)



Without regulariser

With regulariser

Location of $\theta_n(T)$ for $n = 1,2,\ldots,N$ with same initialisations

$\Longrightarrow$ The confusion comes from small/finite N, T (relative to number of minima)!

Summary / Conclusion:

$$\min_{\theta\in\Theta} \ell(\theta) \qquad \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \int \ell(\theta)dQ(\theta) \qquad \min_{Q\in\mathcal{P}(\mathbb{R}^J)} \left\{ \int \ell(\theta)dQ(\theta) + \lambda D(Q,P) \right\}$$

Step 1: probabilistic lifting **Step 2:** convexification through regularisation

1. Non-convex, finite-dimensional (FD) => convex, infinite-dimensional (ID)
2. Build ID gradient descent (GD) algorithm!
   (tells us about interplay of Bayes & Deep ensembles)
3. Practically useful? => Yes for quite small NNs & with sufficient computational budget, no for larger ones

Work available as preprint:

https://arxiv.org/abs/2305.15027