Gradient Flows for Statistical Computation

Trends and Trajectories

Sam Power, University of Bristol

“Advanced Langevin Methods for Bayesian Sampling”, BayesComp,
National University Of Singapore, 19 June 2025

Main ideas today

- Many statistical tasks reduce to solution of an optimisation problem
- Many common methods for these problems have ‘gradient’ structure
- Identifying these commonalities is useful for analysis, synthesis, progress

Collaborators

feel free to stop me at any point

Statistical Inference

Statistical Computing

Optimisation Problems

Three Main Characters

- Optimisation over Parameter Spaces ("$\Theta \subseteq \mathbf{R}^d$")
- Optimisation over Measure Spaces ("$\mathscr{P}\left(\mathscr{X}\right)$"; $\mathscr{X} \subseteq \mathbf{R}^d$)
- Optimisation over 'Hybrid' Spaces ("$\Theta \times \mathscr{P}\left(\mathscr{X}\right)$")

Optimisation over Parameter Spaces

- Maximum Likelihood Estimation ('MLE'): $\max_\theta \sum_{i \in [N]} \log p_\theta(y_i)$
  - maybe incorporate a penalty term ('penalised MLE')
  - maybe use a more general loss ('M-Estimation')
- Variational Approximation : $\min_\theta \mathrm{KL}\left(p_\theta, \pi\right)$
  - e.g. $\theta = (\mathrm{m}, \mathrm{C})$, $p_\theta(\mathrm{d}x) = \mathcal{N}\left(\mathrm{d}x; \mathrm{m}, \mathrm{C}\right)$; "best Gaussian fit"

Optimisation over Measure Spaces

- Sampling from an unnormalised distribution $\pi \propto \exp(-V)$

$$\min_{\mu} \mathrm{KL}(\mu, \pi) \sim \min_{\mu} \left\{ \mathbf{E}_{\mu}[V] + \mathscr{H}(\mu) \right\}$$

with $\mathscr{H}(\mu) = \int \left(\mu \log \mu - \mu \right)$ (special).

- (Nonparametric) Mean-Field Approximation

$$\min_{\mu_1, \cdots, \mu_d} \mathrm{KL}(\mu_1 \otimes \cdots \otimes \mu_d, \pi) \sim \min_{\mu} \left\{ \mathbf{E}_{\mu_1 \otimes \cdots \otimes \mu_d}[V] + \sum_{i \in [d]} \mathscr{H}(\mu_i) \right\}$$

- (integral probability metrics, information-theoretic divergences, etc.)

Optimisation over Hybrid Spaces

- Latent Variable Models: sample $\left[x \mid \theta, y\right]$, optimise $\left[\theta \mid x; y\right]$ (EM)
- Unnormalised Models: sample $\left[x \mid \theta\right]$, optimise θ (CD / MC-MLE)
- Distributed Inference: sample local posterior, tilt for consensus (~EP)
- Opinion: more prevalent than you might expect; worth taking seriously

press me on the examples

Optimisation by Local Search

Optimisation in Metric Spaces

Metrics

- Nothing too fancy - just want enough structure to ‘do good calculus’
- For parameter optimisation, $\Theta \subseteq \mathbf{R}^d$ can carry Euclidean metric.
- For measure optimisation, $\mathscr{P}\left(\mathscr{X}\right)$ can carry transport metric.
- For hybrid optimisation, $\Theta \times \mathscr{P}\left(\mathscr{X}\right)$ can carry ‘hybrid’ metric

$$\mathsf{d}_{\mathrm{hyb}}\left(\left(\theta,\mu\right),\left(\theta',\mu'\right)\right) = \sqrt{\|\theta-\theta'\|_2^2 + \mathscr{T}_2^2\left(\mu,\mu'\right)}$$

Conceptual Optimisation Framework

OPT $\quad \min_{x \in \mathcal{X}} f(x)$

PPM $\quad x_0 \mapsto \arg\min_{x \in \mathcal{X}} \left\{ f(x) + \frac{1}{2h} \cdot \mathsf{d}\left(x, x_0\right)^2 \right\}$

FLOW $\quad \dot{x}_t = -\nabla f\left(x_t\right)$

Specify a Metric Structure

Receive an Optimisation Algorithm

A Word on Gradients

- For parametric optimisation, the gradient is the ‘usual’ gradient
- For measure-valued optimisation, the gradient is a ‘Wasserstein’ gradient
- For hybrid optimisation, you take part of each, ‘as expected’

