Proximal Interacting Particle Langevin Algorithms

Francesca R. Crucinio

King's College London

Joint work with Deniz Akyildiz, Paula Cordero Encinar, Mark Girolami, Tim Johnston, Sotirios Sabanis

Outline

Latent Variable Models (LVM)

Consider the following data-generating process

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

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

Given a data point y we want to find $\theta_\star$ maximising the marginal log-likelihood

$$\log p_\theta(y) = \log \int_{\mathbb{R}^{d_x}} p_\theta(x, y) \mathrm{d}x,$$

where $p_\theta(x, y) = p_\theta(x) p_\theta(y|x)$.

Expectation Maximisation (EM)

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

$$Q(\theta|\theta^{(n)}) = \int_{\mathbb{R}^{d_x}} \log p_\theta(x, y) p_{\theta^{(n)}}(x|y) \mathrm{d}x,$$

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

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

An Optimisation Point of View

Our aim is to find $\theta_\star$ maximising

$$k(\theta) := p_\theta(y) = \int p_\theta(x, y)\mathrm{d}x = \int e^{-U(\theta,x)}\mathrm{d}x,$$

with $U(\theta, x) := -\log p_\theta(x, y)$.

This is a well-studied problem in optimisation, one solution is to find a **measure** which concentrates around $\theta_\star$ and use standard tools to **sample** from this measure.

E.g. **simulated annealing**, set $k(\theta)^N$ and let $N \to \infty$.

Simulated Annealing for LVM

The extended target

$$\pi^N(\theta, x_1, x_2, ..., x_N) \propto \exp\left(-\sum_{i=1}^{N} U(\theta, x_i)\right)$$

admits as θ-marginal

$$\pi_\Theta^N(\theta) \propto \int_{\mathbb{R}^{d_x}} ... \int_{\mathbb{R}^{d_x}} \exp\left(-\sum_{i=1}^{N} U(\theta, x_i)\right) \mathrm{d}x_1 \mathrm{d}x_2 ... \mathrm{d}x_N$$
$$= \left(\int_{\mathbb{R}^{d_x}} e^{-U(\theta,x)} \mathrm{d}x\right)^N = k(\theta)^N,$$

which as $N \to \infty$ concentrates on $\theta_\star$.

Outline

Langevin Dynamics

The Langevin diffusion

$$\mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

has invariant measure $\pi \propto e^{-U}$.

Langevin Dynamics

The Langevin diffusion

$$\mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

has invariant measure $\pi \propto e^{-U}$.
The diffusion

$$\mathrm{d}X_t = -\nabla U(X_t)\mathrm{d}t + \sqrt{2/\beta}\mathrm{d}W_t$$

has invariant measure $\pi_\beta \propto e^{-\beta U}$, where β is known as the *inverse temperature parameter*.
As $\beta \to \infty$, π_β concentrates around its modal points.

Interacting Particle Langevin Algorithm (IPLA)

To sample from our target measure we use the following interacting particle system (IPS) of N particles

$$\mathrm{d}\boldsymbol{\theta}_t^N = -\frac{1}{N}\sum_{j=1}^{N} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}_t^N, \boldsymbol{X}_t^{j,N})\mathrm{d}t + \sqrt{\frac{2}{N}}\mathrm{d}\boldsymbol{B}_t^{0,N}, \tag{1}$$

$$\mathrm{d}\boldsymbol{X}_t^{i,N} = -\nabla_x U(\boldsymbol{\theta}_t^N, \boldsymbol{X}_t^{i,N})\mathrm{d}t + \sqrt{2}\mathrm{d}\boldsymbol{B}_t^{i,N}, i = 1, 2, ..., N.$$

Although (1) is an IPS, we consider it a diffusion evolving on $\mathbb{R}^{d_x} \times (\mathbb{R}^{d_\theta})^N$ and use techniques from **Langevin-based algorithms**.

