Interacting Particle Langevin Algorithm for Maximum Marginal Likelihood Estimation

Tim Johnston [1] and Francesca R. Crucinio [2]

Joint work with Deniz Akyildiz [3], Mark Girolami [4,5] and Sotirios Sabanis [1,4,6]

[1] School of Mathematics, University of Edinburgh

[2] CREST, ENSAE Paris

[3] Department of Mathematics, Imperial College London

[4] The Alan Turing Institute

[5] Department of Engineering, University of Cambridge

[6] National Technical University of Athens

24 March 2023

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

Latent Variable Models

Applications

- Inference with incomplete data (Dempster et al., 1977)
- classification tasks, generative modeling, dimension reduction
- ...

Methods

- expectation maximisation (EM) (Dempster et al., 1977)
- variational methods (Carlin and Louis, 2000; Kingma and Welling, 2013; Burda et al., 2016)
- simulated annealing

EM and Variants

EM is a standard way to maximise the marginal likelihood $p_\theta(y)$ consisting of

E-step w.r.t. *latent variables* x:
compute $\log p_\theta(y) = \int_{\mathbb{R}^{d_x}} \log p_\theta(x, y)\mathrm{d}x$ for fixed θ

M-step w.r.t. *parameters* θ:
maximise $\log p_\theta(y)$

Requires tractability of both steps. Extensions include

- approximations of the **E-step**:
  - simple Monte Carlo (Wei and Tanner, 1990; Sherman et al., 1999)
  - stochastic approximations and MCMC (Delyon et al., 1999; Celeux and Diebolt, 1992)
  - unadjusted schemes (De Bortoli et al., 2021)
- numerical optimisation algorithms for the **M-step** (Meng and Rubin, 1993; Liu and Rubin, 1994); require

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

Simulated Annealing for LVM

Consider $p(\theta)$ an instrumental prior and define the posterior for θ

$$p(\theta|y) \propto p(\theta)p_\theta(y)^\beta.$$

One can define the extended target

$$p_\beta(\theta, x_{1:\beta}) \propto p(\theta) \prod_{i=1}^{\beta} p_\theta(x_i, y).$$

Let $\{\beta_t\}_{t\geq 1}$ be an integer sequence diverging to infinity. We can define a tempered sequence of distribution p_{β_t} which as $t \to \infty$ concentrates on $\theta^\star$ (and then sample from it using, e.g., sequential Monte Carlo; Doucet et al. (2002); Johansen et al. (2008)).

An Interacting Particle System for LVM (Kuntz et al., 2023)

Assume a fixed observation $y \in \mathbb{R}^{d_y}$ and define the negative log-likelihood as

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

Kuntz et al. (2023) build an interacting particle system (IPS) build an interacting particle system (IPS) to maximise $p_\theta(y)$ w.r.t. θ

$$d\boldsymbol{\theta}_t^N = -\frac{1}{N}\sum_{j=1}^{N} \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}_t^N, \boldsymbol{X}_t^{j,N})dt,$$

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

$\Rightarrow$ E-step and M-step are performed together but clearly distinguisheable.

An Optimisation Point of View

Our aim is to find $\theta^\star$ maximising

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

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.

Interacting Particle System

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

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

The key property of (1) is that we can **calculate the invariant measure explicitly**.

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

Langevin Dynamics

An **overdamped Langevin diffusion**

$$dX_t = -\nabla u(X_t)dt + \sqrt{2}dW_t \tag{2}$$

has invariant measure $\pi \sim e^{-u}$, and moreover the diffusion

$$dX_t = -\nabla u(X_t)dt + \sqrt{2/\beta}dW_t \tag{3}$$

has invariant measure $\pi_\beta \sim e^{-\beta u}$, where β is known as the **inverse temperature parameter**. It's well known that π_β **concentrates** around it's modal points as $\beta \to \infty$ (and therefore as the noise goes to 0).

