Particle methods for latent variable models

Francesca R. Crucinio, ESOMAS, University of Turin & Collegio Carlo Alberto

Joint work with: Deniz Akyildiz, Nicolas Chopin, Paula Cordero Encinar, Mark Girolami, Tim Johnston, Anna Korba, Sotirios Sabanis.

Latent Variable Models & Expectation Maximisation

Latent Variable Models (LVM)

Consider the following data-generating process

$$x \sim p_\theta(\cdot)$$
$$y \sim p_\theta(\cdot|x)$$

for some parameter $\theta \in^{d_\theta}$, where $x \in \mathcal{X}$ is a latent variable which cannot be observed.

Empirical Bayes: Given a data point y we want to find $\theta^\star$ maximising the marginal log-likelihood

$$\log p_\theta(y) = \log \int p_\theta(x) p_\theta(y|x) dx = \log \int p_\theta(x, y) dx.$$

Latent Variable Models: Examples

Gaussian Mixture Model

$$p_\theta(y) = \alpha \mathcal{N}(y; \theta_1, 1) + (1 - \alpha)\mathcal{N}(y; \theta_2, 1)$$

- $x \sim p_\theta$ describes the allocation to mixture components
- $y \sim p_\theta(\cdot|x)$ is sampled given the selected component

Energy Based Model

$$p_\theta(x, y) = \frac{1}{Z(\alpha)} e^{-U_\alpha(x)} \mathcal{N}(y; g_\beta(x), \sigma^2 \mathrm{Id})$$

- g_β is a generator function parametrised by β
- $\frac{1}{Z(\alpha)} e^{-U_\alpha(x)}$ is an EB prior

Expectation Maximisation (EM)

E-step w.r.t. *latent variables* x: compute for fixed θ

$$Q(\theta|\theta^{(n)}) = \int \log p_\theta(x,y) p_{\theta^{(n)}}(x|y) dx,$$

with

$$p_{\theta^{(n)}}(x|y) = \frac{p_{\theta^{(n)}}(x,y)}{p_{\theta^{(n)}}(y)} = \frac{p_{\theta^{(n)}}(y|x) p_{\theta^{(n)}}(x)}{p_{\theta^{(n)}}(y)}$$

M-step w.r.t. *parameters* θ: maximise $Q(\cdot|\theta^{(n)})$

An Optimisation Point of View

An Optimisation Point of View

Neal and Hinton (1998) show that

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

is equivalent to

$$\begin{aligned}(\theta^\star, p_{\theta^\star}(x|y)) &= \arg\min_{(\theta,\mu)} \mathrm{KL}(\mu|p_\theta(\cdot, y)) \\ &= \arg\min_{(\theta,\mu)} \left[\int \log\left(\mu(x)\right)\mu(x)dx - \int \log p_\theta(x, y)\mu(x)dx\right]\end{aligned}$$

An Optimisation Point of View: One line proof

$$\begin{aligned}\mathrm{KL}(\mu|p_\theta(\cdot,y)) &= \int \mu(x)\log\mu(x)dx - \int \mu(x)\log p_\theta(x,y)dx + \log p_\theta(y) - \log p_\theta(y)\\ &= \int \mu(x)\log\mu(x)dx - \int \mu(x)\log\frac{p_\theta(x,y)}{p_\theta(y)}dx - \log p_\theta(y)\\ &= \mathrm{KL}(\mu|p_\theta(\cdot|y)) - \log p_\theta(y)\end{aligned}$$

- the first term is 0 if $\mu = p_\theta(\cdot|y)$
- this leaves us with

$$\min_\theta \mathrm{KL}(\mu|p_\theta(\cdot,y)) = \min_\theta\left(-\log p_\theta(y)\right) = \max_\theta \log p_\theta(y)$$

An Optimisation Point of View

- **Aim 1:** sample from the posterior $p_{\theta^\star}(x|y)$
- **Aim 2:** estimate the MLE $\theta^\star$

1 Latent Variable Models & Expectation Maximisation

2 An Optimisation Point of View

- Optimising θ
- Optimising the posterior
- Joining the dots...

