Gradient Flows for Statistical Computation

Trends and Trajectories

Sam Power, University of Bristol

Workshop on ‘Large scale behaviour of interacting diffusions’, 18 Sep 2024

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

$$\boxed{\text{feel free to stop me at any point}}$$

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

$$\min_{\theta} \mathsf{KL}\left(p_{\theta}, \pi\right)$$

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

- (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"

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.

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

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

- For the ambitious: { Riemannian, (Kernel) Stein, (Ensemble) Kalman, $\cdots$ }

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 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:

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

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

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

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

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

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

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

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

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

- For measure optimisation, consider

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

- For hybrid optimisation, consider

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

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

Warm-Up: Hamiltonian Flows

- 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

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

- 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 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

- For measure optimisation, obtain

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

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

- For hybrid optimisation, obtain

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

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