Gradient Flows and Marginal Maximum Likelihood Estimation

From EM and ULA to PGD and back again

Adam M. Johansen (University of Warwick)

Joint work with Rocco Caprio*, Jen Ning Lim, Juan Kuntz* and Sam Power

* From whom I have stolen and mangled many slides.

Newcastle—October 2nd, 2026

a.m.johansen@warwick.ac.uk

http://go.warwick.ac.uk/amjohansen/talks/





Overview

Gradient flows and related analytic techniques are useful:

- Developing novel algorithms
- Understanding existing ones

Outline:

- Latent variables models and marginal inference
- Gradient flows
- The PGD Algorithm
- Functional analysis and non-asymptotic control
- Adding momentum
- Conclusions

Latent Variable Models and Marginal Inference

Inference in Latent Variable Models

Problem: Given some (fixed) data y, and probabilistic model $p_\theta(x, y)$ relating latent variables x and data y, parametrized by θ: infer θ and characterise $p_\theta(x|y)$.

Assumption: x and θ take values in $\mathbb{R}^{d_x}$ and $\mathbb{R}^{d_\theta}$, respectively.

A Simple Example
Often, data is assumed to be generated as

$$y_m = f_\theta(x_m) + \eta_m, \quad x_m \sim p_\theta^x(\mathrm{d}x), \quad \forall m = 1, \ldots, M;$$

where $\{\eta_m\}$ are i.i.d. Gaussian noises. In this case,

$$p_\theta(x, y) = \prod_{m=1}^{M} p^\eta(y_m - f_\theta(x_m)) p_\theta^x(x_m).$$

ML Example: Generator networks for unsupervised learning

Non-linear extensions of factor analysis.

Assume that each point in a dataset is generated by:

(A) sampling latent variables from an isotropic Gaussian prior,
(B) mapping them through a neural network,
(C) and adding Gaussian noise.

We consider their use for image datasets (MNIST and CelebA), where

- observed variables y: 1024 pixels per image,
- latent variables x: 64 per image,
- parameters θ: the network's parameters (dimension $\approx 350,000$).

Marginal Maximum Likelihood / Empirical Bayes / Type-II ML

Problem: Given some (fixed) data y, and probabilistic model $p_\theta(x, y)$ relating latent variables x and data y, parametrized by θ: infer θ and characterise $p_\theta(x|y)$.

Approach: Empirical Bayes (Robbins, 1956): search for a $\theta_\star$ that explains the data well, and use that to infer x:

1. Find $\theta_\star$ maximizing the marginal likelihood, $\int p_\theta(x, y)\mathrm{d}x =: p_\theta(y)$:

$$\theta_\star := \arg\max_\theta \int p_\theta(x, y)\mathrm{d}x.$$

2. Use $\theta_\star$ to obtain

$$p_{\theta_\star}(x|y) = \frac{p_{\theta_\star}(x, y)}{p_{\theta_\star}(y)}$$

and perform inference on x via this 'optimal posterior'.

The EM algorithm (Dempster et al., 1977)

The canonical approach to this problem:

Given some (θ_0, q_0), iterate

- M-step

$$\theta_{k+1} = \arg\max_{\theta} \int \log(p_\theta(x, y)) q_k(\mathrm{d}x)$$

- E-step

$$q_{k+1} = p_{\theta_{k+1}}(\cdot|y)$$

Under general conditions,

$$\theta_k \to \theta_* \quad \text{and} \quad q_k \to p_{\theta_*}(\cdot|y) \quad \text{as} \quad k \to \infty,$$

where θ_* is a stationary point of $\theta \mapsto p_\theta(y)$ (i.e. $\nabla_\theta p_{\theta_*}(y) = 0$).

The EM algorithm as optimization

Proposition (Neal and Hinton (1998))

EM is a **coordinate-wise minimization** *procedure on* $\mathcal{M} := \mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$*:*

$$\theta_{k+1} \in \operatorname*{argmin}_{\theta \in \mathbb{R}^{d_\theta}} \mathcal{F}(\theta, q_k),$$

$$q_{k+1} \in \operatorname*{argmin}_{q \in \mathcal{P}(\mathbb{R}^{d_x})} \mathcal{F}(\theta_{k+1}, q),$$

where $\mathcal{F} : \mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x}) \to \mathbb{R}$ is the *free energy* functional:

$$\mathcal{F}(\theta, q) := \int \log\left(\frac{q(x)}{p_\theta(x, y)}\right) q(\mathrm{d}x)$$

and

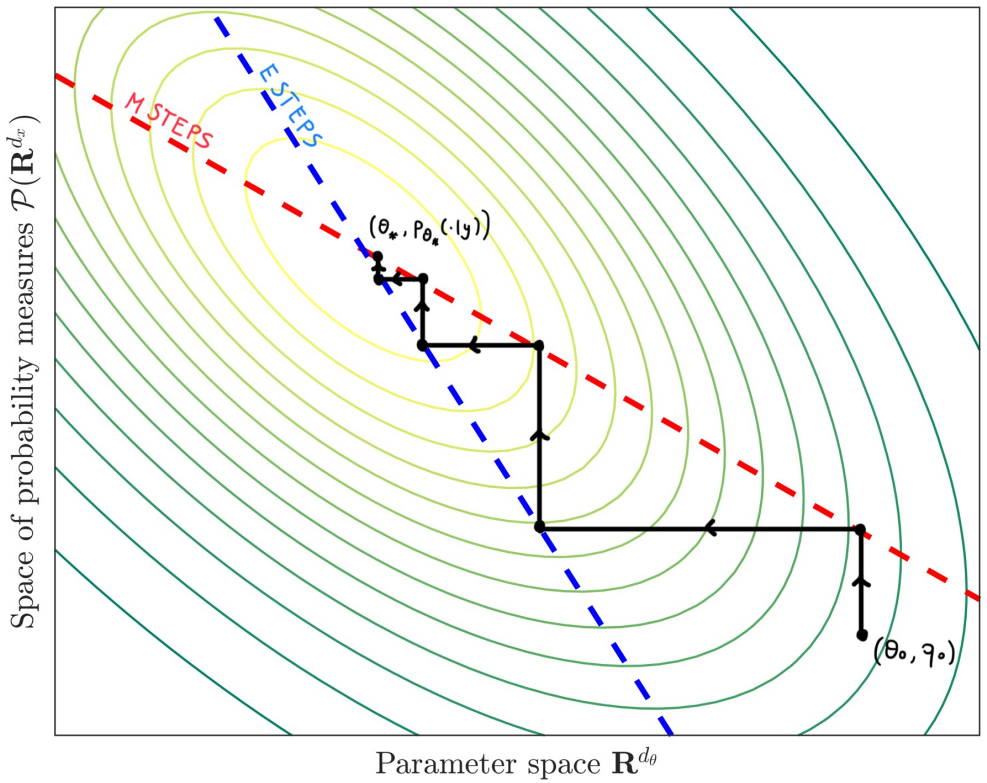
$$\mathcal{M}_\star := \operatorname{argmin} F = \{(\theta_\star, p_{\theta_\star}(\cdot|y)) : \theta_\star \in \arg\max p_\theta(y)\}.$$