Algorithm

Euler–Maruyama discretisation of Langevin IPS with stepsize γ

$$\theta_{n+1}^N = \theta_n^N - \frac{\gamma}{N}\sum_{j=1}^N \nabla_\theta U(\theta_n^N, X_n^{j,N}) + \sqrt{\frac{2}{N}}\xi_{n+1}^{0,N}$$

$$X_{n+1}^{i,N} = X_n^{i,N} - \gamma \nabla_x U(\theta_n^N, X_n^{i,N}) + \sqrt{2}\,\xi_{n+1}^{i,N}$$

Assumptions

A1. (Lipschitz) Let $v = (\theta, x)$ and $v' = (\theta', x')$. We assume that there exist $L > 0$ such that

$$\|\nabla U(v) - \nabla U(v')\| \leq L\|v - v'\|.$$

A2. (Convexity) Let $v = (\theta, x)$. Then, there exists $\mu > 0$ such that

$$\left\langle v - v', \nabla U(v) - \nabla U(v') \right\rangle \geq \mu\|v - v'\|^2,$$

for all $v, v' \in \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_x}$.

Main Convergence Result

$$
\begin{aligned}
\mathbb{E}[\|\theta_n^N - \theta_\star\|^2]^{1/2} \leq & \sqrt{\frac{d_\theta}{N\mu}} \\
& + e^{-\mu n\gamma}\left(\mathbb{E}[\|Z_0^N - z_\star\|^2]^{1/2} + \left(\frac{d_x N + d_\theta}{N\mu}\right)^{1/2}\right) \\
& + C(1 + \sqrt{d_\theta/N + d_x})\gamma^{1/2},
\end{aligned}
$$

- $z_\star = (\bar{\theta}_\star, N^{-1/2}x_\star, \ldots, N^{-1/2}x_\star)$ and $(\bar{\theta}_\star, x_\star)$ is the minimiser of U
- Z_0^N is the initial condition
- $C > 0$ is a constant independent of $n, N, \gamma, d_\theta, d_x$

Proof Idea

We split the error of the algorithm as

$$W_2(\delta_{\theta_\star}, \mathcal{L}(\theta_n^N)) \leq W_2(\delta_{\theta_\star}, \pi_\Theta^N) + W_2(\pi_\Theta^N, \mathcal{L}(\boldsymbol{\theta}_{\gamma n})) + W_2(\mathcal{L}(\boldsymbol{\theta}_{\gamma n}), \mathcal{L}(\theta_n^N))$$

- $W_2(\delta_{\theta_\star}, \pi_\Theta^N)$ is the *concentration* of the invariant measure on $\theta_\star$
- $W_2(\pi_\Theta^N, \mathcal{L}(\boldsymbol{\theta}_{\gamma n}))$ is the *convergence* of the continuous time process to its invariant measure
- $W_2(\mathcal{L}(\boldsymbol{\theta}_{\gamma n}), \mathcal{L}(\theta_n^N))$ is error due to *time discretisation*

Toy Example

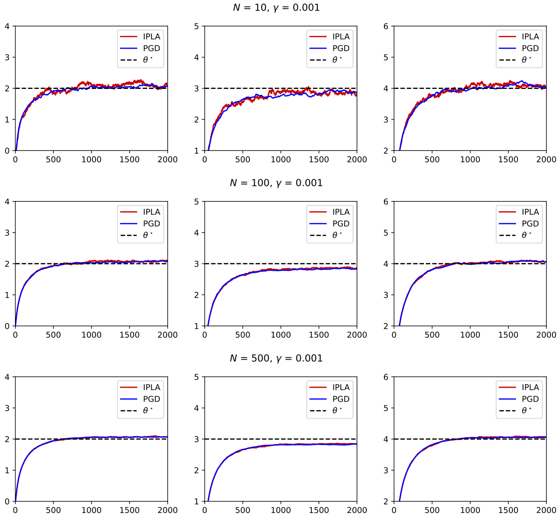
Bayesian logistic regression LVM where for $\theta \in \mathbb{R}^{d_\theta}$