One can easily show this using the **Fokker–Planck equation**

Discretisation is known as **Unadjusted Langevin Algorithm (ULA)**, applications: sampling in Durmus and Moulines (2017), CHAU et al. (2021), Brosse et al. (2019) and as part of EM method in De Bortoli et al. (2021).

Target Measure

$$d\boldsymbol{\theta}_t^N = -\frac{1}{N}\sum_{j=1}^N \nabla_{\boldsymbol{\theta}} U(\boldsymbol{\theta}_t^N, \boldsymbol{X}_t^{j,N})dt + \sqrt{\frac{2}{N}}d\boldsymbol{B}_t^{0,N}, \tag{4}$$

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

Note that

$$\nabla\left(\sum_{i=1}^N U(\theta, x_i)\right) = \left(\sum_{i=1}^N \nabla_\theta U(\theta, x_i), \nabla_x U(\theta, x_1), ..., \nabla_x U(\theta, x_N)\right) \tag{5}$$

so (4) is **almost** a Langevin diffusion.

However factor of $1/N$ in the drift of the theta process and the factor $1/\sqrt{N}$ in the noise **cancel out**, so we obtain that our system has invariant measure

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

Theta-Marginal

As a result, the **theta-marginal** π_Θ^N is given as

$$\pi_\Theta^N(\theta) \propto \int_{\mathbb{R}^{d_x}} \dots \int_{\mathbb{R}^{d_x}} \exp\left(-\sum_{i=1}^{N} U(\theta, x_i)\right) dx_1 dx_2 \dots dx_N$$
$$= \left(\int_{\mathbb{R}^{d_x}} e^{-U(\theta,x)} dx\right)^N = k(\theta)^N.$$

Concentration

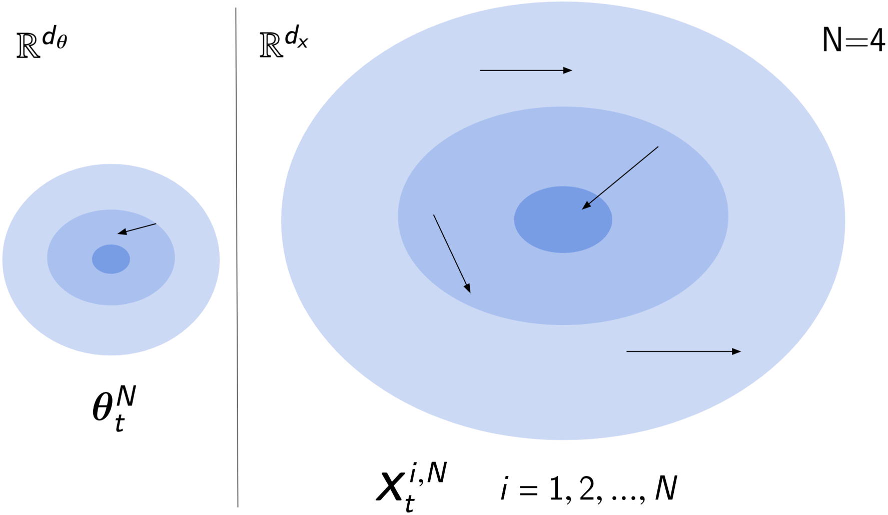
π_Θ^N **concentrates around the maximiser of k** as $N \to \infty$.