The EM algorithm as optimization

$$\theta_{k+1} = \arg\max_{\theta} \int \log(p_{\theta}(x, y)) q_k(\mathrm{d}x) = \operatorname*{argmin}_{\theta \in \mathbb{R}^{d_\theta}} \mathcal{F}(\theta, q_k) \quad \text{(M)}$$

$$q_{k+1} = p_{\theta_{k+1}}(\cdot|y) = \operatorname*{argmin}_{q \in \mathcal{P}(\mathbb{R}^{d_X})} \mathcal{F}(\theta_{k+1}, q) \quad \text{(E)}$$

Hence, EM does something like



The EM algorithm as optimization

EM is generally intractable: what about other optimization approaches for $\mathcal{F}(\theta, q)$?

	$\theta \in \mathbb{R}^{d_\theta}$	$q \in \mathcal{P}(\mathbb{R}^{d_x})$
EM	min	min
Gradient EM	gradient step	min
Langevin EM	min	gradient step
Coordinate-wise Gradient Descent	gradient step	gradient step
Gradient Descent	gradient step	

To make sense of this table, we want to understand what it means to take gradient steps of $\mathcal{F}$ with respect to $(\theta, q) \in \mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$.

Gradient Flows and Particle Gradient Descent

Gradient Flows

Euclidean Given $f : \mathbb{R}^d \to \mathbb{R}$, the ODE

$$dx_t = -\nabla f(x_t)dt, x_0 \in \mathbb{R}^d$$

admits time discretisation

$$x_{t+h} = x_t - h\nabla f(x_t).$$

General Given $F : \mathcal{X} \to \mathbb{R}$, we envisage

$$dX_t = -\mathrm{grad}_{\mathcal{X}} F(X_t)dt, X_0 \in \mathcal{X}$$

and some time discretisation thereof.
Requires an interpretation of $\mathrm{grad}_{\mathcal{X}}$; easy on Riemannian manifolds and more generally via "Otto calculus" (Otto, 2001; Villani, 2009).

Geometry of $\mathcal{P}(\mathbb{R}^{d_x})$

When endowed with the Wasserstein-2 metric d_{W_2}, $\mathcal{P}(\mathbb{R}^{d_x})$ can be thought of as a Riemannian manifold (Otto, 2001; Villani, 2009). Then, for some reasonable functions $\mathcal{G}$ on $\mathcal{P}(\mathbb{R}^{d_x})$ one can make sense of things like

- Gradient of $\mathcal{G}$ in $\mathcal{P}(\mathbb{R}^{d_x})$: $\mathrm{grad}\mathcal{G}(q) = \nabla_x \cdot (q \nabla_x \delta_{\mathcal{P}_2} \mathcal{G}(q))$
- The gradient flow of probability distributions $\partial_t q_t = -\mathrm{grad}\mathcal{G}(q_t)$: $\partial_t q_t = \nabla_x \cdot (q_t \nabla_x \delta_{\mathcal{P}_2} \mathcal{G}(q_t)), \quad q_0 = q.$

where $\delta_{\mathcal{P}_2} \mathcal{G}(q)$ is the functional derivative (first variation) of $\mathcal{G}$.

- If $\mathcal{G}(q) = \mathcal{H}(q) = \int q \log q$, then $\partial_t q_t = \Delta_x q_t$, the heat flow.
- If $\mathcal{G}(q) = \mathrm{KL}(q||\pi) = \int q \log(q/\pi)$, $\pi \propto e^{-V}$, then $\partial_t q_t = \nabla_x \cdot (\nabla_x q_t + q_t \nabla_x V)$, the Langevin diffusion's law:

$$dX_t = -\nabla_x V(X_t) dt + \sqrt{2} dW_t \qquad X_t \sim q_t.$$

Geometry of $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$

When endowed with the product-metric

$$\mathrm{d}((\theta, q), (\theta', q')) = \sqrt{\mathrm{d}_E(\theta, \theta')^2 + \mathrm{d}_{W_2}(q, q')^2}$$

$(\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$ can be thought of as a Riemannian manifold (Kuntz et al., 2023; Caprio et al., 2025)

This allow us to make sense of concepts such as

- Gradient of the free energy $\mathcal{F}$ in $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$:

  $$\mathrm{grad}\mathcal{F}(\theta, q) = \left(-\int \nabla_\theta \log(p_\theta(x, y)) q(dx), -\nabla_x \cdot \{q(\nabla_x \log(q) - \nabla_x \log(p_\theta(\cdot, y))\} \right).$$

- The gradient flow of the free energy $\partial_t(\theta_t, q_t) = -\mathrm{grad}\mathcal{F}(\theta_t, q_t)$:

  $$\partial_t \theta_t = \int \nabla_\theta \log(p_{\theta_t}(x, y)) q_t(dx), \quad \partial_t q_t = \nabla_x \cdot \{q_t(\nabla_x \log(q_t) - \nabla_x \log((p_{\theta_t}(\cdot, y))\}$$

  i.e.,

  $$\partial_t \theta_t = \int \nabla_\theta \log(p_{\theta_t}(x, y)) \mathrm{Law}(X_t)(dx), \quad dX_t = \nabla_x \log(p_{\theta_t}(X_t, y)) dt + \sqrt{2} dW_t$$

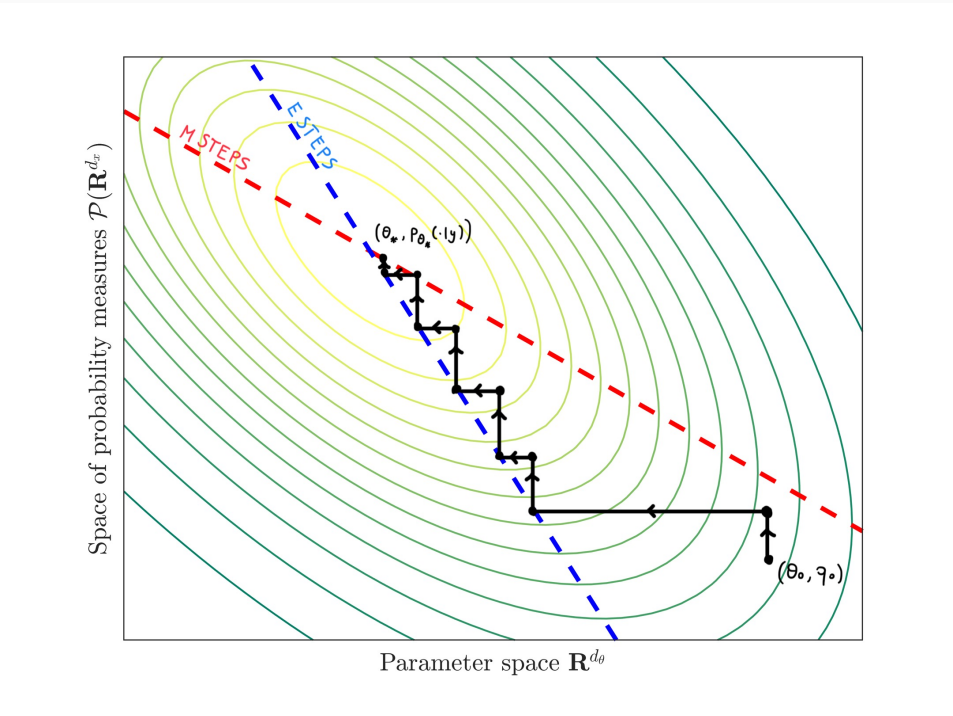
  i.e. the *steepest descent curve connecting an arbitrary* (θ_0, q_0) *to* $(\theta_\star, p_{\theta_\star}(\cdot|y))$.