$$p_\theta(x) = \mathcal{N}(x; \theta, \sigma^2 \mathrm{Id}_{d_x}),$$

$$p_\theta(y|x) = \prod_{j=1}^{d_y} s(v_j^T x)^{y_j}(1 - s(v_j^T x))^{1-y_j},$$

with $d_\theta = d_x$, $s(u) := e^u/(1+e^u)$ the logistic function and $\{v_j\}_{j=1}^{d_y} \in \mathbb{R}^{d_x}$ a set of covariates with corresponding binary responses $\{y_j\}_{j=1}^{d_y} \in \{0, 1\}$.

IPLA vs PGD



Outline

Non-differentiable Targets

Consider the case in which

$$U(\theta, x) = -\log p_\theta(x, y) = g_1(\theta, x) + g_2(\theta, x),$$

with $g_1 \in \mathcal{C}^1$ and g_2 not $\mathcal{C}^1$ but convex and lower semi-continuous.

- Lasso regularisation
- the elastic net
- total-variation norm

What's the need?

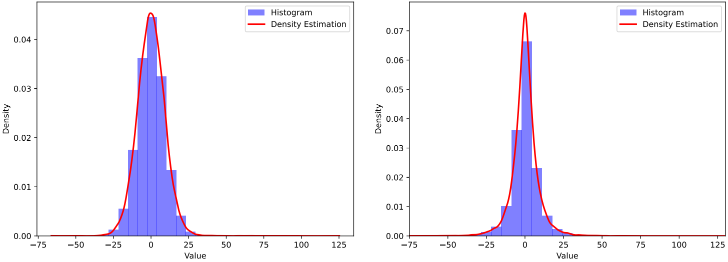


Figure: Normal vs Laplace prior. Histogram and density estimation of the weights of a BNN for a randomly chosen particle from the final (500 steps) cloud of 100 particles.

Proximity map

Proximity map

For U convex, proper and lower semi-continuous and $\lambda > 0$

$$\mathrm{prox}_U^{\lambda}(x) := \arg\min_{z \in \mathbb{R}^d} \left\{ U(z) + \|z - x\|^2/(2\lambda) \right\}.$$

Moves points in the direction of the minimum of U acting as a "gradient".

Moreau-Yosida envelope

Moreau-Yosida envelope

For any $\lambda > 0$, define the λ-Moreau-Yosida approximation of U as

$$U^{\lambda}(x) := \min_{z \in \mathbb{R}^d} \left\{ U(z) + \|z - x\|^2/(2\lambda) \right\}.$$

Take $\pi(x) \propto \exp(-U(x))$. We we define the λ-Moreau-Yosida approximation of π as the following density

$$\pi_{\lambda}(x) \propto \exp(-U^{\lambda}(x))$$

Moreau-Yosida envelope

Moreau-Yosida envelope

For any $\lambda > 0$, define the λ-Moreau-Yosida approximation of U as

$$U^{\lambda}(x) := \min_{z \in \mathbb{R}^d} \left\{ U(z) + \|z - x\|^2 / (2\lambda) \right\}.$$

Take $\pi(x) \propto \exp(-U(x))$. We we define the λ-Moreau-Yosida approximation of π as the following density

$$\pi_\lambda(x) \propto \exp(-U^{\lambda}(x))$$

- converge (pointwise, in TV, ...) to π as $\lambda \to 0$
- π_λ is continuously differentiable with $\nabla \log \pi_\lambda(x) = \lambda^{-1}(x - \mathrm{prox}_U^{\lambda}(x))$

Moreau-Yosida envelope

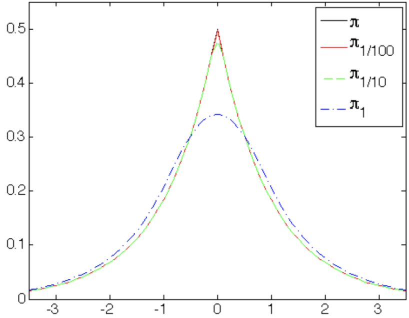


Figure: Moreau-Yoshida envelope for the Laplace distribution $\pi(x) \propto \exp(-|x|)$ (Pereyra, 2016).