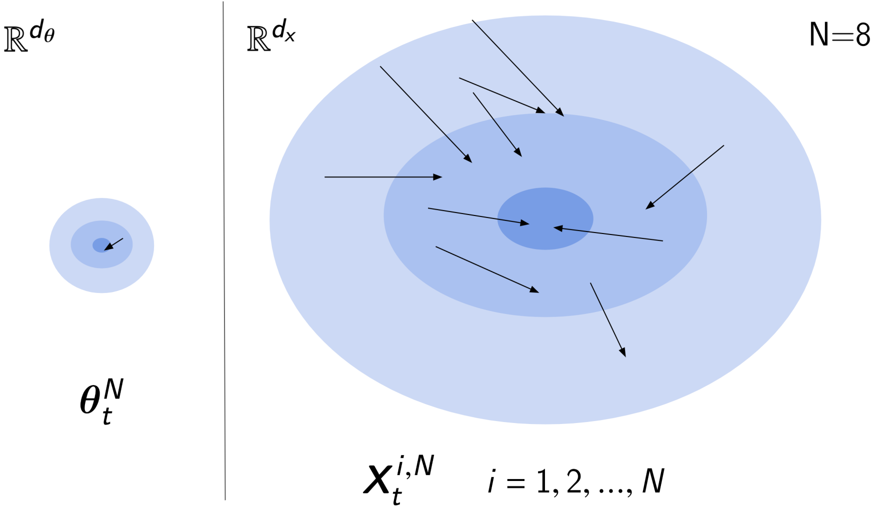
Indeed if $\kappa = -\log k(\theta)$ then $\pi_\Theta^1 \propto e^{-\kappa}$, so $\pi_\Theta^N \propto e^{-N\kappa}$, so it can clearly be seen that N controls the concentration of π_Θ^N just like an **inverse temperature parameter** from Langevin diffusions.

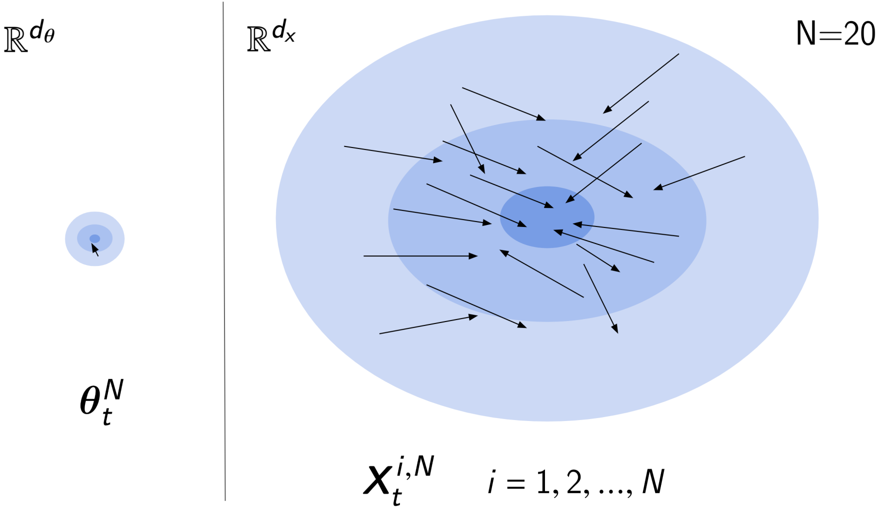
In the convex case:

$$W_2(\delta_{\theta^*}, \pi_\Theta^N) = O(N^{-1/2}) \tag{6}$$

and similar bounds hold in the non-convex setting, see Raginsky et al. (2017).


$\mathbb{R}^{d_\theta}$
$\mathbb{R}^{d_x}$
N=4
θ_t^N
$X_t^{i,N}$ $i = 1, 2, ..., N$


$\mathbb{R}^{d_\theta}$
$\mathbb{R}^{d_x}$
N=8
θ_t^N
$X_t^{i,N} \quad i = 1, 2, \ldots, N$


$\mathbb{R}^{d_\theta}$
$\mathbb{R}^{d_x}$
N=20
θ_t^N
$X_t^{i,N}$ $i = 1, 2, ..., N$

Algorithm

Use non-bold to notate the **discretisation** $(\theta_n^N, X_n^{1,N}, ..., X_n^{1,N})$ with step size γ, specifically

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

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

For algorithm we consider $(\theta_n^N)_{n\geq 0}$ given by above. We only care about the particles $X^{i,N}$, $i = 1, 2, ..., N$ in order to **calculate** $(\theta_n^N)_{n\geq 0}$. So no harm in considering error of **rescaled** system

$$Z_n^N = (\theta_n^N, N^{-1/2} X_n^{1,N}, ..., N^{-1/2} X_n^{N,N}),$$

in order to bound the numerics error $W_2(\mathcal{L}(\boldsymbol{\theta}_{\gamma n}), \mathcal{L}(\theta_n^N))$.