The EM algorithm as optimization: gradient EM

	$\theta \in \mathbb{R}^{d_\theta}$	$q \in \mathcal{P}(\mathbb{R}^{d_x})$
Gradient EM	gradient step	min

$$\theta_{k+1} = \theta_k + h \int \nabla_\theta \log(p_{\theta_k}(x, y)) q_k(\mathrm{d}x) \qquad (\tilde{M})$$

$$q_{k+1} = p_{\theta_{k+1}}(\cdot|y) \qquad (\mathrm{E})$$

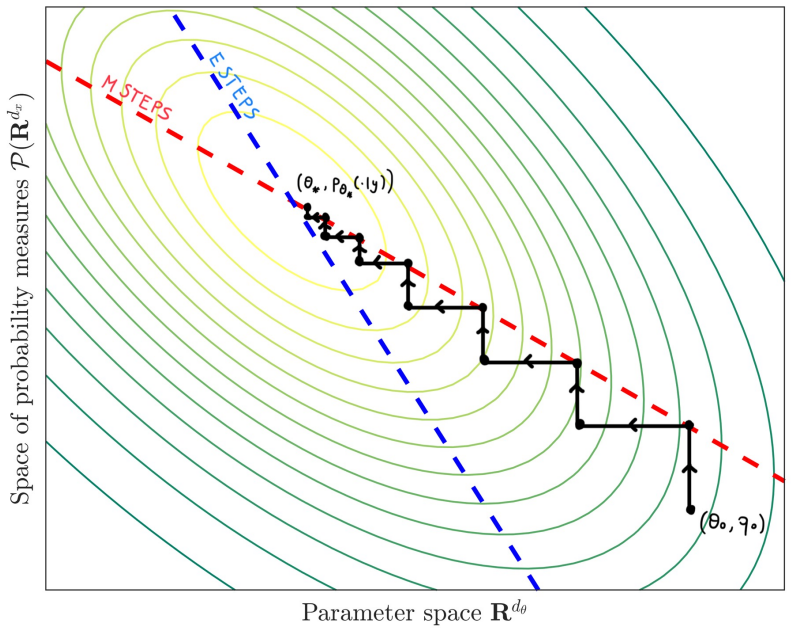


The EM algorithm as optimization: Langevin EM

	$\theta \in \mathbb{R}^{d_\theta}$	$q \in \mathcal{P}(\mathbb{R}^{d_x})$
Langevin EM	min	gradient step

$$\theta_{k+1} = \arg\max_{\theta} \int \log(p_{\theta_k}(x, y)) q_k(\mathrm{d}x) \qquad (\mathrm{M})$$

$$q_{k+1} = \mathrm{Law}(X_k + h\nabla_x \log(p_{\theta_{k+1}}(X_k|y)) + \sqrt{2h}\mathcal{N}_k), \quad \mathrm{Law}(X_k) = q_k \qquad (\tilde{E})$$

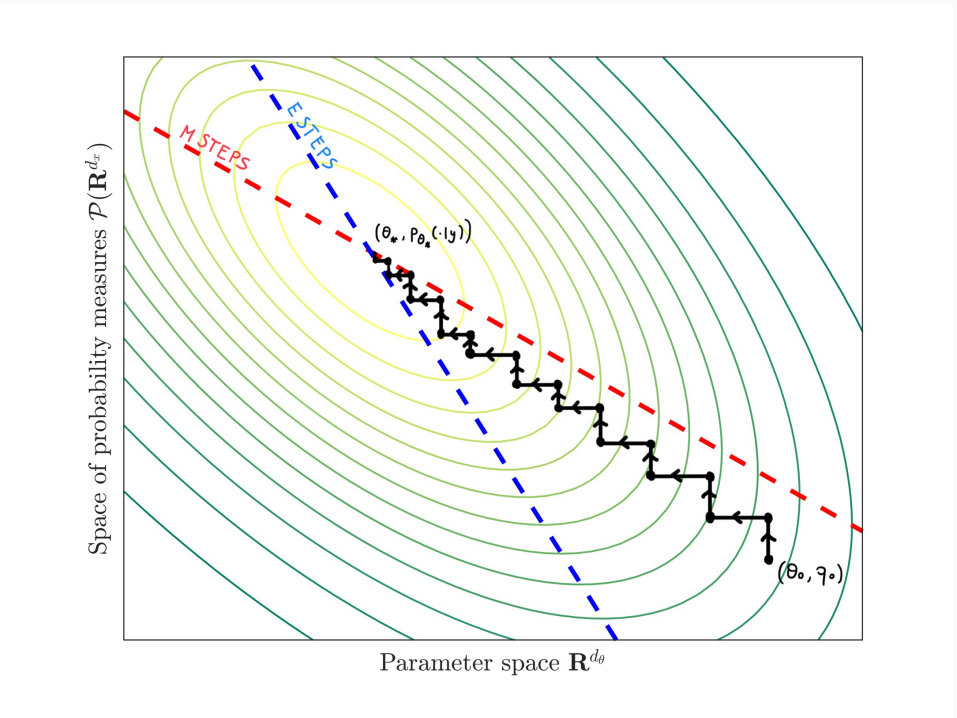


The EM algorithm as optimization: CW Gradient Descent

	$\theta \in \mathbb{R}^{d_\theta}$	$q \in \mathcal{P}(\mathbb{R}^{d_x})$
Coordinate-wise Gradient Descent	gradient step	gradient step

$$\theta_{k+1} = \theta_k + h \int \nabla_\theta \log(p_{\theta_k}(x, y)) q_k(\mathrm{d}x) \qquad (\tilde{M})$$

$$q_{k+1} = \mathrm{Law}(X_k + h\nabla_x \log(p_{\theta_{k+1}}(X_k|y)) + \sqrt{2h}\mathcal{N}_k), \quad \mathrm{Law}(X_k) = q_k \qquad (\tilde{E})$$



A mean field limit of the SOUL algorithm (De Bortoli et al. (2021)).