3 Experiments

Gradient Descent

Consider the **optimisation problem**

$$\min_{z \in \mathbb{R}^d} \mathcal{F}(z)$$

The gradient descent ODE in **Euclidean space** is

$$\dot{x}_t = -\nabla \mathcal{F}(x_t).$$

An Euler discretisation of the above gives the standard gradient descent algorithm

$$x_{n+1} = x_n - \gamma \nabla \mathcal{F}(x_n).$$

Gradient Descent

Consider the **optimisation problem**

$$\min_{z \in \mathbb{R}^d} \mathcal{F}(z)$$

The gradient descent ODE in **Euclidean space** is

$$\dot{x}_t = -\nabla \mathcal{F}(x_t).$$

An Euler discretisation of the above gives the standard gradient descent algorithm

$$x_{n+1} = x_n - \gamma \nabla \mathcal{F}(x_n).$$

Applying this to the Empirical Bayes problem

$$\theta_{n+1} = \theta_n - \gamma \nabla_\theta \, \mathrm{KL}(\mu | p_\theta(\cdot, y))_{|\theta=\theta_n} = \theta_n + \gamma \int \nabla_\theta \log p_\theta(x, y)_{|\theta=\theta_n} \mu(x) dx$$

1 Latent Variable Models & Expectation Maximisation

2 An Optimisation Point of View

- Optimising θ
- Optimising the posterior
- Joining the dots...

3 Experiments

Optimisation over Distributions

Assume θ is fixed. We need to solve

$$\min_{\mu \in \mathcal{P}(\mathcal{X})} \mathrm{KL}(\mu | p_\theta(\cdot|y)).$$

Gradient descent in this space is given by the following gradient flow

$$\dot{\mu}_t = -\nabla_{\mathcal{M}} \mathrm{KL}(\mu_t | p_\theta(\cdot|y))$$

where $\mathcal{M}$ denotes the metric w.r.t. which the gradient is taken.

Choice of metric[1]

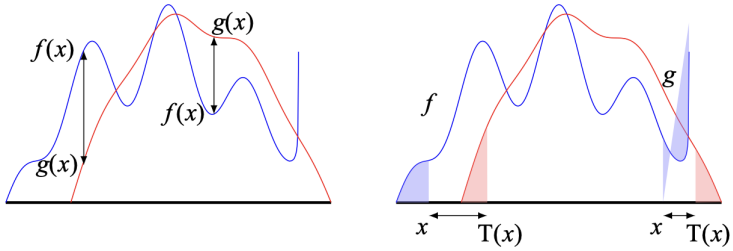


Figure 1: "Vertical" vs "horizontal" distances (the transport T is computed in the picture on the right using 1D considerations, imposing equality between the blue and red areas under the graphs of f and g).

[1] Santambrogio (2017)

Wasserstein Gradient Flow

If the metric is the Wasserstein-2 distance we obtain the **Wasserstein gradient flow PDE** (Jordan et al., 1998)

$$\begin{aligned}\partial_t \mu_t &= -\nabla_{W_2} \mathrm{KL}(\mu_t | p_\theta(\cdot|y)) \\ &= \mathrm{div}\left(\mu_t \nabla \log\left(\frac{\mu_t}{p_\theta(\cdot|y)}\right)\right) \\ &= -\mathrm{div}\left(\mu_t \nabla \log\left(p_\theta(\cdot|y)\right)\right) + \Delta \mu_t.\end{aligned}$$

Wasserstein Gradient Flow

If the metric is the Wasserstein-2 distance we obtain the **Wasserstein gradient flow PDE** (Jordan et al., 1998)

$$\begin{aligned}\partial_t \mu_t &= -\nabla_{W_2} \mathrm{KL}(\mu_t | p_\theta(\cdot|y)) \\ &= \mathrm{div}\left(\mu_t \nabla \log\left(\frac{\mu_t}{p_\theta(\cdot|y)}\right)\right) \\ &= -\mathrm{div}\left(\mu_t \nabla \log\left(p_\theta(\cdot|y)\right)\right) + \Delta \mu_t.\end{aligned}$$

Using the connection between Fokker–Plank PDEs and SDEs we obtain

$$\begin{aligned}dX_t &= \nabla \log p_\theta(X_t|y)dt + \sqrt{2}dB_t \\ &= \nabla \log p_\theta(X_t, y)dt + \sqrt{2}dB_t\end{aligned}$$

which is known as the **Langevin diffusion**.