Wasserstein Gradients: Some Intuition

- Suppose that we are interested in a functional $\mathscr{F} : \mathscr{P}\left(\mathscr{X}\right) \to \mathbf{R}$
- Assume that it carries a Taylor expansion along the path $\mu \rightsquigarrow \mu'$

$$\frac{\mathscr{F}\left((1-t)\cdot\mu + t\cdot\mu'\right) - \mathscr{F}\left(\mu\right)}{t} \approx \int \left(\delta_\mu \mathscr{F}\right)\left(\mu, x\right) \cdot \left\{\mu'\left(\mathrm{d}x\right) - \mu\left(\mathrm{d}x\right)\right\}$$

- Natural to decrease $\mathscr{F}$ by pushing mass towards minima of $\left(\delta_\mu \mathscr{F}\right)\left(\mu, \cdot\right)$, i.e.

$$\dot{X}_t = -\nabla_x \delta_\mu \mathscr{F}\left(\mu_t, X_t\right)$$

Wasserstein Gradients: An Example

- For $\mathscr{V}\left(\mu\right) = \mathbf{E}_{\mu}\left[V\right]$, obtain

$$\delta_{\mu}\mathscr{V}\left(\mu, x\right) = V(x)$$

$$\nabla_{\mathscr{T}}\mathscr{V}\left(\mu, x\right) = \nabla V(x)$$

Wasserstein Gradients: And Another

- For $\mathscr{W}\left(\mu\right) = \frac{1}{2}\int \mu\left(\mathrm{d}x\right)\mu\left(\mathrm{d}y\right)W\left(x,y\right)$, obtain

$$\delta_{\mu}\mathscr{W}\left(\mu,x\right) = \int \mu\left(\mathrm{d}y\right)\cdot W\left(x,y\right)$$

$$\nabla_{\mathscr{T}}\mathscr{W}\left(\mu,x\right) = \int \mu\left(\mathrm{d}y\right)\cdot \nabla_{x}W\left(x,y\right)$$

A Special Case: Entropy

- A particularly special functional is the (shifted, negative) entropy

$$\mathscr{H}(\mu) = \int \left(\mu \log \mu - \mu\right)$$

- This satisfies $\delta_\mu \mathscr{H}\left(\mu, x\right) = \log \mu\left(x\right)$, so we could decrease it by evolving

$$\dot{X}_t = -\nabla_x \log \mu_t\left(X_t\right)$$

- Remarkably, the same path of measures is induced by instead evolving stochastically by

$$\mathrm{d}X_t = \sqrt{2}\mathrm{d}W_t$$

- “The gradient flow of the entropy can be realised by Brownian motion”

How do the Algorithms Look?

Gradient Flows on Parameter Spaces

- Task: $\min_{\theta \in \Theta} f(\theta)$
- ODE:

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

- Time-Discretised Method: “Gradient Method”

Gradient Flows on Measure Spaces (1)

- Task: $\min_{\mu \in \mathscr{P}(\mathscr{X})} \left\{ \mathscr{F}(\mu) + \mathscr{H}(\mu) \right\}$
- PDE:

$$\partial_t \mu_t = -\nabla_{\mathscr{T},\mu} \left\{ \mathscr{F} + \mathscr{H} \right\} (\mu_t)$$

Gradient Flows on Measure Spaces (2)

- Task: $\min_{\mu \in \mathscr{P}(\mathscr{X})} \left\{ \mathscr{F}(\mu) + \mathscr{H}(\mu) \right\}$

- Nonlinear ('McKean-Vlasov') SDE:

$$\mathrm{d}X_t = -\nabla_x \delta_\mu \mathscr{F}\left(\mu_t, X_t\right)\, \mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

Gradient Flows on Measure Spaces (3)

- Task: $\min_{\mu \in \mathscr{P}(\mathscr{X})} \left\{ \mathscr{F}(\mu) + \mathscr{H}(\mu) \right\}$
- Particle SDE:

$$\mathrm{d}X_t^i = -\nabla_x \delta_\mu \mathscr{F}\left(\widehat{\mu}_t^N, X_t^i\right) \mathrm{d}t + \sqrt{2}\mathrm{d}W_t^i, \qquad 1 \le i \le N$$