Moreau-Yosida Langevin Dynamics

Since $\pi_\lambda \propto e^{-U^\lambda}$ is now continuously differentiable, we can write the Langevin diffusion

$$\mathrm{d}X_{\lambda,t} = -\nabla U^\lambda(X_{\lambda,t})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t,$$

or, equivalently,

$$\mathrm{d}X_{\lambda,t} = \lambda^{-1}(\mathrm{prox}_U^\lambda(X_{\lambda,t}) - X_{\lambda,t})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t.$$

The resulting algorithm is known as MY-ULA (Durmus et al., 2018; Pereyra, 2016).

Proximal Interacting Particle Langevin SDE

We consider as target measure π_N^λ, the Moreau-Yosida envelope of π_N and obtain

$$
\begin{aligned}
\mathrm{d}\boldsymbol{\theta}_t^N &= -\frac{1}{N}\sum_{j=1}^{N} \nabla_\theta U^\lambda(\boldsymbol{\theta}_t^N, \boldsymbol{X}_t^{j,N})\mathrm{d}t + \sqrt{\frac{2}{N}}\mathrm{dB}_t^{0,N} \\
\mathrm{d}\boldsymbol{X}_t^{i,N} &= -\nabla_x U^\lambda(\boldsymbol{\theta}_t^N, \boldsymbol{X}_t^{i,N})\mathrm{d}t + \sqrt{2}\mathrm{dB}_t^{i,N}.
\end{aligned}
$$

For regular enough U this SDE approaches the IPLA SDE as $\lambda \to 0$.

Moreau-Yosida Interacting Particle Langevin Algorithm (MYIPLA)

If $U = g_1 + g_2$, we can take $U^\lambda = g_1 + g_2^\lambda$ so that

$$\nabla U^\lambda(v) = \nabla g_1(v) + \lambda^{-1}(v - \mathrm{prox}_{g_2}^\lambda(v))$$

and obtain

$$\mathrm{d}\boldsymbol{\theta}_t^N = \frac{1}{N}\sum_{j=1}^N \left(-\nabla_\theta g_1(\boldsymbol{\theta}_t^N, \boldsymbol{X}_t^{j,N}) + \lambda^{-1}(\mathrm{prox}_{g_2}^\lambda(\boldsymbol{\theta}_t^N, \boldsymbol{X}_t^{j,N})_\theta - \boldsymbol{\theta}_t^N)\right)\mathrm{d}t + \sqrt{\frac{2}{N}}\mathrm{dB}_t^{0,N}$$

$$\mathrm{d}\boldsymbol{X}_t^{i,N} = \left(-\nabla_x g_1(\boldsymbol{\theta}_t^N, \boldsymbol{X}_t^{i,N}) + \lambda^{-1}(\mathrm{prox}_{g_2}^\lambda(\boldsymbol{\theta}_t^N, \boldsymbol{X}_t^{i,N})_x - \boldsymbol{X}_t^{i,N})\right)\mathrm{d}t + \sqrt{2}\mathrm{dB}_t^{i,N}.$$