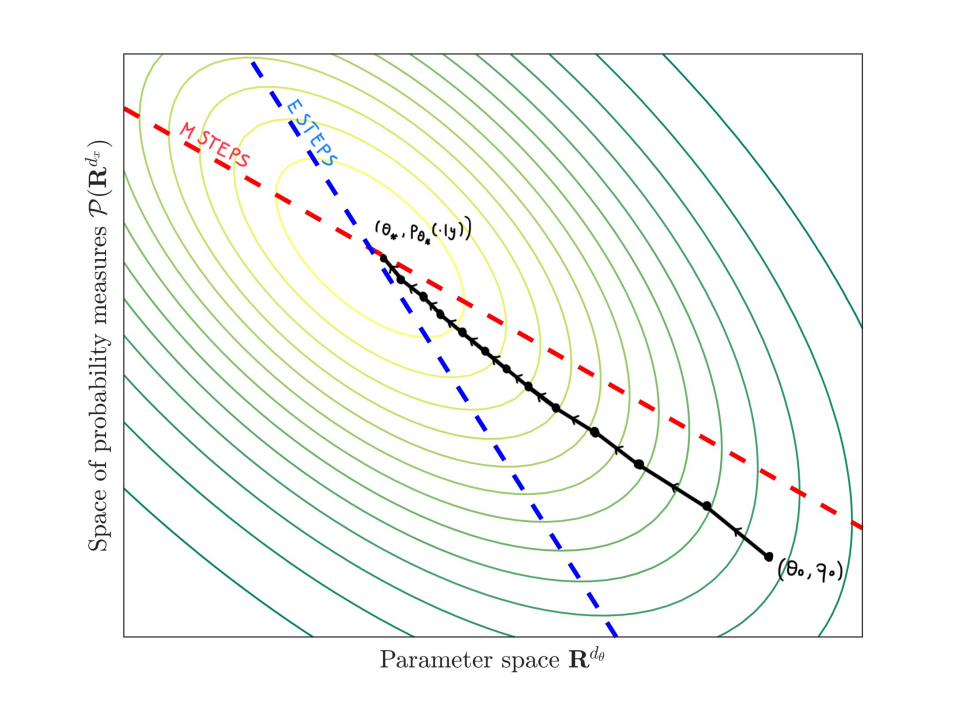
The EM algorithm as optimization: Gradient Descent

	$\theta \in \mathbb{R}^{d_\theta}$	$q \in \mathcal{P}(\mathbb{R}^{d_x})$
Gradient Descent	gradient step	

$$\theta_{k+1} = \theta_k + h \int \nabla_\theta \log(p_{\theta_k}(x, y)) q_k(\mathrm{d}x)$$

$$q_{k+1} = \mathrm{Law}(X_k + h\nabla_x \log(p_{\theta_k}(X_k|y)) + \sqrt{2h}\mathcal{N}_k), \quad \mathrm{Law}(X_k) = q_k$$

($\tilde{M}$-$\tilde{E}$)



A m.f. limit of PGD (Kuntz et al. (2023)), IPLA (Akyildìz et al. (2025)).

Three intractabilities

Gradient flow:

$$\dot{\theta}_t = \int \nabla_\theta \ell(\theta_t, x) q_t(x) dx, \quad \dot{q}_t = -\nabla_x \cdot \left[q_t \nabla_x \log \left(\frac{p_{\theta_t}(\cdot, y)}{q_t} \right) \right]. \tag{1}$$

Three intractabilities:

(A) The continuous time axis.
(B) The integral over $\mathcal{X}$.
(C) The PDE with domain $\mathcal{X}$.

Solution: (1) is the a McKean-Vlasov Fokker-Planck equation satisfied by the law of the following McKean SDE:

$$d\theta_t = \left[\int \nabla_\theta \ell(\theta_t, x) q_t(x) dx \right] dt, \quad dX_t = \nabla_x \ell(\theta_t, X_t) dt + \sqrt{2} dW_t,$$

where q_t denotes X_t's law and W a standard Brownian motion.

Two intractabilities

McKean SDE:

$$d\theta_t = \left[\int \nabla_\theta \ell(\theta_t, x) q_t(x) dx\right] dt, \quad dX_t = \nabla_x \ell(\theta_t, X_t) dt + \sqrt{2} dW_t,$$

Two intractabilities:

(1) The continuous time axis.

(2) The integral over $\mathcal{X}$.

Solution: Generate $N > 0$ particles $X_t^1, \ldots, X_t^N$ with law q_t by solving

$$dX_t^n = \nabla_x \ell(\theta_t, X_t^n) dt + \sqrt{2} dW_t^n \quad \forall n \in [N] := \{1, \ldots, N\},$$

with $W^1, \ldots, W^N$ denoting N independent Brownian motions, and use

$$q_t \approx \frac{1}{N} \sum_{n=1}^N \delta_{X_t^n} \quad \Rightarrow \quad \int \nabla_\theta \ell(\theta_t, x) q_t(x) dx \approx \frac{1}{N} \sum_{n=1}^N \nabla_\theta \ell(\theta_t, X_t^n).$$

One intractability

SDE:

$$d\Theta_t = \left[\frac{1}{N}\sum_{n=1}^{N} \nabla_\theta \ell(\Theta_t, X_t^n)\right] dt,$$
$$dX_t^n = \nabla_x \ell(\Theta_t, X_t^n)dt + \sqrt{2}dW_t^n \ \ \forall n \in [N].$$

One intractability: The continuous time axis.

Solution: Discretize using Euler-Maruyama with step size $h > 0$,

$$\Theta_{k+1} = \theta_k + \frac{h}{N}\sum_{n=1}^{N} \nabla_\theta \ell(\Theta_k, X_k^n), \quad \forall k \in [K],$$
$$X_{k+1}^n = X_k^n + h\nabla_x \ell(\Theta_k, X_k^n) + \sqrt{2h}W_k^n \ \ \forall n \in [N], \ \ k \in [K],$$

where $(W_k^n)_{k\in[K-1],n\in[N]}$ denote independent standard normal r.v.s.