Gradient Flows on Measure Spaces (3)

- Task: $\min_{\mu \in \mathscr{P}(\mathscr{X})} \left\{ \mathscr{F}(\mu) + \mathscr{H}(\mu) \right\}$
- Particle SDE:

$$\mathrm{d}X_t^i = -\nabla_x \delta_\mu \mathscr{F}\left(\widehat{\mu}_t^N, X_t^i\right) \mathrm{d}t + \sqrt{2}\mathrm{d}W_t^i, \qquad 1 \leq i \leq N$$

- **Space-**Time-Discretised Method: (Mean-Field) “Langevin Monte Carlo”

Gradient Flows on Hybrid Spaces (1)

- Task: $\min_{\theta \in \Theta, \mu \in \mathscr{P}(\mathscr{X})} \left\{ \mathscr{F}(\theta, \mu) + \mathscr{H}(\mu) \right\}$

- ODE-PDE:

$$\dot{\theta}_t = -\nabla_\theta \mathscr{F}(\theta_t, \mu_t)$$

$$\partial_t \mu_t = -\nabla_{\mathscr{T}, \mu} \left\{ \mathscr{F} + \mathscr{H} \right\}(\theta_t, \mu_t)$$

Gradient Flows on Hybrid Spaces (2)

- Task: $\min_{\theta\in\Theta,\mu\in\mathscr{P}(\mathscr{X})} \left\{ \mathscr{F}\left(\theta,\mu\right) + \mathscr{H}\left(\mu\right) \right\}$

- ODE-MVSDE:

$$\dot{\theta}_t = -\nabla_\theta \mathscr{F}\left(\theta_t,\mu_t\right)$$

$$\mathrm{d}X_t = -\nabla_x \delta_\mu \mathscr{F}\left(\theta_t,\mu_t,X_t\right)\ \mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

$$\mu_t = \mathsf{Law}\left(X_t\right)$$

Gradient Flows on Hybrid Spaces (3)

- Task: $\min_{\theta\in\Theta,\mu\in\mathscr{P}(\mathscr{X})} \left\{ \mathscr{F}\left(\theta,\mu\right) + \mathscr{H}\left(\mu\right) \right\}$

- ODE-pSDE:

$$\dot{\theta}_t = -\nabla_\theta \mathscr{F}\left(\theta_t, \widehat{\mu}_t^N\right)$$

$$\mathrm{d}X_t^i = -\nabla_x \delta_\mu \mathscr{F}\left(\theta_t, \widehat{\mu}_t^N, X_t^i\right)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t^i$$

$$\widehat{\mu}_t^N = \mathsf{Emp\_Meas}\left(\left\{X_t^1, \cdots, X_t^N\right\}\right)$$

Gradient Flows on Hybrid Spaces (4)

- Task: $\min_{\theta\in\Theta,\mu\in\mathscr{P}(\mathscr{X})} \left\{ \mathscr{F}\left(\theta,\mu\right) + \mathscr{H}\left(\mu\right) \right\}$

- ODE-pSDE:

$$\dot{\theta}_t = -\nabla_\theta \mathscr{F}\left(\theta_t, \widehat{\mu}_t^N\right)$$

$$\mathrm{d}X_t^i = -\nabla_x \delta_\mu \mathscr{F}\left(\theta_t, \widehat{\mu}_t^N, X_t^i\right)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t^i$$

$$\widehat{\mu}_t^N = \mathsf{Emp\_Meas}\left(\left\{X_t^1, \cdots, X_t^N\right\}\right)$$

- Space-Time-Discretised Method: “Particle Gradient Descent”

Some Words on Theory

- In each case, the theoretical picture is very clear for “convex” problems
- In each case, there exists a ‘robust’ notion of convexity / connectedness which yields guarantees for a larger class of problems
- These notions are …
  - quite well-developed on Θ,
  - very well-developed on $\mathscr{P}\left(\mathscr{X}\right)$, and
  - still under development for hybrid spaces

Some Words on Extensions

- Today, I have really spoken about ‘standard’ methods in this area
- In many applications, the ‘standard’ gradient flow is sub-optimal.
  - This is true in both continuous and in discrete time.
- Some intuition has developed for how ‘optimal’ improvements look.
- A common (though not universal) theme seems to involve ‘momentum’.

