Connection Between Optimization and Sampling Algorithms

K.C Zygalakis

School of Mathematics, University of Edinburgh

Maxwell Institute for Mathematical Sciences

The Alan Turing Institute

Analytic and Geometric Approaches to Machine Learning
Bath-LMS Symposium Series
26 July 2021

Collaborators

Jesus M. Sanz-Serna
(UC3M)

- J. M. Sanz-Serna and K. C. Zygalakis. The connections between Lyapunov functions for some optimization algorithms and differential equations. *SIAM Journal on Numerical Analysis*, 59(3), 1542–1565, 2021.
- J. M. Sanz-Serna and K. C. Zygalakis. Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations. *arXiv:2104.12384*, 2021.

Overview

1. Introduction
   - Candidate differential equations
   - Main approach
2. Momentum equation and optimization methods
   - Continuous time
   - Discrete time
   - Analysis of Nesterov method
   - Failure of the heavy ball method
3. Non Asymptotic bounds for numerical integrators of SDEs
   - Continuous time
   - Convergence to equilibrium
   - Discrete time
   - Application to underdamped Langevin dynamics
4. Conclusions

Overview

1. Introduction
   - Candidate differential equations
   - Main approach
2. Momentum equation and optimization methods
   - Continuous time
   - Discrete time
   - Analysis of Nesterov method
   - Failure of the heavy ball method
3. Non Asymptotic bounds for numerical integrators of SDEs
   - Continuous time
   - Convergence to equilibrium
   - Discrete time
   - Application to underdamped Langevin dynamics
4. Conclusions

THE UNIVERSITY *of* EDINBURGH

Statement of two (innocent looking) problems

Optimization

Find the unconstrained minimum of a function $\pi(x)$ in $\mathbb{R}^d$

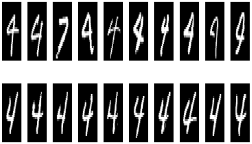
$$\min_{x \in \mathbb{R}^d} \pi(x)$$

Sampling

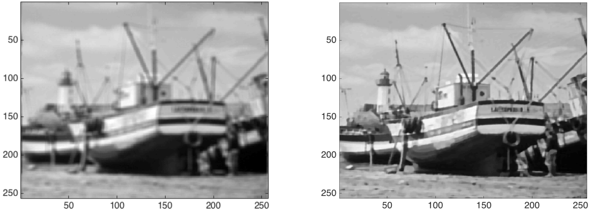
Let $x \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and assume that we want to calculate an expectation with respect to a probability distribution with smooth density $\pi(x)$

$$\pi(g) := \mathbb{E}_\pi(g) = \int_{\mathcal{X}} g(x)\pi(x)dx$$

Numerous applications



(a) Uncertainty quantification for classification methods



(b) Image reconstruction

Gradient flow

Consider the differential equation:

$$\frac{dx}{dt} = -\nabla\pi(x).$$

This has the interesting property that

$$\frac{d\pi(x)}{dt} = -\left\|\nabla\pi(x)\right\|^2 \Rightarrow \lim_{t\to\infty} x(t) = x^*,$$

where x^* is a (unique) minimizer. This makes the equation above central (or at least the simplest choice) for optimization purposes.

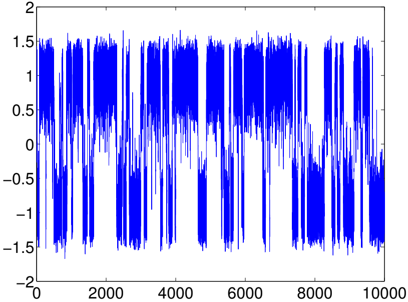
Langevin dynamics

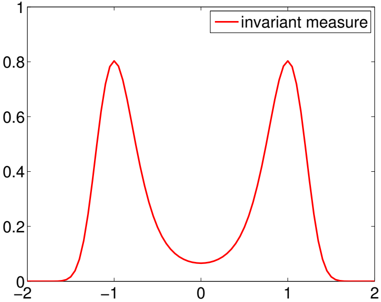
Consider the stochastic differential equation

$$dX_t = \nabla \log \pi(X_t)dt + \sqrt{2}dW_t.$$

Under appropriate assumptions on $\nabla \log \pi(x)$ one can show that its dynamics are ergodic with respect to $\pi(x) : \mathbb{R}^n \mapsto \mathbb{R}$ i.e

$$\lim_{T\to\infty} \frac{1}{T}\int_0^T g(X_s)ds = \mathbb{E}_\pi[g] := \int_{\mathbb{R}^n} g(x)\pi(x)dx.$$





In an ideal world!!!