Particle gradient descent (Kuntz et al., 2023)

Discretizing the gradient flow, we obtain particle gradient descent:

(1) Choose the step size $h > 0$ and particle number $N > 0$:

$$\Theta_{k+1} = \Theta_k + \frac{h}{N}\sum_{n=1}^{N}\nabla_\theta \ell(\Theta_k, X_k^n), \quad \forall k \in [K] := \{1, \dots, K\},$$

$$X_{k+1}^n = X_k^n + h\nabla_x \ell(\theta_k, X_k^n) + \sqrt{2h}W_k^n, \quad \forall n \in [N], \;\; k \in [K],$$

where $(W_k^n)_{k\in[K-1],n\in[N]}$ denote iid standard normal r.v.s., so

$$\Theta_k \approx \theta_{kh}, \qquad \frac{1}{N}\sum_{n=1}^{N}\delta_{X_k^n} \approx q_{kh}, \qquad \forall k > 0.$$

(2) Estimate a stationary point θ_* of the marginal likelihood $\theta \mapsto p_\theta(y)$ and its corresponding posterior $p_{\theta_*}(\cdot|y)$ using

$$\theta_* \approx \Theta_K, \qquad p_{\theta_*}(\cdot|y) \approx \frac{1}{N}\sum_{n=1}^{N}\delta_{X_K^n}.$$

Particle gradient descent

Particle gradient descent:

- avoids accept-reject steps;
- only requires evaluating gradients of $\ell(\theta, x) = \log(p_\theta(x, y))$;
- its cost is $\mathcal{O}(N[\text{eval. cost of } \nabla\ell])$;
- computations can be vectorized across N;
- in big-data settings, we replace $\nabla\ell$ with estimates thereof;
- we improve performance by adapting step sizes.

PGD trains large latent variable models without resorting to VI.

Example: Generator networks

We

- fit the model using PGD,
- $10,000$ (MNIST) and $40,000$ (CelebA) training images,
- and a Google Colab subscription.

Image synthesis

MNIST



CelebA



Figure 1: Synthesized images obtained using the generator trained with PGD.

Image reconstruction

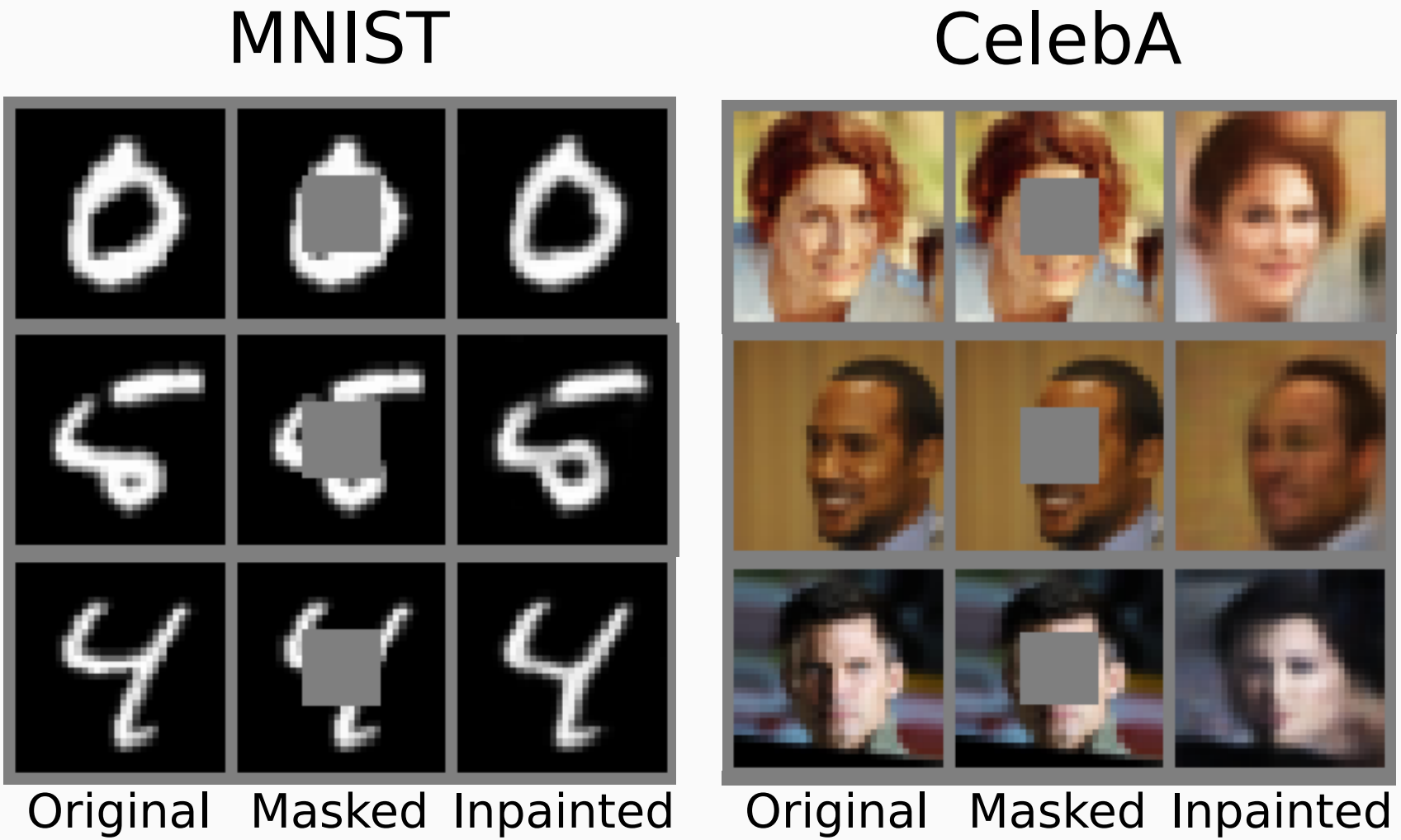


Figure 2: Images reconstructed using the generator trained with PGD.

A Little Functional Analysis

Error bounds via functional inequalities

1. We want to study non-asymptotics for the above algorithms.
2. They are, in a way or another, different methods to minimise $\mathcal{F}$, and their convergence will depend on properties of $\mathcal{F}$.
3. Furthermore, this suggests we should first study their convergence in terms of free energy $\mathcal{F}$.
4. For instance, for EM
   - M-step
   $$\theta_{k+1} = \arg\max_{\theta} \int \log(p_{\theta_k}(x, y)) q_k(\mathrm{d}x)$$
   - E-step
   $$q_{k+1} = p_{\theta_{k+1}}(\cdot|y)$$

   this suggests we should study the *joint* convergence of both θ_k and q_k in terms of free energy, $\mathcal{F}(\theta_k, q_k)$.

Error bounds via functional inequalities (Caprio et al., 2025)

We introduce a functional inequality that relates $\mathcal{F}$ with its *gradient*. Let

$$\mathcal{I}(\theta, q) := \left|\left| \int \nabla_\theta \log(p_\theta(x,y)) q(x) dx \right|\right|^2 + \int \left|\left| \nabla_x \log\left(\frac{q(x)}{p_\theta(x,y)}\right) \right|\right|^2 q(x) dx.$$

The functional $\mathcal{I}$ can be interpreted as $||\text{grad}\mathcal{F}||^2$, where $\text{grad}\mathcal{F}$ is the gradient of the free energy in the geometry induced by the Euclidean and Wasserstein metrics on $\mathbb{R}^{d_\theta} \times \mathcal{P}_2(\mathbb{R}^{d_x})$.