Momentum-Enriched Dynamics

- A common theme involves ‘momentum’.
- Some pattern-matching indicates how to introduce this momentum.
  - Introduce a momentum and a suitable extended objective functional
  - Instead of the gradient flow, work with the *Conformal Hamiltonian Flow*
- My discussion here is brief and informal, but can be made formal.

Momentum-Enriched Algorithms

- Some pattern-matching indicates how to introduce this momentum.
  - For parameter optimisation, recover *Nesterov's Fast Gradient Method*
  - For measure optimisation, recover *Kinetic Langevin Monte Carlo*
  - For hybrid case, we propose *Momentum Particle Gradient Descent*
- Promising mixture of theoretical and practical benefits

Some Take-Aways

- Optimisation problems are widespread in statistical tasks
  - … and often involve more than ‘just’ fixed-dimensional parameters.
- It is often possible to solve such problems “with gradient descent”
  - … and we can even systematically concoct improvements on GD.
- Identifying these commonalities is useful for analysis, synthesis, progress
  - … and many interesting questions still remain.

Particle algorithms for maximum likelihood training of latent variable models

Juan Kuntz **Jen Ning Lim** **Adam M. Johansen**
Department of Statistics, University of Warwick.

Abstract

Neal and Hinton (1998) recast maximum likelihood estimation of any given latent variable model as the minimization of a free energy functional F, and the EM algorithm as coordinate descent applied to F. Here, we explore alternative ways to optimize the functional. In particular, we identify various gradient flows associated with F and show that their limits coincide with F's stationary points. By discretizing the flows, we obtain practical particle-based algorithms for maximum likelihood estimation in broad classes of latent variable models. The novel algorithms scale to high-dimensional settings and perform well in numerical experiments.

(S2) obtain the corresponding *posterior distribution*,

$$p_{\theta_*}(x|y) := \frac{p_{\theta_*}(x,y)}{p_{\theta_*}(y)}.$$

Perhaps the most well-known method for tackling (S1,2) is the *expectation maximization* (EM) algorithm (Dempster et al., 1977): starting from an initial guess θ_0, alternate,

(E) compute $q_k := p_{\theta_k}(\cdot|y)$,
(M) solve for $\theta_{k+1} := \arg\max_{\theta\in\Theta} \int \ell(\theta, x) q_{k+1}(x) dx$,

where $\ell(\theta, x) := \log(p_\theta(x, y))$ denotes the log-likelihood. Under general conditions (McLachlan, 2007, Chap. 3), θ_k converges to a stationary point θ_* of the marginal likelihood and q_k to the corresponding posterior $p_{\theta_*}(\cdot|y)$. In cases where the above steps are not analytically tractable, it is common to approximate (E) using Monte Carlo (or Markov chain Monte Carlo if $p_\theta(\cdot|y)$ cannot be sampled

AISTATS 2023

Momentum Particle Maximum Likelihood

Jen Ning Lim [1] **Juan Kuntz** [2] **Samuel Power** [3] **Adam M. Johansen** [1]

Abstract

Maximum likelihood estimation (MLE) of latent variable models is often recast as the minimization of a free energy functional over an extended space of parameters and probability distributions. This perspective was recently combined with insights from optimal transport to obtain novel particle-based algorithms for fitting latent variable models to data. Drawing inspiration from prior works which interpret 'momentum-enriched' optimization algorithms as discretizations of ordinary differential equations, we propose an analogous dynamical-systems-inspired approach to minimizing the free energy functional. The result is a dynamical system that blends elements of Nesterov's Accelerated Gradient method, the underdamped Langevin diffusion, and particle methods. Under suitable assumptions, we prove that the continuous-time system minimizes the functional. By discretizing the system, we obtain a practical algorithm for MLE in latent variable models. The algorithm outperforms existing particle methods in numerical experiments and compares favourably with other MLE algorithms.

by constructing an objective defined over an extended space, whose optima are in one-to-one correspondence with those of the MLE problem. To this end, we define the 'free energy' functional:

$$\mathcal{E}(\theta, q) := \int \log\left(\frac{q(x)}{p_\theta(y, x)}\right) q(x)\, \mathrm{d}x. \tag{1}$$