Langevin based algorithms

Simple Euler–Maruyama discretisation leads to the **Unadjusted Langevin Algorithm** (ULA; Durmus et al. (2019))

$$X_{n+1} = X_n + \gamma \nabla \log p_\theta(X_n, y) + \sqrt{2\gamma}\xi_{n+1}$$

where $(\xi_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. d-dimensional standard Gaussian random variables.

Langevin based algorithms

Simple Euler–Maruyama discretisation leads to the **Unadjusted Langevin Algorithm** (ULA; Durmus et al. (2019))

$$X_{n+1} = X_n + \gamma \nabla \log p_\theta(X_n, y) + \sqrt{2\gamma}\xi_{n+1}$$

where $(\xi_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. d-dimensional standard Gaussian random variables.

Many others:

- Metropolis adjusted Langevin algorithm (MALA; Roberts and Tweedie (1996))
- Random walk Metropolis (RWM; Gelman et al. (1997))

Fisher–Rao Gradient Flow

If the metric is the Fisher–Rao metric (or Hellinger) we obtain the **Fisher–Rao gradient flow PDE**

$$\partial_t \mu_t = \mu_t \left(\log \left(\frac{p_\theta(\cdot|y)}{\mu_t} \right) - \mathbb{E}_{\mu_t} \left[\log \left(\frac{p_\theta(\cdot|y)}{\mu_t} \right) \right] \right)$$

which has analytic solution (Chen et al., 2023)

$$\mu_t(x) \propto \mu_0(x)^{e^{-t}} p_\theta(x|y)^{1-e^{-t}} \propto \mu_0(x)^{e^{-t}} p_\theta(x,y)^{1-e^{-t}}$$

Fisher–Rao with importance sampling

Time discretisation of the FR gradient flow:

$$\mu_n(x) \propto \mu_0(x)^{e^{-t_n}} p_\theta(x, y)^{1-e^{-t_n}}$$

Fisher–Rao with importance sampling

Time discretisation of the FR gradient flow:

$$\mu_n(x) \propto \mu_0(x)^{e^{-t_n}} p_\theta(x,y)^{1-e^{-t_n}}$$

We can obtain μ_n from μ_{n-1}:

$$\mu_n(x) \propto \mu_0(x)^{e^{-t_n}} p_\theta(x,y)^{1-e^{-t_n}} = \frac{p_\theta(x,y)^{1-e^{-t_n}}}{p_\theta(x,y)^{1-e^{-t_{n-1}}}} p_\theta(x,y)^{1-e^{-t_{n-1}}} \frac{\mu_0(x)^{e^{-t_n}}}{\mu_0(x)^{e^{-t_{n-1}}}} \mu_0(x)^{e^{-t_{n-1}}}$$

$$= \left(\frac{p_\theta(x,y)}{\mu_{n-1}(x)}\right)^{1-e^{-(t_{n-1}-t_n)}} \mu_{n-1}(x)$$

Fisher–Rao with importance sampling

Time discretisation of the FR gradient flow:

$$\mu_n(x) \propto \mu_0(x)^{e^{-t_n}} p_\theta(x,y)^{1-e^{-t_n}}$$

We can obtain μ_n from μ_{n-1}:

$$\mu_n(x) \propto \mu_0(x)^{e^{-t_n}} p_\theta(x,y)^{1-e^{-t_n}} = \frac{p_\theta(x,y)^{1-e^{-t_n}}}{p_\theta(x,y)^{1-e^{-t_{n-1}}}} p_\theta(x,y)^{1-e^{-t_{n-1}}} \frac{\mu_0(x)^{e^{-t_n}}}{\mu_0(x)^{e^{-t_{n-1}}}} \mu_0(x)^{e^{-t_{n-1}}}$$

$$= \left(\frac{p_\theta(x,y)}{\mu_{n-1}(x)}\right)^{1-e^{-(t_{n-1}-t_n)}} \mu_{n-1}(x)$$