- There is nothing to be done...
- Discretize the candidate differential equations and go
  - *Optimization*: Go to infinity as quickly as possible (in terms of function evaluations).
  - *Sampling*: Go to infinity as quickly as possible (in terms of function evaluations). Once there produce samples that are i.i.d.

In real life...

- Starting from the differential equation and discretising might not be enough in terms of mimicking the rate of convergence to equilibrium.
- Going to infinity as quickly as possible implies that you can use arbitrary large time-steps in your numerical discretization.
- Reality unfortunately comes back to bite you, as time-steps restrictions appear once you discretize your (stochastic) differential equation.

Overview

1. Introduction
   - Candidate differential equations
   - Main approach
2. Momentum equation and optimization methods
   - Continuous time
   - Discrete time
   - Analysis of Nesterov method
   - Failure of the heavy ball method
3. Non Asymptotic bounds for numerical integrators of SDEs
   - Continuous time
   - Convergence to equilibrium
   - Discrete time
   - Application to underdamped Langevin dynamics
4. Conclusions

Continuous time formulation

$$
\begin{aligned}
\dot{\xi}(t) &= \bar{A}\xi(t) + \bar{B}u(t), \\
y(t) &= \bar{C}\xi(t), \\
u(t) &= \nabla f(y(t)).
\end{aligned}
$$

where $\xi(t) \in \mathbb{R}^n$ is the state, $y(t) \in \mathbb{R}^d (d \leq n)$ the output, and $u(t) = \nabla f(y(t))$ the continuous feedback input. Fixed points of the system satisfy

$$0 = \bar{A}\xi^\star, \quad y^\star = \bar{C}\xi^\star, \quad u^\star = \nabla f(y^\star);$$

in our context $u^\star = 0$ and $y^\star = x^\star$.

[1] M.Fazlyab, A. Ribeiro, M. Morari, V. M. Preciado, *SIAM J. Optim.*, 28(3), 2654–2689, (2018).

The class $\mathcal{F}(m, L)$

1. $\langle x-y, \nabla f(x) - \nabla f(y) \rangle \geq m \|x-y\|^2$.
2. $\|\nabla f(x) - \nabla f(y)\|^2 \leq L^2 \|x-y\|^2$.

An equivalent way of expressing these equations are the following quadratic constraints:

1. $$\begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^T \begin{bmatrix} -mI_d & \frac{1}{2}I_d \\ \frac{1}{2}I_d & 0_d \end{bmatrix} \begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \geq 0.$$
2. $$\begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix}^T \begin{bmatrix} L^2 I_d & 0_d \\ 0_d & -I_d \end{bmatrix} \begin{bmatrix} x-y \\ \nabla f(x) - \nabla f(y) \end{bmatrix} \geq 0.$$

Three examples

1. *Gradient flow:* $\dot{x} = -\nabla f(x)$.

$$\bar{A} = 0_{d\times d}, \quad \bar{B} = -I_{d\times d}, \quad \bar{C} = I_{d\times d}.$$

2. *Momentum equation-convex:* $\ddot{x} + \frac{r}{t}\dot{x} + \nabla f(x) = 0$

$$\bar{A} = \begin{bmatrix} 0_d & 0_d \\ \frac{r-1}{t} I_d & -\frac{r-1}{t} I_d \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} -\frac{t}{r-1} I_d \\ 0_d \end{bmatrix}, \quad \bar{C} = \begin{bmatrix} 0_d & I_d \end{bmatrix}.$$

3. *Momentum equation-strongly convex:* $\ddot{x} + \bar{b}\sqrt{m}\dot{x} + \nabla f(x) = 0$.

$$\bar{A} = \begin{bmatrix} -\bar{b}\sqrt{m} I_d & 0_d \\ \sqrt{m} I_d & 0_d \end{bmatrix}, \quad \bar{B} = \begin{bmatrix} -(1/\sqrt{m}) I_d \\ 0_d \end{bmatrix}, \quad \bar{C} = \begin{bmatrix} 0_d & I_d \end{bmatrix}.$$

[2] W. Su, S. Boyd, E. J. Candés NIPS 2014: 2510-2518, (2014).

