Gradient Flows for Statistical Computation

Trends and Trajectories

Sam Power, University of Bristol

London Symposium on Information Theory,
University of Cambridge, 15 May 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

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

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

- (Nonparametric) Mean-Field Approximation

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

- (other objectives involving integral probability metrics, density ratio 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”

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

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

Optimisation on Metric Spaces

Conceptual Optimisation Framework

1. Abstract Optimisation Problem

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

2. Proximal Point Update

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

3. Gradient Flow

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

4. (Discretisation)

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”
- Extremely well-understood for uniformly-convex f
- Further theory when $f - \inf f \lesssim \|\nabla f\|^2$; ‘Polyak-Łojasiewicz Inequality’

Gradients on Measure Spaces?

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

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

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

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

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”

Gradient Flows on Measure Spaces

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

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

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

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

Gradient Flows on Hybrid Spaces

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

- Continuous Dynamics:

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

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

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

- Very well-understood for uniformly-geodesically-convex $\mathscr{F}$

- Further theory currently missing; Open Questions

questions on the ‘basic’ 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”

Enriched Objective Functions

- For parameter optimisation, consider

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

- For measure optimisation, consider

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

More Enriched Objective Functions

- For hybrid optimisation, consider

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

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

Warm-Up: Hamiltonian Flows

- A sketchy 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

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

- For measure optimisation, obtain (omitting entropy term)

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

- For hybrid optimisation, obtain

$$\dot{\theta}_t = \varphi_t, \qquad \dot{\varphi}_t = -\nabla_\theta \mathscr{F}\left(\theta_t, \mu_t\right)$$
$$\mathrm{d}X_t = P_t\,\mathrm{d}t, \qquad \mathrm{d}P_t = \nabla_x \delta_\mu \mathscr{F}\left(\theta_t, \mu_t, X_t\right)\,\mathrm{d}t$$

- But … *why bother*?

Conformal Hamiltonian Flows

- A recurrent phenomenon: it can be interesting to blend Hamiltonian circulation with gradient-type damping *only on the momentum term*

- With some consistency, this appears to yield improved (and even optimal) methods

- 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

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

  - ≈ Nesterov's "Fast Gradient Method", rate-optimal for convex minimisation

- For measure optimisation, obtain

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

  - ≈ (Kinetic, Underdamped, …) Langevin Monte Carlo, improving upon LMC in many cases, "plausibly" optimal

- For hybrid optimisation, obtain

$$\begin{aligned} \dot{\theta}_t &= \varphi_t, & \dot{\varphi}_t &= -\nabla_\theta \mathscr{F}\left(\theta_t, \mu_t\right) - \gamma \cdot \varphi_t \\ \mathrm{d}X_t &= P_t\,\mathrm{d}t, & \mathrm{d}P_t &= -\nabla_x \delta_\mu \mathscr{F}\left(\theta_t, \mu_t, X_t\right)\,\mathrm{d}t - \gamma \cdot P_t\,\mathrm{d}t + \sqrt{2 \cdot \gamma}\,\mathrm{d}W_t \end{aligned}$$

  - ≈ our "Momentum Particle Gradient Descent", which empirically outperforms the original PGD

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,
  - Which other practical tasks can be fruitfully interpreted as optimisation?
  - Should we ever look *beyond* gradient and conformal Hamiltonian flows?

The Papers

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.

1. Introduction

In this work, we study parameter estimation for (probabilistic) latent variable models $p_\theta(y, x)$ with parameters $\theta \in \mathbb{R}^{d_\theta}$, unobserved (or latent) variables $x \in \mathbb{R}^{d_x}$, and observed variables $y \in \mathbb{R}^{d_y}$ (which we treat as fixed throughout). The type II maximum likelihood approach (Good, 1983) estimates θ by maximizing the *marginal likelihood* $p_\theta(y) := \int p_\theta(y, x)\,\mathrm{d}x$. However, for most models of practical interest, the integral has no known closed-form expressions and we are unable to optimize the marginal likelihood directly.

This issue is often overcome (e.g., Neal and Hinton (1998)) 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

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

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 is the classical Expectation Maximization (EM) algorithm applicable to models for which the posterior distributions $p_\theta(\cdot|y)$ are available in closed form. As shown in Neal and Hinton (1998), its iterates coincide with those of coordinate descent applied to the free energy $\mathcal{E}$. Recently, Kuntz et al. (2023) sought analogues of gradient descent (GD) applicable to the free energy $\mathcal{E}$. In particular, building on ideas popular in optimal transport (e.g., see Ambrosio et al. (2005)), they identified an appropriate notion for the free energy's gradient $\nabla\mathcal{E}$, discretized the corresponding gradient flow $(\dot\theta_t, \dot q_t) = -\nabla\mathcal{E}(\theta_t, q_t)$, and obtained a practical algorithm they called particle gradient descent (PGD).

In optimization, GD is well-known to be suboptimal: other practical first-order methods achieve better worst-case convergence guarantees and practical performance (Nemirovskij and Yudin, 1983; Nesterov, 1983). A common feature among algorithms that achieve the optimal 'accelerated' convergence rates is the presence of 'momentum' effects in the dynamics of the algorithm. Roughly speaking,

[1]University of Warwick [2]Polygeist [3]University of Bristol. Correspondence to: Jen Ning Lim <Jen-Ning.Lim@warwick.ac.uk>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 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.