Algorithm

Euler–Maruyama discretisation of proximal Langevin IPS with stepsize γ

$$\theta_{n+1}^N = \Big(1 - \frac{\gamma}{\lambda}\Big)\theta_n^N + \frac{\gamma}{N}\sum_{i=1}^N \Big(- \nabla_\theta g_1(\theta_n^N, X_n^{i,N}) + \frac{1}{\lambda}\operatorname{prox}_{g_2}^{\lambda}(\theta_n^N, X_n^{i,N})_\theta \Big) + \sqrt{\frac{2\gamma}{N}}\xi_{n+1}^{0,N}$$

$$X_{n+1}^{i,N} = \Big(1 - \frac{\gamma}{\lambda}\Big)X_n^{i,N} - \gamma \nabla_x g_1(\theta_n^N, X_n^{i,N}) + \frac{\gamma}{\lambda}\operatorname{prox}_{g_2}^{\lambda}(\theta_n^N, X_n^{i,N})_x + \sqrt{2\gamma}\, \xi_{n+1}^{i,N}$$

Assumptions – g_1

A1. (Lipschitz) Let $v = (\theta, x)$ and $v' = (\theta', x')$. We assume that there exist $L > 0$ such that

$$\|\nabla g_1(v) - \nabla g_1(v')\| \leq L\|v - v'\|.$$

A2. (Convexity) Let $v = (\theta, x)$. Then, there exists $\mu > 0$ such that

$$\left\langle v - v', \nabla g_1(v) - \nabla g_1(v') \right\rangle \geq \mu\|v - v'\|^2,$$

for all $v, v' \in \mathbb{R}^{d_\theta} \times \mathbb{R}^{d_x}$.

Assumptions – g_2

A3. g_2 is proper, convex, lower semi-continuous and lower bounded.

A4. Let $v = (\theta, x)$ and $v' = (\theta', x')$. We assume that

$$\|g_2(v) - g_2(v')\| \leq \|g_2\|_{\text{Lip}} \|v - v'\|.$$

Main Convergence Result

$$
\begin{aligned}
\mathbb{E}[\|\theta_n^N - \theta_\star\|^2]^{1/2} \leq & \frac{\lambda}{\mu}\Big(\frac{\|g_2\|_{\mathrm{Lip}}^2}{2} A + B\Big) + \sqrt{\frac{d_\theta}{N\mu}} \\
& + e^{-\mu n \gamma}\left(\mathbb{E}[\|Z_0^N - z_\star\|^2]^{1/2} + \Big(\frac{d_x N + d_\theta}{N\mu}\Big)^{1/2}\right) \\
& + C(1 + \sqrt{d_\theta/N + d_x})\gamma^{1/2},
\end{aligned}
$$

- $z_\star = (\bar{\theta}_\star, N^{-1/2}x_\star, \ldots, N^{-1/2}x_\star)$ and $(\bar{\theta}_\star, x_\star)$ is the minimiser of U^λ
- Z_0^N is the initial condition
- $A, B, C > 0$ is a constant independent of $n, N, \gamma, \lambda, d_\theta, d_x$

Proof Idea

Overall, we split the error of the algorithm as

$$W_2(\delta_{\theta_\star}, \mathcal{L}(\theta_n^N)) \leq \|\theta_\star - \theta_{\star,\lambda}\| + W_2(\delta_{\theta_{\star,\lambda}}, \mathcal{L}(\theta_n^N))$$

- $\|\theta_\star - \theta_{\star,\lambda}\|$ distance between the minimiser of $p_\theta(y)$ and that of its MY envelope
- $W_2(\delta_{\theta_\star}, \mathcal{L}(\theta_n^N))$ combines concentration, convergence and time discretisation error

Toy Example

Bayesian logistic regression LVM where for $\theta \in \mathbb{R}^{d_\theta}$

$$p_\theta(x) = \prod_{i=1}^{d_x} \text{Laplace}(x_i|\theta, 1),$$

$$p_\theta(y|x) = \prod_{j=1}^{d_y} s(v_j^T x)^{y_j}(1 - s(v_j^T x))^{1-y_j},$$

with $d_\theta = d_x$, $s(u) := e^u/(1 + e^u)$ the logistic function and $\{v_j\}_{j=1}^{d_y} \in \mathbb{R}^{d_x}$ a set of covariates with corresponding binary responses $\{y_j\}_{j=1}^{d_y} \in \{0, 1\}$.