Rescaling

Similarly, when calculating distance distance $W_2(\pi_\Theta^N, \mathcal{L}(\boldsymbol{\theta}_{\gamma n}))$ between continuous $\boldsymbol{\theta}_t$ process and the marginal of the invariant measure, no harm in considering the **rescaled continuous** system

$$\boldsymbol{Z}_t^N = (\boldsymbol{\theta}_t^N, N^{-1/2}\boldsymbol{X}_t^{1,N}, ..., N^{-1/2}\boldsymbol{X}_t^{N,N})$$

Indeed, one has

$$\|\boldsymbol{Z}_t^N - \tilde{\boldsymbol{Z}}_t^N\|^2 = \|z_0 - \tilde{z}_0\|^2 - \frac{2}{N}\sum_{i=1}^{N}\int_0^t \langle \nabla U(\boldsymbol{V}_s^{i,N}) - \nabla U(\tilde{\boldsymbol{V}}_s^{i,N}), \boldsymbol{V}_s^{i,N} - \tilde{\boldsymbol{V}}_s^{i,N}\rangle ds, \quad (7)$$

if $\boldsymbol{Z}_t^N$ and $\tilde{\boldsymbol{Z}}_t^N$ are two versions driven by the same noise, and

$$\boldsymbol{V}_s^{i,N} = (\boldsymbol{\theta}_t^N, \boldsymbol{X}_t^{i,N}), \quad \tilde{\boldsymbol{V}}_s^{i,N} = (\tilde{\boldsymbol{\theta}}_t^N, \tilde{\boldsymbol{X}}_t^{i,N}) \quad (8)$$

Rescaling

Using rescaling we get numerics bounds and convergence to equilibrium bounds that **don't scale with** N. Furthermore one has the useful property

$$\|\boldsymbol{Z}_t^N\|^2 = \frac{1}{N}\sum_{i=1}^{N} \|(\boldsymbol{\theta}_t^N, \boldsymbol{X}_t^{i,N})\|^2 = \frac{1}{N}\sum_{i=1}^{N} \|\boldsymbol{V}_t^{i,N}\|^2 \tag{9}$$

$$\|Z_n^N\|^2 = \frac{1}{N}\sum_{i=1}^{N} \|(\theta_n^N, X_n^{i,N})\|^2 = \frac{1}{N}\sum_{i=1}^{N} \|V_n^{i,N}\|^2 \tag{10}$$

In general using rescaling means one doesn't have to consider $\nabla_\theta U$ and $\nabla_x U$ seperately.

Contraction

Specifically, under **convexity assumption**

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

one has

$$\begin{aligned}
\|\boldsymbol{Z}_t^N - \tilde{\boldsymbol{Z}}_t^N\|^2 &= \|z_0 - \tilde{z}_0\|^2 - \frac{2}{N}\sum_{i=1}^N \int_0^t \langle \nabla U(\boldsymbol{V}_s^{i,N}) - \nabla U(\tilde{\boldsymbol{V}}_s^{i,N}), \boldsymbol{V}_s^{i,N} - \tilde{\boldsymbol{V}}_s^{i,N} \rangle ds \\
&\leq \|z_0 - \tilde{z}_0\|^2 - \frac{2\mu}{N}\sum_{i=1}^N \int_0^t \|\boldsymbol{V}_s^{i,N} - \tilde{\boldsymbol{V}}_s^{i,N}\|^2 ds \\
&\leq \|z_0 - \tilde{z}_0\|^2 - 2\mu \int_0^t \|\boldsymbol{Z}_s^N - \tilde{\boldsymbol{Z}}_s^N\|^2 ds
\end{aligned}$$

and therefore one obtains **exponential convergence to invariant measure**.