(Continuous) Lyapunov functions

Consider

$$V(\xi(t), t) = \alpha(t)(f(y(t)) - f(y_*)) + (\xi(t) - \xi_*)P(t)(\xi(t) - \xi_*)$$

and assume that we can find $\alpha(t), P(t) \succeq 0$ such that

$$V(\xi(t), t) \leq V(\xi(t_0), t_0)$$

then

$$0 \leq f(y(t)) - f(y_*) \leq V(\xi(t_0, t_0))/\alpha(t) = \mathcal{O}(1/\alpha(t))$$

A small calculation

By differentiating the Lyapunov function we have

$$\begin{aligned}\dot{V} &= \dot{\alpha}(t)(f(y(t)) - f(y_*)) \\ &+ \alpha(t)(\nabla f(y(t)))^T \dot{y}(t) \\ &+ 2(\xi(t) - \xi_*)^T P(t)\dot{\xi}(t) \\ &+ (\xi(t) - \xi_*)^T \dot{P}(t)(\xi(t) - \xi_*)^T\end{aligned}$$

Setting $e(t) = [(\xi(t) - \xi_*)^T (u(t) - u_*)^T]$ and using the strong convexity properties of f $(f \in \mathcal{F}_{m,L})$ we can obtain

$$\dot{V(t)} \leq e^T(t)(\cdots)e(t)$$

and if the matrix inside the parenthesis is negative definite then we are done.

A theorem for the (continuous) Lyapunov function

(Continuous) convergence to the minimizer

Suppose that there exist $\lambda > 0$, $\bar{P} \succeq 0$, and $\sigma \geq 0$ that satisfy

$$\bar{T} = \bar{M}^{(0)} + \bar{M}^{(1)} + \lambda \bar{M}^{(2)} + \sigma \bar{M}^{(3)} \preceq 0$$

where

$$\bar{M}^{(0)} = \begin{bmatrix} \bar{P}\bar{A} + \bar{A}^{\mathcal{T}}\bar{P} + \lambda\bar{P} & \bar{P}\bar{B} \\ \bar{B}^{\mathcal{T}}\bar{P} & 0 \end{bmatrix},$$

$$\bar{M}^{(1)} = \frac{1}{2}\begin{bmatrix} 0 & (\bar{C}\bar{A})^{\mathcal{T}} \\ \bar{C}\bar{A} & \bar{C}\bar{B} + \bar{B}^{\mathcal{T}}\bar{C}^{\mathcal{T}} \end{bmatrix},$$

$$\bar{M}^{(2)} = \begin{bmatrix} \bar{C}^{\mathcal{T}} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{m}{2} I_d & \frac{1}{2} I_d \\ \frac{1}{2} I_d & 0 \end{bmatrix} \begin{bmatrix} \bar{C} & 0 \\ 0 & I_d \end{bmatrix},$$

$$\bar{M}^{(3)} = \begin{bmatrix} \bar{C}^{\mathcal{T}} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{mL}{m+L} I_d & \frac{1}{2} I_d, \\ \frac{1}{2} I_d & -\frac{1}{m+L} I_d \end{bmatrix} \begin{bmatrix} \bar{C} & 0 \\ 0 & I_d \end{bmatrix}.$$

Then the following inequality holds for $f \in \mathcal{F}_{m,L}$, $t \geq 0$,

$$f(y(t)) - f(y^\star) \leq e^{-\lambda t}\left(f(y(0)) - f(y^\star) + (\xi(0) - \xi^\star)^{\mathcal{T}}\bar{P}(\xi(0) - \xi^\star)\right).$$

Gradient flow vs momentum equations

- *Gradient flow:* We have that $\lambda = 2m$.
- *Momentum equations:* We have that $\lambda = \sqrt{m}$, when $\bar{b} = 2$

Some observations:

1. The momentum dynamics accelerate the convergence to equilibrium ($\sqrt{m} \gg m$ when $m \ll 1$.)
2. The value of $\bar{b} = 2$ in fact maximizes the decay rate of f for an arbitrary $f \in \mathcal{F}_{m,L}$.