If we have $X_{n-1}^1, \dots X_{n-1}^N \sim \mu_{n-1}$ we can approximate μ_n by **importance sampling** with weights

$$W_n^i = \left(\frac{p_\theta(X_n^i, y)}{\mu_{n-1}(X_n^i)}\right)^{1-e^{-(t_{n-1}-t_n)}}$$

Wasserstein or Fisher–Rao?

Wasserstein

- requires gradients ($\nabla_x \log p_\theta(x, y)$)
- works well only on some classes of distributions (but is extremely fast on those)

Fisher–Rao

- can be used when $p_\theta(x, y)$ is not differentiable in x (and even in discrete spaces!)
- generally slower

1 Latent Variable Models & Expectation Maximisation

2 An Optimisation Point of View

- Optimising θ
- Optimising the posterior
- Joining the dots...

3 Experiments

Joining the dots...

In empirical Bayes we want to solve

$$\min_{(\theta,\mu)} \mathrm{KL}(\mu | p_\theta(\cdot, y))$$

- θ-**update**:

$$\theta_{n+1} = \theta_n - \gamma \nabla_\theta \,\mathrm{KL}(\mu | p_\theta(\cdot, y))_{|\theta=\theta_n} = \theta_n + \gamma \int \nabla_\theta \log p_\theta(x, y)_{|\theta=\theta_n} \mu(x) dx$$

  but μ is unknown and evolves in time

- **posterior update**:

$$\min_{\mu} \mathrm{KL}(\mu | p_\theta(\cdot | y))$$

  but θ is unknown and evolves in time

Joining the dots...

In empirical Bayes we want to solve

$$\min_{(\theta,\mu)} \mathrm{KL}(\mu|p_\theta(\cdot, y))$$

- θ-**update**:

$$\theta_{n+1} = \theta_n - \gamma \nabla_\theta \,\mathrm{KL}(\mu|p_\theta(\cdot, y))_{|\theta=\theta_n} = \theta_n + \gamma \int \nabla_\theta \log p_\theta(x, y)_{|\theta=\theta_n} \mu(x) dx$$

  but μ is unknown and evolves in time $\Rightarrow$ use Monte Carlo to approximate μ

- **posterior update**:

$$\min_{\mu} \mathrm{KL}(\mu|p_\theta(\cdot|y))$$

  but θ is unknown and evolves in time $\Rightarrow$ Plug in θ_n

Gradient Descent & Wasserstein Gradient Flow[2]

- θ-**update**

$$\theta_{n+1} = \theta_n + \gamma \int \nabla_\theta \log p_\theta(x, y)_{|\theta=\theta_n} \mu(x) dx \approx \theta_n + \gamma \frac{1}{N} \sum_{i=1}^{N} \nabla_\theta \log p_{\theta_n}(X_n^i, y)$$

- **posterior update** (N copies)

$$X_{n+1}^i = X_n^i + \gamma \nabla_x \log p_{\theta_n}(X_n^i, y) + \sqrt{2\gamma}\xi_{n+1} \qquad i = 1, \ldots, N$$

where $(\xi_n)_{n\in\mathbb{N}}$ is a sequence of i.i.d. d-dimensional standard Gaussian random variables.