Overall, we split the **error** of the algorithm as

$$W_2(\delta_{\theta^*}, \mathcal{L}(\theta_n)) \leq W_2(\delta_{\theta^*}, \pi_\Theta^N) + W_2(\pi_\Theta^N, \mathcal{L}(\boldsymbol{\theta}_{\gamma n})) + W_2(\mathcal{L}(\boldsymbol{\theta}_{\gamma n}), \mathcal{L}(\theta_n^N)) \quad (11)$$

where

- $W_2(\delta_{\theta^*}, \pi_\Theta^N)$ is the difference between the **invariant measure** and θ^*
- $W_2(\pi_\Theta^N, \mathcal{L}(\boldsymbol{\theta}_{\gamma n}))$ is the difference between the **continuous time process** and the **invariant measure**
- $W_2(\mathcal{L}(\boldsymbol{\theta}_{\gamma n}), \mathcal{L}(\theta_n^N))$ is the difference between the **discretisation** and the **continuous time process**

Rescaling Bounds

As mentioned before, we bound the **second** and **third** errors in (11) as

$$W_2(\pi_\Theta^N, \mathcal{L}(\boldsymbol{\theta}_{\gamma n})) \leq W_2(\pi_{*,z}^N, \mathcal{L}(\boldsymbol{Z}_{\gamma n})) \text{ (inv. measure/ cont. process)} \tag{12}$$

$$W_2(\mathcal{L}(\boldsymbol{\theta}_{\gamma n}), \mathcal{L}(\theta_n^N)) \leq W_2(\mathcal{L}(\boldsymbol{Z}_{\gamma n}), \mathcal{L}(Z_n^N)) \text{ (cont. process/ discr. process)} \tag{13}$$

where $\pi_{*,z}^N$ is the rescaled invariant measure.

Assumptions

Our assumptions:

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

$$\|\nabla U(v) - \nabla U(v')\| \leq L_\theta \|\theta - \theta'\| + L_x \|x - x'\|.$$

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

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

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

Main Convergence Result

Under these assumptions we obtain

$$\mathbb{E}\left[\|\theta_n - \theta^\star\|^2\right]^{1/2} \leq \sqrt{\frac{2d_\theta}{N\mu}} + e^{-\mu n\gamma}\left(\|z_0 - z^\star\| + \left(\frac{d_x N + d_\theta}{N\mu}\right)^{1/2}\right) + C(1 + d_\theta/N + d_x)\gamma^{1/2}. \tag{14}$$

where γ is the stepsize, N is the number of particles and C is a constant. However the three error bounds

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

have been adressed in different contexts under much **more general conditions**. The calculation of error bounds under weaker conditions will be a topic for future research.

Conclusions

To conclude:

1. we solve the problem of maximising

$$k(\theta) := \int_{\mathbb{R}^{d_x}} e^{-U(\theta,x)}dx \tag{15}$$

   by **sampling** from a distribution that concentrates around the maximiser θ^*
2. to do this we use the following IPS with **invariant measure** equal to such a distribution

$$d\boldsymbol{\theta}_t^N = -\frac{1}{N}\sum_{j=1}^{N}\nabla_{\boldsymbol{\theta}}U(\boldsymbol{\theta}_t^N,\boldsymbol{X}_t^{j,N})dt + \sqrt{\frac{2}{N}}d\boldsymbol{B}_t^{0,N}, \tag{16}$$

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

3. for theoretical purposes we use the **rescaling**

$$\boldsymbol{Z}_t^N = (\boldsymbol{\theta}_t^N, N^{-1/2}\boldsymbol{X}_t^{1,N}, ..., N^{-1/2}\boldsymbol{X}_t^{N,N})$$

Conclusions

- we draw from the EM, SA and IPS literature to obtain a scheme with nonasymptotic error bounds
- this particle-based method is scalable (Kuntz et al., 2023) and can be applied to nonconvex models too
- other discretisation schemes/sampling methods possible
- usual subsampling/minibatching

Thank you!