[3] A.C. Wilson, B. Recht, M. I. Jordan, *J. Mach. Learn. Res.* 22 1-34, (2021)

Discrete time

$$
\begin{aligned}
\xi_{k+1} &= A\xi_k + Bu_k, \\
u_k &= \nabla f(y_k), \\
y_k &= C\xi_k, \\
x_k &= E\xi_k.
\end{aligned}
$$

(Discrete) Lyapunov functions

Consider

$$V_k(\xi) = \rho^{-2k}\left(a_0(f(x_k) - f(x^\star)) + (\xi_k - \xi^\star)^\mathcal{T} P(\xi_k - \xi^\star)\right),$$

and assume that we can find $a_0 > 0, P \succeq 0$ such that

$$V_{k+1}(\xi_{k+1}) \leq V_k(\xi_k),$$

we can then conclude

$$f(x_k) - f(x^\star) \leq \rho^{2k} \frac{V_0(\xi_0)}{a_0}.$$

If $\rho < 1$, we have found a convergence rate for $f(x_k)$ towards the optimal value $f(x^\star)$.

A theorem for the (discrete) Lyapunov function

(Discrete) convergence to miminizer

Suppose that there exist $a_0 > 0$, $P \succeq 0$, $\ell > 0$, and $\rho \in [0,1)$ such that

$$T = M^{(0)} + a_0 \rho^2 M^{(1)} + a_0(1-\rho^2) M^{(2)} + \ell M^{(3)} \preceq 0,$$

where

$$M^{(0)} = \begin{bmatrix} A^{\mathcal{T}} PA - \rho^2 P & A^{\mathcal{T}} PB \\ B^{\mathcal{T}} PA & B^{\mathcal{T}} PB \end{bmatrix}, \quad M^{(1)} = N^{(1)} + N^{(2)}, \quad M^{(2)} = N^{(1)} + N^{(3)}, \quad M^{(3)} = N^{(4)},$$

with

$$N^{(1)} = \begin{bmatrix} EA - C & EB \\ 0 & I_d \end{bmatrix}^{\mathcal{T}} \begin{bmatrix} \frac{L}{2} I_d & \frac{1}{2} I_d \\ \frac{1}{2} I_d & 0 \end{bmatrix} \begin{bmatrix} EA - C & EB \\ 0 & I_d \end{bmatrix},$$

$$N^{(2)} = \begin{bmatrix} C - E & 0 \\ 0 & I_d \end{bmatrix}^{\mathcal{T}} \begin{bmatrix} -\frac{m}{2} I_d & \frac{1}{2} I_d \\ \frac{1}{2} I_d & 0 \end{bmatrix} \begin{bmatrix} C - E & 0 \\ 0 & I_d \end{bmatrix},$$

$$N^{(3)} = \begin{bmatrix} C^{\mathcal{T}} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{m}{2} I_d & \frac{1}{2} I_d \\ \frac{1}{2} I_d & 0 \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & I_d \end{bmatrix},$$

$$N^{(4)} = \begin{bmatrix} C^{\mathcal{T}} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{mL}{m+L} I_d & \frac{1}{2} I_d \\ \frac{1}{2} I_d & -\frac{1}{m+L} I_d \end{bmatrix} \begin{bmatrix} C & 0 \\ 0 & I_d \end{bmatrix}.$$

Then, for $f \in \mathcal{F}_{m,L}$, the sequence $\{x_k\}$ satisfies $f(x_k) - f(x^\star) \leq \frac{a_0(f(x_0) - f(x^\star)) + (\xi_0 - \xi^\star)^{\mathcal{T}} P (\xi_0 - \xi^\star)}{a_0} \rho^{2k}$.