Toy Example: Results

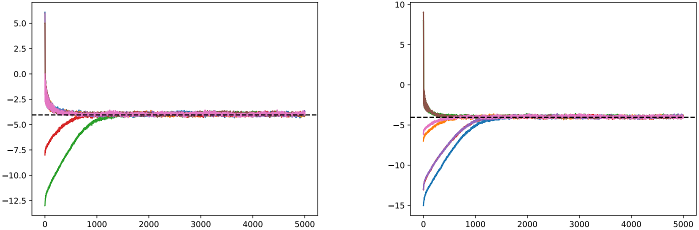


Figure: MYIPLA. Approximate vs Iterative solver for prox.

Bayesian Neural Network: Laplace prior

Bayesian two-layer neural network to classify MNIST images.
The latent variables are the weights, $w \in \mathbb{R}^{d_w := 40 \times 784}$, of the input layer and those, $v \in \mathbb{R}^{d_v := 2 \times 40}$, of the output layer.

$$
\begin{aligned}
p(l|f, x) &\propto \exp\left(\sum_{j=1}^{40} v_{lj} \tanh\left(\sum_{i=1}^{784} w_{ji} f_i\right)\right) \\
p_\alpha(w) &= \prod_i \mathrm{Laplace}(w_i | 0, e^{2\alpha}) \\
p_\beta(v) &= \prod_i \mathrm{Laplace}(v_i | 0, e^{2\beta})
\end{aligned}
$$

with $\theta = (\alpha, \beta)$.

Bayesian Neural Network: Laplace prior

Prior	% of zero weights		Thresholds		Error (%)	LPD
	Layer 1	Layer 2	Layer 1	Layer 2		
Laplace	74	48	0.2	0.2	7	-0.23
Normal	74	48	0.5	1.1	15	-0.74
	16	15	0.2	0.2	16	-0.78

Bayesian Neural Network: Laplace prior

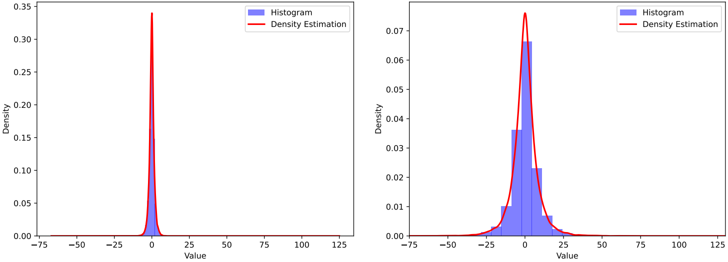


Figure: MYIPLA vs IPLA prior. Histogram and density estimation of the weights of a BNN with Laplace prior for a randomly chosen particle from the final (500 steps) cloud of 100 particles.

Conclusions I

We propose a family of algorithms to find the MLE in LVM which exploits

- scaling of Langevin diffusions
- optimisation perspective
- combines expectation and maximisation steps
- allows for non-differentiable prior/likelihoods
- returns approximations of both $\theta_\star$ and $p_{\theta_\star}(x|y)$

Conclusions II

There's more to do!

- other algorithms to sample from π^N can be constructed
- for ProxIPLA other discretisations exists (as well as a PGD equivalent)
- it should be possible to extend the analysis to the non-convex case

Both papers are on arXiv:

- IPLA `https://arxiv.org/abs/2303.13429`
- ProxIPLA `https://arxiv.org/abs/2406.14292`

Both papers are on arXiv:

- IPLA `https://arxiv.org/abs/2303.13429`
- ProxIPLA `https://arxiv.org/abs/2406.14292`

Thank you!

Bibliography I

Alain Durmus, Éric Moulines, and Marcelo Pereyra. Efficient Bayesian Computation by Proximal Markov Chain Monte Carlo: When Langevin Meets Moreau. *SIAM Journal on Imaging Sciences*, 11(1):473–506, 2018.

Marcelo Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.