Definition (Extended log-Sobolev inequality; xLSI)

The statistical model $(p_\theta(x,y))_{\theta \in \mathbb{R}^{d_\theta}}$ satisfies the extended log-Sobolev inequality with constant $\lambda > 0$ if

$$2\lambda[\mathcal{F}(\theta, q) - \inf \mathcal{F}] \leq \mathcal{I}(\theta, q) \qquad \text{(xLSI)}$$

holds for all $(\theta, q) \in \mathcal{M}_2$.

Error bounds via functional inequalities

A few remarks on the xLSI:

1. If, in a sense, $\mathbb{R}^{d_\theta} = \{\theta\}$, this is the log-Sobolev inequality for $p_\theta(\cdot|y)$:

$$2\lambda \mathrm{KL}(q||p_\theta(\cdot|y)) \leq \int \left|\left|\nabla_x \log\left(\frac{q(x)}{p_\theta(x|y)}\right)\right|\right|^2 q(x)dx;$$

2. If, in a sense, $\mathcal{P}_2(\mathbb{R}^{d_x}) = \{p_\theta(\cdot|y)\}$, this is the Polyak–Łojasiewicz inequality for $f(\theta) := -\log(p_\theta(y))$:

$$2\lambda[f(\theta) - \inf f] \leq ||\mathrm{grad} f(\theta)||^2;$$

3. The xLSI is a property of a *statistical model* / parametrized family of distributions.

Error bounds via functional inequalities

Just as the extended LSI generalises the standard log-Sobolev inequality, we have generalisations of various other key results and inequalities:

- An 'extended' Talagrand (1996) inequality:

$$\lambda \mathrm{d}((\theta, q), \mathcal{M}_\star)^2 \leq 2(\mathcal{F}(\theta, q) - \mathcal{F}_\star), \quad \mathcal{M}_\star := \operatorname{argmin} \mathcal{F} \quad (2)$$

- An 'extended' Otto and Villani (2000) Theorem:

$$\text{xLSI} \Rightarrow \text{the extended Talagrand inequality.}$$

Error bounds via functional inequalities: nice to have

- An 'extended' HWI inequality (Otto and Villani, 2000): if $\nabla^2 \log(p_\theta(x,y)) \preceq -\iota I_{d_\theta+d_x}$ for some $\iota \in \mathbb{R}$,

  $$\mathcal{F}(\theta,q) \leq \mathcal{F}(\theta',q') + \mathrm{d}((\theta,q),(\theta',q'))\sqrt{\mathcal{I}(\theta,q)} - \frac{\iota}{2}\mathrm{d}((\theta,q),(\theta',q'))$$

- An extended Bakry and Émery (1985) Theorem: if $\nabla^2 \log(p_\theta(x,y)) \preceq -\iota I_{d_\theta+d_x}$ with $\iota > 0$, the xLSI holds with $\lambda = \iota > 0$.

- Preservation results: if the model $(p_\theta(x,y))_{\theta\in\Theta}$ satisfy the xLSI with constant λ, and "$\tilde{p}_\theta(x,y)$ is some perturbation of $p_\theta(x,y)$" then also $(\tilde{p}_\theta(x,y))_{\theta\in\Theta}$ satisfy the extended log-Sobolev inequality with some explicit constant $\tilde{\lambda}$ (generalisation of things like Holley and Stroock (1987)'s lemma ...)

Error bounds via functional inequalities

Two assumptions we will consider (separately, or together).

We say the model $(p_\theta(x, y))_{\theta \in \mathbb{R}^{d_\theta}}$ is

1. L_θ-smooth if
   $||\nabla_\theta \log(p_\theta(x, y)) - \nabla_\theta \log(p_{\theta'}(x, y))|| \leq L_\theta \, ||\theta - \theta'||$;

2. L_x-smooth if
   $||\nabla_x \log(p_\theta(x, y)) - \nabla_x \log(p_\theta(x', y))|| \leq L_x \, ||x - x'||$;

and set $L := \max(L_\theta, L_x)$.

Error bounds for exact algorithms

Theorem ((Caprio and Johansen, 2025))
If statistical model $(p_\theta(x, y))_{\theta \in \mathbb{R}^{d_\theta}}$ satisfy the xLSI + smoothness:

- *For gradient descent:*

$$\lambda \mathrm{d}((\theta_k, q_k), \mathcal{M}_\star)^2 \leq 2e^{-\lambda \sum_{j=1}^k h_j}(\mathcal{F}(\theta_0, q_0) - \mathcal{F}_\star) + 12L^2 d_x \sum_{j=1}^k h_j^2 e^{-\lambda \sum_{i=j+1}^k h_i}.$$

- *For alternating minimisation (EM):*

$$\lambda \mathrm{d}((\theta_k, q_k), \mathcal{M}_\star)^2 \leq 2 \inf \left\{ e^{-\lambda(k/L_\theta + \sum_{j=1}^k h_j)}(\mathcal{F}(\theta_0, q_0) - \mathcal{F}_\star) + B \sum_{j=1}^k h_j^2 e^{-\lambda(k/L_\theta + \sum_{i=j+1}^k h_i)} \right\}$$

with all $h_j \leq L_x/4$.

Error bound for PGD

Theorem (Caprio et al. (2025))

Under Assumptions of differentiability, smoothness and strong log concavity, if $X_0^{1,h}, \dots, X_0^{N,h}$ *are drawn independently from a distribution* q_0 *in* $\mathcal{P}_2(\mathbb{R}^{d_x})$ *and* $h \leq 1/(\lambda + L)$*, then*

$$\mathrm{d}((\Theta_K^{N,h}, Q_K^{N,h}), (\theta_*, Q_*^N)) \tag{3}$$

$$\leq \sqrt{h} A_{0,h} + \frac{L\sqrt{2}}{\lambda\sqrt{N}} \sqrt{B_0 + \frac{2d_x}{\lambda}} + \mathrm{d}((\theta_0, q_0), (\theta_*, \pi_{\theta_*})) e^{-h\lambda K} \tag{4}$$

for all K *in* $\mathbb{N}$*; where* $B_0 := ||\theta_0||^2 + \mathbb{E}\left[||X_0||^2\right]$ *and*

$$A_{0,h} := \sqrt{\frac{4h + 4/\iota}{\iota} 220 L^2 \left(L^2 h \left[B_0 + \frac{2d_x}{\lambda} \right] + d_x \right)} \quad \textit{with} \quad \iota := \frac{2L\lambda}{L + \lambda}.$$

Can be improved for conditionally iid models.

