Gradient Flows for Statistical Computation

Trends and Trajectories

Sam Power, University of Bristol

Probability and Statistics Seminar Series, Università di Pavia, 4 June 2025

Some background

- Lecturer in Statistical Science $\subseteq$ School of Maths @ Bristol
- Trained as a Mathematician (MMath @ Oxford + PhD @ Cambridge)
- Most often thinking about topics related to algorithms for statistics:
  - { Simulation, Particle Methods, Estimation, Optimisation, … }
- Interested in analysis and synthesis of algorithms; applications

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

Game Plan

- Describe a diverse variety of relevant statistical optimisation tasks
- Describe a consistent framework for solving them computationally
- Identify some ‘standard’ methods which come from this framework
  - … and explain how some extensions can be derived
- Identify some open questions arising from these new methods

Collaborators

feel free to stop me at any point

ask me about references

Examples of Statistical 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 add 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. $\theta = (m, C)$, $p_\theta(\mathrm{d}x) = \mathcal{N}(\mathrm{d}x; m, C)$; “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 $\mathscr{H}(\mu) = \int \mu \log \mu - \mu.$

- (Nonparametric) Mean-Field Approximation

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

- ‘Quadratic Free Energy Minimisation’

$$\min_{\mu} \left\{ \mathbf{E}_{\mu}[V] + \frac{1}{2} \mathbf{E}_{\mu \otimes \mu}[W] + \mathscr{H}(\mu) \right\}$$

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

Optimisation over Hybrid Spaces

- Basic Example: Deconvolution
  - Model: draw $X \sim p_\theta$, but only *observe* $Y \sim \mathcal{N}\left(X, \sigma^2\right)$
  - In principle, can ‘just’ do MLE …
    - … but here, $p_\theta(y),\ \nabla_\theta \log p_\theta(y)$ are likely unavailable
  - Coupled problem: impute $\left[x \mid \theta, y\right]$, optimise $\left[\theta \mid x; y\right]$
  - More generally: “EM Algorithm”, “Latent Variable Models”

Optimisation over Hybrid Spaces

- { ‘Energy-Based’ / ‘Unnormalised’ / ‘Pre-Normalised’ } Models
  - Specify $p_\theta(y) \propto \exp\left(-V(y;\theta)\right)$; leave $Z(\theta)$ defined implicitly
  - In principle, can ‘just’ do MLE …
    - … but here, $p_\theta(y)$, $\nabla_\theta \log p_\theta(y)$ are likely unavailable
  - Coupled problem: Sample $x \sim p_\theta$, then optimise θ based on x, y
    - “Contrastive Divergence”, “MC-MLE”

Additional Comments on Hybrid Spaces

- Increasingly, clear that many problems have this two-scale structure
  - Adaptive MCMC (sample from π, optimise parameters of dynamics)
  - Distributed Inference (sample ‘locally’, ‘tilt parameters’ for consensus)
  - See also “MCMC-Driven Learning” chapter by Bouchard-Côté+++
    - “Markovian Optimisation-Integration” framework
- IMO: Worthy of serious consideration; not just hypothetical / edge case.

last chance to ask about examples

Metric Structures in Statistical Optimisation

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 Kantorovich metric.
- For hybrid optimisation, $\Theta \times \mathscr{P}\left(\mathscr{X}\right)$ can carry ‘hybrid’ metric

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

Optimisation on Metric Spaces

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

From Gradient Flows to Algorithms

Gradient Flows on Parameter Spaces

- Task: $\min_{\theta \in \Theta} f(\theta)$
- Continuous Dynamics: $\dot{\theta}_t = -\nabla_\theta f\left(\theta_t\right)$
- Time-Discretised Method: “Gradient Method”

Gradient Flows on Parameter Spaces

- Time-Discretised Method: “Gradient Method”
- Extremely well-understood when f is uniformly-convex
- Further theory for ‘gradient-dominated’ f; ‘Polyak-Łojasiewicz Inequality’

Gradient Flows on Measure Spaces

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

- Continuous Dynamics: (abstract gradient form)

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

Gradient Flows on Measure Spaces

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

- Continuous Dynamics: (PDE form)

$$\partial_t \mu_t = \mathrm{div}_x \left(\mu_t \nabla_x \delta_\mu \mathscr{F}(\mu_t) \right) + \Delta_x \mu_t$$

Gradient Flows on Measure Spaces

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

- Continuous Dynamics: (McKean-Vlasov form, nonlinear)

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

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

Gradient Flows on Measure Spaces

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

- Continuous Dynamics: (McKean-Vlasov form, particle)