The infimum of $\mathcal{E}$ over the *extended space* $\mathbb{R}^{d_\theta} \times \mathcal{P}(\mathbb{R}^{d_x})$ (where $\mathcal{P}(\mathbb{R}^{d_x})$ denotes the space of probability distributions over $\mathbb{R}^{d_x}$) coincides with negative of the log marginal likelihood's supremum:

$$\mathcal{E}^* := \inf_{(\theta,q)} \mathcal{E}(\theta, q) = -\sup_{\theta \in \mathbb{R}^{d_\theta}} \log p_\theta(y). \tag{2}$$

To see this, note that

$$\mathcal{E}(\theta, q) = -\log p_\theta(y) + \mathsf{KL}(q, p_\theta(\cdot \mid y)), \tag{3}$$

where KL denotes the Kullback–Leibler divergence and $p_\theta(\cdot|y) := p_\theta(y, \cdot)/p_\theta(y)$ the posterior distribution. So, if (θ^*, q^*) minimizes the free energy, then θ^* maximizes the marginal likelihood and $q^*(\cdot) = p_{\theta^*}(\cdot|y)$. In other words, by minimizing the free energy, we solve our MLE problem.

This perspective motivates the search for practical procedures that minimize the free energy $\mathcal{E}$. One such example

ICML 2024

Error bounds for particle gradient descent, and extensions of the log-Sobolev and Talagrand inequalities

Rocco Caprio[†] Juan Kuntz[‡] Samuel Power[§] Adam M. Johansen[†]

April 12, 2024

Abstract

We prove non-asymptotic error bounds for particle gradient descent (PGD) [26], a recently introduced algorithm for maximum likelihood estimation of large latent variable models obtained by discretizing a gradient flow of the free energy. We begin by showing that, for models satisfying a condition generalizing both the log-Sobolev and the Polyak–Łojasiewicz inequalities (LSI and PŁI, respectively), the flow converges exponentially fast to the set of minimizers of the free energy. We achieve this by extending a result well-known in the optimal transport literature (that the LSI implies the Talagrand inequality) and its counterpart in the optimization literature (that the PŁI implies the so-called quadratic growth condition), and applying it to our new setting. We also generalize the Bakry–Émery Theorem and show that the LSI/PŁI generalization holds for models with strongly concave log-likelihoods. For such models, we further control PGD's discretization error, obtaining non-asymptotic error bounds. While we are motivated by the study of PGD, we believe that the inequalities and results we extend may be of independent interest.

Keywords: latent variable models, maximum marginal likelihood, gradient flows, log-Sobolev inequality, Polyak–Łojasiewicz inequality, Talagrand inequality, quadratic growth condition.

Error bounds for particle gradient descent, and extensions of the log-Sobolev and Talagrand inequalities

Rocco Caprio[†] Juan Kuntz[‡] Samuel Power[§] Adam M. Johansen[†]

April 12, 2024

Abstract

We prove non-asymptotic error bounds for particle gradient descent (PGD) [26], a recently introduced algorithm for maximum likelihood estimation of large latent variable models obtained by discretizing a gradient flow of the free energy. We begin by showing that, for models satisfying a condition generalizing both the log-Sobolev and the Polyak–Łojasiewicz inequalities (LSI and PŁI, respectively), the flow converges exponentially fast to the set of minimizers of the free energy. We achieve this by extending a result well-known in the optimal transport literature (that the LSI implies the Talagrand inequality) and its counterpart in the optimization literature (that the PŁI implies the so-called quadratic growth condition), and applying it to our new setting. We also generalize the Bakry–Émery Theorem and show that the LSI/PŁI generalization holds for models with strongly concave log-likelihoods. For such models, we further control PGD's discretization error, obtaining non-asymptotic error bounds. While we are motivated by the study of PGD, we believe that the inequalities and results we extend may be of independent interest.

Keywords: latent variable models, maximum marginal likelihood, gradient flows, log-Sobolev inequality, Polyak–Łojasiewicz inequality, Talagrand inequality, quadratic growth condition.

Journal of Machine Learning Research, in press,

Some Take-Aways

- Optimisation problems are widespread in statistical tasks
  - … and often involve more than ‘just’ fixed-dimensional parameters.
- It is often possible to solve such problems “with gradient descent”
  - … and we can even systematically concoct improvements on GD.
- Identifying these commonalities is useful for analysis, synthesis, progress
  - … and many interesting questions still remain.