One Extension: Adding Momentum and MPGD

Adding momentum in the Euclidean setting

If GD for a function f is analogous to solving the ordinary differential equation (ODE)

$$\dot{\theta}_t = -\nabla_\theta f\left(\theta_t\right) \tag{5}$$

with $\nabla_\theta f$ denoting f's Euclidean gradient, then a 'momentum' method is akin to solving a second-order ODE like

$$\dot{\theta}_t = \eta m_t, \quad \dot{m}_t = -\gamma\eta m_t - \nabla_\theta f\left(\theta_t\right), \tag{6}$$

where γ, η denote positive hyperparameters.

And in the latent variable setting

Enriching both components of the space with momentum variables:

- replace θ and x in $\mathbb{R}^{d_\theta}$ and $\mathbb{R}^{d_x}$
- with (θ, m) and (x, u) in $\mathbb{R}^{2d_\theta}$ and $\mathbb{R}^{2d_x}$;
- and substitute $\mathcal{P}(\mathbb{R}^{d_x})$ with $\mathcal{P}(\mathbb{R}^{2d_x})$.

Define a Hamiltonian on the enriched space $\mathbb{R}^{2d_\theta} \times \mathcal{P}(\mathbb{R}^{2d_x})$:

$$\mathcal{F}(\theta, m, q) := \int \log\left(\frac{q(x,u)}{\rho_{\theta,\eta_x}(x,u)}\right) q(x,u)\,\mathrm{d}x\,\mathrm{d}u + \frac{\eta_\theta}{2}\|m\|^2.$$

where (η_θ, η_x) denote positive hyperparameters and $\rho_{\theta,\eta_x}(x,u) := p_\theta(y,x)\, r_{\eta_x}(u)$ with $r_{\eta_x} := \mathcal{N}(0, \eta_x^{-1} I_{d_x})$.

A latent variable Hamiltonian flow

The $(\gamma_\theta, \gamma_x)$-damped Hamiltonian flow of $\mathcal{F}$:

$$(\dot{\theta}_t, \dot{m}_t) = -\mathrm{D}_{\gamma_\theta} \nabla_{(\theta,m)} \mathcal{F}(\theta_t, m_t, q_t), \tag{7}$$

$$\dot{q}_t = \nabla_{(x,u)} \cdot \left(q_t \mathrm{D}_{\gamma_x} \nabla_{(x,u)} \delta_q \mathcal{F}[\theta_t, m_t, q_t] \right). \tag{8}$$

is the Fokker–Planck equation of a McKean–Vlasov SDE:

$$\mathrm{d}\theta_t = \eta_\theta m_t \, \mathrm{d}t, \tag{9}$$

$$\mathrm{d}m_t = \left[\int \nabla_\theta \ell(\theta_t, x) \, q_{t,X}(\mathrm{d}x) - \gamma_\theta \eta_\theta m_t \right] \mathrm{d}t, \tag{10}$$

$$\mathrm{d}X_t = \eta_x U_t \, \mathrm{d}t, \tag{11}$$

$$\mathrm{d}U_t = [\nabla_x \ell(\theta_t, X_t) - \gamma_x \eta_x U_t] \, \mathrm{d}t + \sqrt{2\gamma_x} \, \mathrm{d}W_t, \tag{12}$$

where $q_{t,X} := \mathrm{Law}(X_t)$.

Dynamics of the Momentum-PGD particle system

For $i \in [M]$, we have, for $\mathcal{E}(\theta, q) := \int \log\left(\frac{q(x)}{p_\theta(y,x)}\right) q(x)\,\mathrm{d}x$:

$$\begin{aligned}
\tilde{\theta}_k =& \tilde{\theta}_{k-1} + \frac{\iota_\theta(h)}{\gamma_\theta}\tilde{m}_{k-1} - \frac{1}{\gamma_\theta}\left(h - \frac{\iota_\theta(h)}{\gamma_\theta \eta_\theta}\right)\nabla_\theta \mathcal{E}\left(\bar{\theta}_{k-1}, \tilde{q}^M_{k-1,X}\right)\\
\tilde{m}_k =& (1-\iota_\theta(h))\tilde{m}_{k-1} - \frac{\iota_\theta(h)}{\gamma_\theta\eta_\theta}\nabla_\theta \mathcal{E}\left(\bar{\theta}_{k-1}, \tilde{q}^M_{k-1,X}\right)\\
\tilde{X}^i_k =& \tilde{X}^i_{k-1} + \frac{\iota_x(h)}{\gamma_x}\tilde{U}^i_{k-1} + \frac{1}{\gamma_x}\left(h - \frac{\iota_x(h)}{\gamma_x\eta_x}\right)\nabla_x \ell\left(\bar{\bar{\theta}}_{k-1}, \tilde{X}^i_{k-1}\right) + L^{XX}_\Sigma \xi^i_k\\
\tilde{U}^i_k =& (1-\iota_x(h))\tilde{U}^i_{k-1} + \frac{\iota_x(h)}{\gamma_x\eta_x}\nabla_x \ell\left(\bar{\bar{\theta}}_{k-1}, \tilde{X}^i_{k-1}\right) + L^{XU}_\Sigma \xi^i_k + L^{UU}_\Sigma \xi'^{,i}_k,
\end{aligned}$$

where $\iota_x(h) := 1 - \exp(-\eta_x\gamma_x h)$, $\left\{\xi^i_k, \xi'^{,i}_k\right\}_{i\in[M]} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0_{d_x}, I_{d_x})$, and $\left(L^{XX}_\Sigma, L^{XU}_\Sigma, L^{UU}_\Sigma\right)$ are scalars.

We set

$$\bar{\theta}_{k-1} := \tilde{\theta}_{k-1} + \frac{\iota_\theta(h)}{\gamma_\theta}\tilde{m}_{k-1}, \quad \bar{\bar{\theta}}_{k-1} = \tilde{\theta}_k. \tag{13}$$