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

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

Gradient Flows on Measure Spaces

- **Space-**Time-Discretised Method: (Mean-Field) “Langevin Monte Carlo”
- Extremely well-understood when $\mathscr{F}$ is uniformly-geodesically-convex
- Further theory for ‘well-connected’ $\mathscr{F}$; “Sobolev-type” inequalities

Gradient Flows on Hybrid Spaces

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

- Continuous Dynamics: (abstract gradient form)

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

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

Gradient Flows on Hybrid Spaces

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

- Continuous Dynamics: (coupled ODE-PDE form)

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

$$\partial_t \mu_t = \mathrm{div}_x\left(\mu_t \nabla_x \delta_\mu \mathscr{F}\left(\theta_t,\mu_t\right)\right) + \Delta_x \mu_t$$

Gradient Flows on Hybrid Spaces

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

- Continuous Dynamics: (McKean-Vlasov form, nonlinear)

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

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

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

Gradient Flows on Hybrid Spaces

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

- Continuous Dynamics: (McKean-Vlasov form, particle)

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

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

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

Gradient Flows on Hybrid Spaces

- Space-Time-Discretised Method: “Particle Gradient Descent”
- Very well-understood when $\mathscr{F}$ is uniformly-geodesically-convex:
  - PoC, versions of { BÉ, LSI, TH, HWI, … }
- Further theory currently missing; Open Questions

questions about ‘core’ methods?

Some New Directions

Beyond ‘Standard’ Gradient Flows

- 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’.
  - “lifting the problem to the cotangent bundle”

Beyond ‘Standard’ Gradient Flows

- We start off with some notion of an objective function.
- We then identify dynamics which can minimise that function for us.
- How shall this pipeline change when ‘momentum’ enters the picture?
- New objective function, new dynamics.
- Minimisers of new objective map onto minimisers of old objective.

Enriched Objective Functions

- For parameter optimisation, consider

$$\min_{(\theta,\varphi)\in\mathscr{T}^{\star}\Theta}\left\{h\left(\theta,\varphi\right):=f\left(\theta\right)+\frac{1}{2}\cdot\left\|\varphi\right\|_{2}^{2}\right\}$$

Enriched Objective Functions

- For measure optimisation, consider

$$\min_{\nu \in \mathscr{P}\left(\mathscr{T}^{\star}\mathscr{X}\right)} \left\{ H(\nu) := \mathscr{F}(\nu) + \mathscr{H}(\nu) + \mathbf{E}_{\nu}\left[\frac{1}{2} \cdot \|P\|^2\right] \right\}$$

Enriched Objective Functions

- For hybrid optimisation, consider

$$\min_{(\theta,\varphi),\nu} \left\{ \mathsf{H}\left(\theta,\varphi,\nu\right) := \mathscr{F}\left(\theta,\nu\right) + \mathscr{H}\left(\nu\right) + \frac{1}{2}\cdot\|\varphi\|_2^2 + \mathbf{E}_\nu\left[\frac{1}{2}\cdot\|P\|^2\right] \right\}$$

Enriched Objective Functions

$$\min_{(\theta,\varphi)\in\mathscr{T}^\star\Theta}\left\{h\left(\theta,\varphi\right):=f(\theta)+\frac{1}{2}\cdot\|\varphi\|_2^2\right\}$$

$$\min_{\nu\in\mathscr{P}\left(\mathscr{T}^\star\mathscr{X}\right)}\left\{H(\nu):=\mathscr{F}(\nu)+\mathscr{H}(\nu)+\mathbf{E}_\nu\left[\frac{1}{2}\cdot\|P\|^2\right]\right\}$$

$$\min_{(\theta,\varphi),\nu}\left\{\mathsf{H}\left(\theta,\varphi,\nu\right):=\mathscr{F}(\theta,\nu)+\mathscr{H}(\nu)+\frac{1}{2}\cdot\|\varphi\|_2^2+\mathbf{E}_\nu\left[\frac{1}{2}\cdot\|P\|^2\right]\right\}$$

- N.B. These choices are not “automatic” / “canonical”, but appear to make sense in many examples.

