Gradient Flows for Statistical Computation

Trends and Trajectories

Sam Power, University of Bristol

Statistics Seminar,
University of Edinburgh, 16 November 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

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’)

Optimisation over Parameter Spaces

- Variational Approximation:

  with π known up to a normalising constant,

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

  - 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 $\pi \propto \exp(-V)$

$$\min_{\mu} \mathrm{KL}(\mu, \pi)$$

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

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

Optimisation over Measure Spaces

- (Nonparametric) Mean-Field Approximation

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

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

Optimisation over Measure Spaces

- ‘Quadratic Free Energy Minimisation’

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

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

  - think of e.g. { covariance regularisation, kernel methods, … }

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, Distributed Inference, … (ask me!)
  - see also “MCMC-Driven Learning” chapter by Bouchard-Côté+++
- IMO: worthy of serious attention

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, $\Theta \subseteq \mathbf{R}^d$ with Euclidean metric.
- For measure optimisation, $\mathscr{P}\left(\mathscr{X}\right)$ with 2-Wasserstein metric.
- For hybrid optimisation, $\Theta \times \mathscr{P}\left(\mathscr{X}\right)$ with ‘hybrid’ metric

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

Optimisation on Metric Spaces

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

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 a Taylor expansion along $\mu \rightsquigarrow \mu'$

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

- 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

- Recall

$$\mathscr{H}(\mu) = \int \left(\mu \log \mu - \mu\right) \qquad \rightsquigarrow \qquad \delta_{\mu}\mathscr{H}\left(\mu, x\right) = \log \mu\left(x\right)$$

- Hence, descend $\mathscr{H}$ by microscopically evolving

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

- At the *population level*, equivalent to do

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

- "*The gradient flow of the entropy can be realised by Brownian motion*"

From Gradient Flows to Algorithms

Gradient Flows on Parameter Spaces

$$\min_{\theta \in \Theta} f(\theta)$$

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

"*Gradient Method*"

Gradient Flows on Parameter Spaces

$$\max_{\theta \in \Theta} \log p_\theta \left(y \right)$$

$$\dot{\theta}_t = \nabla_\theta \log p_{\theta_t} \left(y \right)$$

“*MLE by Gradient Flow*”

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

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

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

Gradient Flows on Measure Spaces

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

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

Gradient Flows on Measure Spaces

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

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

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

Gradient Flows on Measure Spaces

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

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

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

Gradient Flows on Measure Spaces

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

$$\mathrm{d}X_t = -\nabla V(X_t)\ \mathrm{d}t - \left(\int \nabla_1 W(X_t, y)\ \mu_t(\mathrm{d}y) \right)\ \mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

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

Gradient Flows on Measure Spaces

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

$$\mathrm{d}X_t^i = -\nabla V\left(X_t^i\right)\,\mathrm{d}t - \frac{1}{N-1}\sum_{j \neq i} \nabla_1 W\left(X_t^i, X_t^j\right)\,\mathrm{d}t + \sqrt{2}\mathrm{d}W_t^i$$

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

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

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

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

Gradient Flows on Hybrid Spaces

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

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

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

Gradient Flows on Hybrid Spaces

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

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

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

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

Gradient Flows on Hybrid Spaces

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

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

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

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

Gradient Flows on Hybrid Spaces

$$\max_{\theta\in\Theta,\mu\in\mathscr{P}(\mathscr{X})} \left\{ \mathbf{E}_{\mu}\left[\log p_{\theta}\left(X,y\right)\right] - \mathscr{H}\left(\mu\right) \right\}$$

$$\dot{\theta}_t = \int \nabla_{\theta} \log p_{\theta_t}\left(x,y\right)\ \mu_t\left(\mathrm{d}x\right)$$

$$\mathrm{d}X_t = \nabla_x \log p_{\theta_t}\left(X_t,y\right)\ \mathrm{d}t + \sqrt{2}\mathrm{d}W_t$$

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

Gradient Flows on Hybrid Spaces

$$\max_{\theta\in\Theta,\mu\in\mathscr{P}(\mathscr{X})}\left\{\mathbf{E}_{\mu}\left[\log p_{\theta}\left(X,y\right)\right]-\mathscr{H}\left(\mu\right)\right\}$$

$$\dot{\theta}_t=\frac{1}{N}\sum_i\nabla_{\theta}\log p_{\theta_t}\left(X_t^i,y\right)$$

$$\mathrm{d}X_t^i=\nabla_x\log p_{\theta_t}\left(X_t^i,y\right)\,\mathrm{d}t+\sqrt{2}\mathrm{d}W_t^i$$

Gradient Flows on Hybrid Spaces

- Space-Time-Discretised Method: “*Particle Gradient Descent*”
- Very well-understood when $\mathscr{F}$ is uniformly-geodesically-convex
- Further theory currently missing / in development; Open Questions

questions about ‘core’ methods?

Some New Directions

Beyond ‘Standard’ Gradient Flows

- Typically, 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’.

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

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

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

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

Warm-Up: Hamiltonian Flows

- An odd idea in isolation:
  - don't descend the objective, but instead conserve the '*Hamiltonian*'
- Introduce skew-symmetric matrix

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

- In abstract terms: with $z = (x, p)$

go from $\dot{x} = -\nabla f(x)$ to $\dot{z} = \mathbf{J}\nabla H(z),$

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

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

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

with ‘expected’ modifications for particle version

okay, but why bother with this?

Conformal Hamiltonian Flows

- Experience suggests to (linearly) blend
  - Hamiltonian circulation, and
  - gradient-type damping *only on the momentum term*
- This routinely yields improved methods

Conformal Hamiltonian Flows

- Blend Hamiltonian circulation with gradient-type *momentum damping*
- 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

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

so what does this look like?

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*"

Conformal Hamiltonian Flows in Action

- For measure optimisation, obtain

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

with ‘expected’ modifications for particle version

  - $\approx$ (*Kinetic*, *Underdamped*, …) Langevin Monte Carlo

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

  - $\approx$ *Momentum* Particle Gradient Descent

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 tasks can 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

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 ROCCO.CAPRIO@WARWICK.AC.UK
Department of Statistics
University of Warwick
Coventry, CV4 7AL, UK

Juan Kuntz

Samuel Power SAM.POWER@BRISTOL.AC.UK
School of Mathematics
University of Bristol
Bristol, BS8 1UG, UK

Adam M. Johansen A.M.JOHANSEN@WARWICK.AC.UK
Department of Statistics
University of Warwick
Coventry, CV4 7AL, UK

Editor: Alexandre Bouchard

Abstract

We derive non-asymptotic error bounds for particle gradient descent (PGD, Kuntz et al. (2023)), 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 the flow converges exponentially fast to the free energy's minimizers for models satisfying a condition that generalizes both the log-Sobolev and the Polyak–Łojasiewicz inequalities (LSI and PŁI, respectively). 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 the extension to our new setting. We also generalize the Bakry–Émery Theorem and show that the LSI/PŁI extension holds for models with strongly concave log-likelihoods. For such models, we further control PGD's discretization error and obtain the 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.

JMLR 2025

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.