[2] Kuntz et al. (2023), Akyildiz et al. (2025)

Gradient Descent & Fisher–Rao Gradient Flow[3]

- θ-**update**

$$\theta_{n+1} = \theta_n + \gamma \int \nabla_\theta \log p_\theta(x, y)_{|\theta=\theta_n} \mu(x) dx \approx \theta_n + \gamma \sum_{i=1}^{N} W_n^i \nabla_\theta \log p_{\theta_n}(X_n^i, y)$$

- **posterior update** $X_n^1, \dots X_n^N \sim \mu_n$ use importance sampling with weights

$$W_{n+1}^i = \left(\frac{p_{\theta_n}(X_n^i, y)}{\mu_n(X_n^i)} \right)^{1 - e^{-(t_{n-1} - t_n)}}$$

[3] Crucinio (2025)

Experiments

Bayesian logistic regression

$$p_\theta(x) = \mathcal{N}(x; \theta, \mathrm{Id}_{d_x}) \qquad p_\theta(y|x) = \prod_{j=1}^{d_y} s(v_j^T x)^{y_j}(1 - s(v_j^T x))^{1-y_j}$$

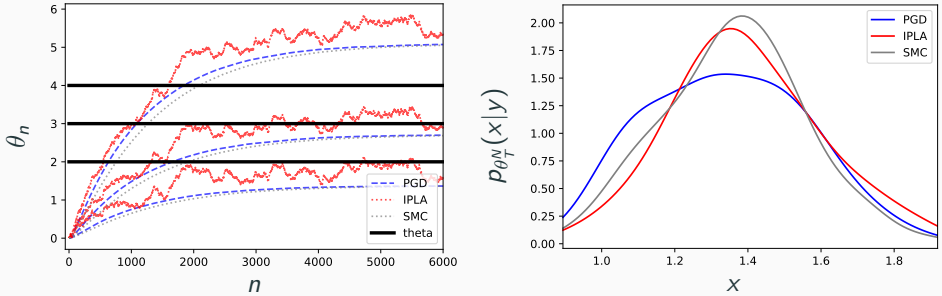
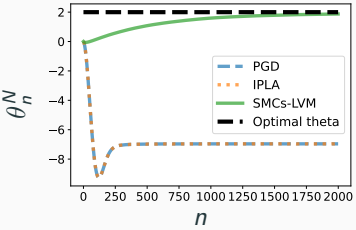
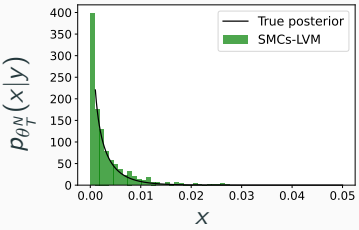


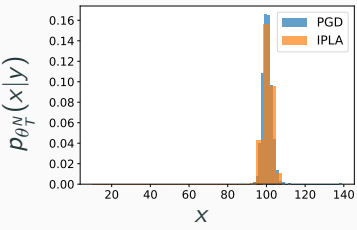
Figure 1: θ-iterates and first component of the approximate posterior. Comparing with IPLA and PGD, which also minimise $\mathrm{KL}(\mu|p_\theta(\cdot, y))$ but use gradient based methods to sample from the posterior.

Non-smooth Likelihood

$$p_\theta(x) = \mathrm{Gamma}(x; \alpha, \beta) \qquad p_\theta(y|x) = \mathcal{N}(y; \theta, x^{-1})$$







Gaussian Mixture Model

$$p_\theta(y) = \alpha\mathcal{N}(y;\theta,1) + (1-\alpha)\mathcal{N}(y;-\theta,1)$$

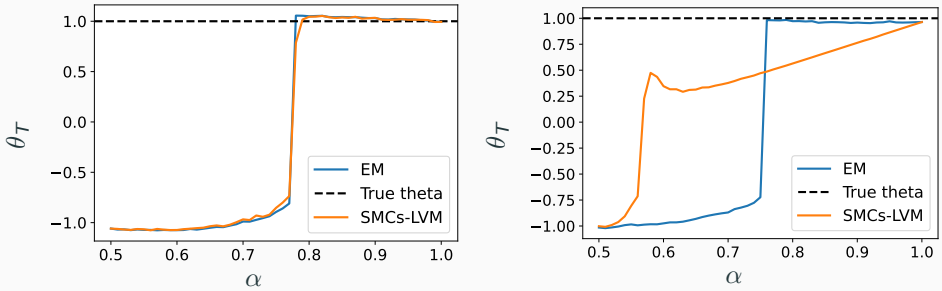


Figure 2: Left: uniform proposal. Right: α-informed proposal

Stochastic Block Model[4]

- allocation probabilities: $\mathbb{P}(x = q) = p_q$
- edge probabilities: $y_{ij}|x_i, x_j \sim \text{Bernoulli}(\nu_{x_i x_j})$

The set of parameters is $\theta = \left((p_q)_{q=1}^{Q}, (\nu_{ql})_{q,l=1}^{Q}\right)$.