Image synthesis revisited

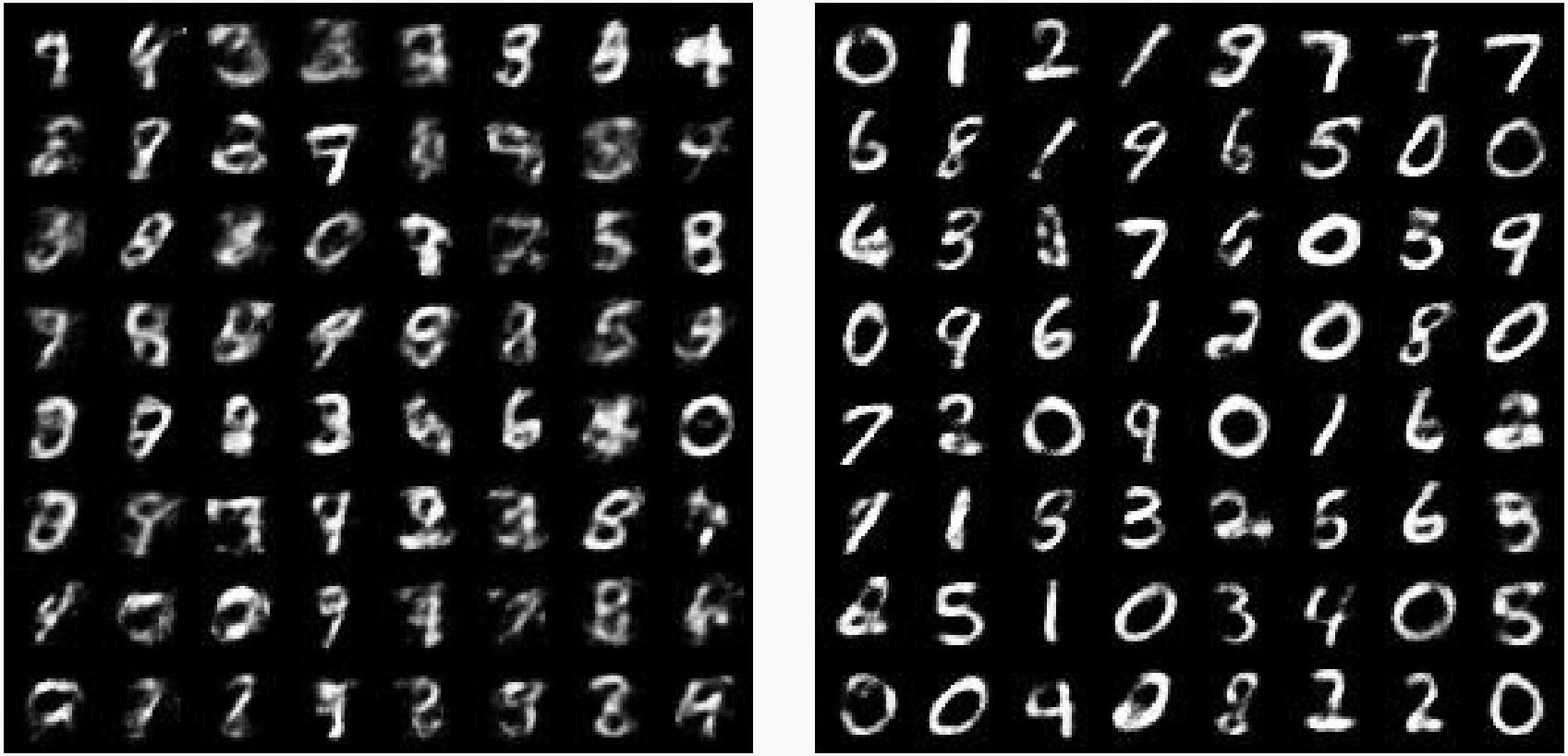


Figure 3: Synthesized images obtained using the generator trained with PGD (left) and MPGD (right).

Conclusions

Conclusions

Specifics

- Gradient flows generalised naturally to hybrid spaces.
- They provide scalable approximations of empirical Bayes.
- Most extensions also generalise.
- Ongoing work includes: methodological application to state space models (with Crucinio and Lin); extension to post Bayesian inference (with Caprio and Chadwick).

Generalities

- One can take inspiration from optimisation (and sampling) to solve the maximum marginal likelihood estimation problem, and to study related algorithms theoretically;
- Connections between EM (and its friends) with functional inequalities run deep.
- Functional analysis; gradient flows and related topics still have more to tell us about computational statistics.

Some key references

- Otto calculus and later developments: Villani (2009); Santambrogio (2015)

- PGD and the heuristic framework: Kuntz et al. (2023)
- A formal analysis of PGD: Caprio et al. (2025)
- Analysis of EM using gradient methods: Caprio and Johansen (2025)
- Adding momentum to PGD: Lim et al. (2024)

- A pragmatic alternative: Akyildìz et al. (2025) and several follow-up papers

References i

Ö. Deniz Akyildìz, Francesca R. Crucinio, Mark Girolami, Tim Johnston, and Sotirios Sabanis. Interacting particle Langevin algorithm for maximum marginal likelihood estimation. *ESAIM: Probability and Statistics*, 29:243 – 280, 2025.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84*. Lecture Notes in Mathematics Springer, 1985.

R. Caprio and A. M. Johansen. Fast convergence of the Expectation Maximization algorithm under a logarithmic Sobolev inequality. *Biometrika*, 112(4):asaf061, 2025. doi: 10.1093/biomet/asaf061.

Rocco Caprio, Juan Kuntz, Samuel Power, and Adam M Johansen. Error bounds for particle gradient descent, and extensions of the log-Sobolev and Talagrand inequalities. *Journal of Machine Learning Research*, 26 (103):1–38, 2025. URL `https://jmlr.org/papers/v26/24-0437.html`.

Valentin De Bortoli, Alain Durmus, Marcelo Pereyra, and Ana F Vidal. Efficient stochastic optimisation by unadjusted Langevin Monte Carlo: Application to maximum marginal likelihood and empirical Bayesian estimation. *Statistics and Computing*, 31:1–18, 2021.

Arthur P. Dempster, Nan M. Laird, and Donald B. Rubin. Maximum likelihood from incomplete data via the EM Algorithm. *Journal of the Royal Statistical Society, Series B*, 39:2–38, 1977.

Richard Holley and Daniel W Stroock. Logarithmic sobolev inequalities and stochastic ising models. *Journal of Statistical Physics*, 46:1159–1194, 1987.

Juan Kuntz, Jen Ning Lim, and Adam M. Johansen. Particle algorithms for maximum likelihood training of latent variable models. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206, pages 5134–5180, 2023.

References ii

J. N. Lim, J. Kuntz, S. Power, and A. M. Johansen. Momentum particle maximum likelihood. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp, editors, *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 29816–29871. PMLR, 21–27 Jul 2024. URL `https://proceedings.mlr.press/v235/lim24b.html`.

Radford M. Neal and Geoffrey E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Springer Netherlands, 1998.

Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. *Communications in Partial Differential Equations*, 26:101–174, 2001.

Felix Otto and Cédric Villani. Generalization of an inequality by Talagrand and links with the logarithmic Sobolev inequality. *Journal of Functional Analysis*, 173:361–400, 2000.

Herbert Robbins. An empirical Bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 3.1, pages 157–164, 1956.

Filippo Santambrogio. *Optimal Transport for Applied Mathematicians: Calculus of Variations, PDEs, and Modeling*. Birkhäuser/Springer, 2015.

Michel Talagrand. Transportation cost for Gaussian and other product measures. *Geometric & Functional Analysis*, 6:587–600, 1996.

Cédric Villani. *Optimal Transport: Old and New*. Springer Science & Business Media, 2009.

Thank you for listening!



https://go.warwick.ac.uk/amjohansen/talks