Bibliography I

Nicolas Brosse, Alain Durmus, Éric Moulines, and Sotirios Sabanis. The tamed unadjusted langevin algorithm. *Stochastic Processes and their Applications*, 129(10):3638–3663, 2019. ISSN 0304-4149. doi: https://doi.org/10.1016/j.spa.2018.10.002. URL https://www.sciencedirect.com/science/article/pii/S0304414918305635.

Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. In *4th International Conference on Learning Representations (ICLR)*, 2016.

Bradley P Carlin and Thomas A Louis. Empirical Bayes: Past, present and future. *Journal of the American Statistical Association*, 95(452):1286–1289, 2000.

Gilles Celeux and Jean Diebolt. A stochastic approximation type EM algorithm for the mixture problem. *Stochastics: An International Journal of Probability and Stochastic Processes*, 41(1-2):119–134, 1992.

NGOC HUY CHAU, Eric Moulines, Miklós Rásonyi, Sotirios Sabanis, and Ying Zhang. On stochastic gradient langevin dynamics with dependent data streams: The fully nonconvex case. *SIAM Journal on Mathematics of Data Science*, 3:959–986, 09 2021. doi: 10.1137/20M1355392.

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(3), May 2021. ISSN 0960-3174. doi: 10.1007/s11222-020-09986-y. Publisher Copyright: © 2021, The Author(s). Copyright: Copyright 2021 Elsevier B.V., All rights reserved.

Valentin De Bortoli, Alain Durmus, Marcelo Pereyra, and Ana F Vidal. Efficient stochastic optimisation by unadjusted Langevin Monte Carlo. *Statistics and Computing*, 31(3):1–18, 2021.

Bernard Delyon, Marc Lavielle, and Eric Moulines. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Statistics*, pages 94–128, 1999.

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 (Statistical Methodology)*, 39:2–38, 1977.

Arnaud Doucet, Simon J Godsill, and Christian P Robert. Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statistics and Computing*, 12(1):77–84, 2002.

Alain Durmus and Éric Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3): 1551 – 1587, 2017. doi: 10.1214/16-AAP1238. URL https://doi.org/10.1214/16-AAP1238.

Adam M Johansen, Arnaud Doucet, and Manuel Davy. Particle methods for maximum likelihood estimation in latent variable models. *Statistics and Computing*, 18(1):47–57, 2008.

Diederik P Kingma and Max Welling. Auto-encoding variational Bayes. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2013.

Juan Kuntz, Jen Ning Lim, and Adam M Johansen. Particle algorithms for maximum likelihood training of latent variable models. *AISTATS*, 2023.

Bibliography II

Chuanhai Liu and Donald B Rubin. The ECME algorithm: a simple extension of EM and ECM with faster monotone convergence. *Biometrika*, 81(4): 633–648, 1994.

Xiao-Li Meng and Donald B Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 1993.

Maxim Raginsky, Alexander Rakhlin, and Matus Telgarsky. Non-convex learning via stochastic gradient Langevin dynamics: a nonasymptotic analysis. In Satyen Kale and Ohad Shamir, editors, *Proceedings of the 2017 Conference on Learning Theory*, volume 65 of *Proceedings of Machine Learning Research*, pages 1674–1703. PMLR, 07–10 Jul 2017.

Robert P Sherman, Yu-Yun K Ho, and Siddhartha R Dalal. Conditions for convergence of Monte Carlo EM sequences with an application to product diffusion modeling. *The Econometrics Journal*, 2(2):248–267, 1999.

Greg CG Wei and Martin A Tanner. A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.