[4] Image from Wikipedia

Stochastic Block Model

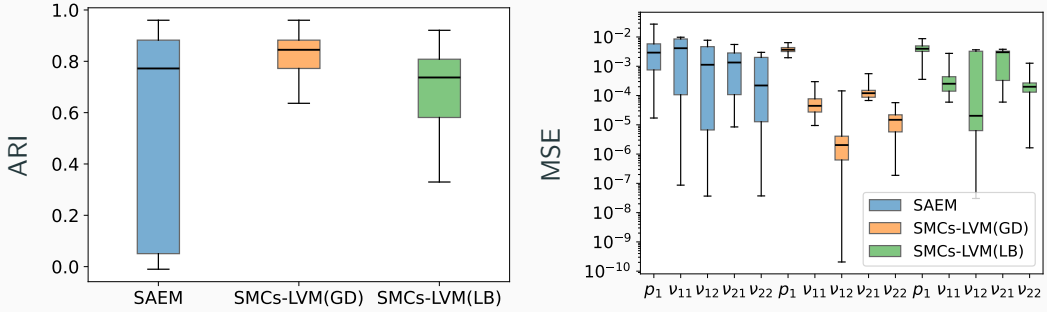


Figure 3: Distribution of ARI and MSE for 50 repetitions of SAEM and SMCs-LVM with logarithmic barrier (LB) and gradient descent (GD) update for θ for the stochastic block model on synthetic data.

Conclusions

- Perform optimisation and sampling simultaneously (unlike EM) can be beneficial...
- ...but more expensive than EM
- Gradient based sampling works well...
- ...but cannot handle discrete latent spaces...
- ... importance sampling can...
- ... but is more expensive than gradient based methods

Conclusions

- Perform optimisation and sampling simultaneously (unlike EM) can be beneficial...
- ...but more expensive than EM
- Gradient based sampling works well...
- ...but cannot handle discrete latent spaces...
- ... importance sampling can...
- ... but is more expensive than gradient based methods

Thank you!

References

Ö. Deniz Akyildiz, Francesca Romana Crucinio, Mark Girolami, Tim Johnston, and Sotirios Sabanis. Interacting Particle Langevin Algorithm for Maximum Marginal Likelihood Estimation. *ESAIM: PS*, 29:243–280, 2025.

Yifan Chen, Daniel Zhengyu Huang, Jiaoyang Huang, Sebastian Reich, and Andrew M Stuart. Sampling via gradient flows in the space of probability measures. *arXiv preprint arXiv:2310.03597*, 2023.

Francesca Romana Crucinio. A mirror descent approach to maximum likelihood estimation in latent variable models. *Journal of Computational and Graphical Statistics*, (just-accepted):1–19, 2025.

Alain Durmus, Szymon Majewski, and Błażej Miasojedow. Analysis of Langevin Monte Carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

A. Gelman, W. R. Gilks, and G. O. Roberts. Weak convergence and optimal scaling of random walk Metropolis algorithms. *The Annals of Applied Probability*, 7(1), February 1997. ISSN 1050-5164. doi: 10.1214/aoap/1034625254. URL `https://projecteuclid.org/journals/annals-of-applied-probability/volume-7/issue-1/Weak-convergence-and-optimal-scaling-of-random-walk-Metropolis-algorithms/10.1214/aoap/1034625254.full`.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the Fokker–Planck equation. *SIAM Journal on Mathematical Analysis*, 29(1):1–17, 1998.

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

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, 1998.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, pages 341–363, 1996.

Filippo Santambrogio. {Euclidean, metric, and Wasserstein} gradient flows: an overview. *Bulletin of Mathematical Sciences*, 7(1):87–154, 2017.