Gradient Flows for Statistical Computation

Trends and Trajectories

Sam Power, University of Bristol

Post-Bayes Workshop, University College London, 16 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
- Post-Bayes as a rich source of motivating applications

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 : $\min_{\theta} \mathsf{KL}\left(p_{\theta}, \pi\right)$
  - e.g. $\theta = (\mathsf{m}, \mathsf{C})$, $p_{\theta}(\mathrm{d}x) = \mathcal{N}\left(\mathrm{d}x; \mathsf{m}, \mathsf{C}\right)$; "best Gaussian fit"

Optimisation over Measure Spaces

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

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

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

- (Nonparametric) Mean-Field Approximation

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

- (integral probability metrics, information-theoretic divergences, etc. - see Zheyang's talk!)

Optimisation over Hybrid Spaces

- Latent Variable Models: impute $\left[x \mid \theta, y\right]$, 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)
- <u>Opinion</u>: more prevalent than you might expect; worth taking seriously

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

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

Conceptual Optimisation Framework

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

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

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

Specify a Metric Structure

Receive an Optimisation Algorithm

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

- Task: $\min_{\mu \in \mathcal{P}(\mathcal{X})} \left\{ \mathcal{F}\left(\mu\right) + \mathcal{H}\left(\mu\right) \right\}$
- PDE: $\partial_t \mu_t = -\nabla_{\mathcal{T},\mu} \left\{ \mathcal{F} + \mathcal{H} \right\} \left(\mu_t\right)$
- SDE: $\mathrm{d}X_t = -\nabla_x \delta_\mu \mathcal{F}\left(\mu_t, X_t\right)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t$
- **Space-**Time-Discretised Method: (Mean-Field) “Langevin Monte Carlo”

Gradient Flows on Hybrid Spaces

- Task: $\min_{\theta \in \Theta, \mu \in \mathscr{P}(\mathscr{X})} \left\{ \mathscr{F}\left(\theta, \mu\right) + \mathscr{H}\left(\mu\right) \right\}$
- ODE-PDE: $\dot{\theta}_t = -\nabla_\theta \mathscr{F}\left(\theta_t, \mu_t\right), \partial_t \mu_t = -\nabla_{\mathscr{T},\mu} \left\{ \mathscr{F} + \mathscr{H} \right\} \left(\theta_t, \mu_t\right)$
- ODE-SDE: $\dot{\theta}_t = -\nabla_\theta \mathscr{F}\left(\theta_t, \mu_t\right), \mathrm{d}X_t = -\nabla_x \delta_\mu \mathscr{F}\left(\theta_t, \mu_t, X_t\right) \mathrm{d}t + \sqrt{2}\mathrm{d}W_t$
- Space-Time-Discretised Method: “Particle Gradient Descent”

A Word 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 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.