Warm-Up: Hamiltonian Flows

- An odd idea in isolation: conserve the ‘Hamiltonian’
- Introduce skew-symmetric matrix

$$\mathbf{J} = \begin{pmatrix} 0 & \mathbf{I} \\ -\mathbf{I} & 0 \end{pmatrix}$$

- In abstract terms: instead of

$$\dot{x} = -\nabla f(x),$$

- take $z = (x, p)$ and do

$$\dot{z} = \mathbf{J}\nabla H(z)$$

Hamiltonian Flows in Action

- For parameter optimisation, obtain

$$\dot{\theta}_t = \varphi_t, \qquad \dot{\varphi}_t = -\nabla f\left(\theta_t\right)$$

Hamiltonian Flows in Action

- For measure optimisation, obtain (omitting entropy term)

$$\mathrm{d}X_t = P_t\,\mathrm{d}t, \qquad \mathrm{d}P_t = -\nabla_x \delta_\mu \mathcal{F}\left(\mu_t, X_t\right)\,\mathrm{d}t$$

with ‘expected’ modifications for particle version

Hamiltonian Flows in Action

- For hybrid optimisation, obtain

$$\dot{\theta}_t = \varphi_t, \qquad \dot{\varphi}_t = -\nabla_\theta \mathcal{F}\left(\theta_t, \mu_t\right)$$

$$\mathrm{d}X_t = P_t\,\mathrm{d}t, \qquad \mathrm{d}P_t = \nabla_x \delta_\mu \mathcal{F}\left(\theta_t, \mu_t, X_t\right)\,\mathrm{d}t$$

with ‘expected’ modifications for particle version

okay, but why bother with this?

Conformal Hamiltonian Flows

- A recurrent phenomenon: it can be interesting to blend
  - Hamiltonian circulation, and
  - gradient-type damping *only on the momentum term*
- With some consistency, this appears to yield improved methods

Conformal Hamiltonian Flows

- It can be interesting to blend Hamiltonian circulation with gradient-type damping *only on the momentum term*

- The key matrix is then (for some $\gamma > 0$)

$$\mathsf{D}_\gamma = \begin{pmatrix} 0 & -\mathbf{I} \\ \mathbf{I} & \gamma \cdot \mathbf{I} \end{pmatrix}$$

and we will (formally) construct dynamics according to

$$\dot{z} = -\mathsf{D}_\gamma \nabla H(z)$$

Conformal Hamiltonian Flows in Action

- For parameter optimisation, obtain

$$\dot{\theta}_t = \varphi_t, \qquad \dot{\varphi}_t = -\nabla f\left(\theta_t\right) - \gamma \cdot \varphi_t$$

  - $\approx$ Nesterov's "Fast Gradient Method", rate-optimal for convex minimisation

Conformal Hamiltonian Flows in Action

- For measure optimisation, obtain

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

with ‘expected’ modifications for particle version

  - $\approx$ (Kinetic, Underdamped, …) Langevin Monte Carlo, improving upon LMC in many cases, “plausibly” optimal

Conformal Hamiltonian Flows in Action

- For hybrid optimisation, obtain

$$\dot{\theta}_t = \varphi_t, \qquad \dot{\varphi}_t = -\nabla_\theta \mathcal{F}\left(\theta_t, \mu_t\right) - \gamma \cdot \varphi_t$$

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

with ‘expected’ modifications for particle version

  - $\approx$ our “Momentum Particle Gradient Descent”, which empirically outperforms the original PGD; some supporting theory

Recap and Open Questions

Main ideas today

- 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.

Some Open Questions

- For optimisation problems on hybrid spaces,
  - Can we strengthen the theory outside of the uniformly-convex case?
  - Can we develop good principles for numerical discretisation?
  - What more shall be learned from “pure” optimisation and sampling?
- For momentum-enrichment,
  - How should we systematically construct ‘enriched’ objective functions?

Some Further Questions

- In general,
  - Can other practical tasks be fruitfully interpreted through optimisation?
  - Should we ever look *beyond* gradient and conformal Hamiltonian flows?

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

Momentum Particle Maximum Likelihood

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

[1]University of Warwick
[2]Polygeist
[3]University of Bristol

December 13, 2023

Abstract

Maximum likelihood estimation (MLE) of latent variable models is often recast as an optimization problem over the extended space of parameters and probability distributions. For example, the Expectation Maximization (EM) algorithm can be interpreted as coordinate descent applied to a suitable free energy functional over this space. Recently, this perspective has been combined with insights from optimal transport and Wasserstein gradient flows to develop particle-based algorithms applicable to wider classes of models than standard EM.

Drawing inspiration from prior works which interpret 'momentum-enriched' optimisation algorithms as discretizations of ordinary differential equations, we propose an analogous dynamical systems-inspired approach to minimizing the free energy functional over the extended space of parameters and probability distributions. The result is a dynamic system that blends elements of Nesterov's Accelerated Gradient method, the underdamped Langevin diffusion, and particle methods.

Under suitable assumptions, we establish quantitative convergence of the proposed system to the unique minimiser of the functional in continuous time. We then propose a numerical discretization of this system which enables its application to parameter estimation in latent variable models. Through numerical experiments, we demonstrate that the resulting algorithm converges faster than existing methods and compares favourably with other (approximate) MLE algorithms.

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.