A family of algorithms

$$x_{k+1} = x_k + \beta(x_k - x_{k-1}) - \alpha\nabla f(y_k),$$
$$y_k = x_k + \gamma(x_k - x_{k-1}),$$

1. For $\beta = \gamma = 0$ we recover the gradient descent

$$x_{k+1} = x_k - \alpha f(x_k).$$

2. For $\gamma = \beta$ we recover the Nesterov method.
3. For $\gamma = 0$, $\beta \neq 0$ we recover the heavy ball method.

Nesterov method

We introduce $\delta = \sqrt{m\alpha}$ and $d_k = \frac{1}{\delta}(x_k - x_{k-1})$, so we can re-write our algorithm as:

$$\begin{aligned} d_{k+1} &= \beta d_k - \frac{\alpha}{\delta}\nabla f(y_k), \\ x_{k+1} &= x_k + \delta\beta d_k - \alpha\nabla f(y_k), \\ y_k &= x_k + \delta\beta d_k. \end{aligned}$$

Setting $\xi_k = [d_k^{\mathcal{T}}, x_k^{\mathcal{T}}]^{\mathcal{T}} \in \mathbb{R}^{2d}$ we can express the algorithm in the discrete form with

$$A = \begin{bmatrix} \beta I_d & 0 \\ \delta\beta I_d & I_d \end{bmatrix}, \quad B = \begin{bmatrix} -(\alpha/\delta) I_d \\ -\alpha I_d \end{bmatrix}, \quad C = \begin{bmatrix} \delta\beta I_d & I_d \end{bmatrix}, \quad E = \begin{bmatrix} 0 & I_d \end{bmatrix}.$$

Dimension reduction

- The matrix A is a a Kronecker product of a 2×2 matrix and I_d,

$$A = \begin{bmatrix} \beta & 0 \\ \delta\beta & 1 \end{bmatrix} \otimes I_d;$$

- The matrices B, C and E have a similar Kronecker product structure.
- It is then natural to consider symmetric matrices P of the form

$$P = \widehat{P} \otimes I_d, \qquad \widehat{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix},$$

- T will also have a Kronecker product structure

$$T = \widehat{T} \otimes I_d, \qquad \widehat{T} = \begin{bmatrix} t_{11} & t_{12} & t_{13} \\ t_{12} & t_{22} & t_{23} \\ t_{13} & t_{23} & t_{33} \end{bmatrix}.$$

Structure of $\widehat{T}$

We have

$$\begin{aligned}
t_{11} &= \beta^2 p_{11} + 2\delta\beta^2 p_{12} + \delta^2\beta^2 p_{22} - \rho^2 p_{11} - \delta^2\beta^2 m/2,\\
t_{12} &= \beta p_{12} + \delta\beta p_{22} - \rho^2 p_{12} - \delta\beta m/2 + \rho^2\delta\beta m/2,\\
t_{13} &= -\delta^{-1}\alpha\beta p_{11} - 2\alpha\beta p_{12} - \delta\alpha\beta p_{22} + \delta\beta/2,\\
t_{22} &= p_{22} - \rho^2 p_{22} - m/2 + \rho^2 m/2,\\
t_{23} &= -\delta^{-1}\alpha p_{12} - \alpha p_{22} + 1/2 - \rho^2/2,\\
t_{33} &= \delta^{-2}\alpha^2 p_{11} + 2\delta^{-1}\alpha^2 p_{12} + \alpha^2 p_{22} + \alpha^2 L/2 - \alpha.
\end{aligned}$$

Our task is to find $\rho \in [0,1)$, p_{11}, p_{12}, and p_{22} that lead to $\widehat{T} \preceq 0$ and $\widehat{P} \succeq 0$ (which imply $T \preceq 0$ and $P \succeq 0$).

Solution

The algebra becomes simpler if we represent β and ρ^2 as:

$$\beta = 1 - b\delta, \quad \rho^2 = 1 - r\delta.$$

Note that we are interested in $r \in (0, 1/\delta]$ so as to get $\rho^2 \in [0,1)$. Going through the algebra we find

$$\widehat{P} = \begin{bmatrix} p_{11} & p_{12} \\ p_{12} & p_{22} \end{bmatrix} = \frac{m}{2} \begin{bmatrix} (1-r\delta)^2 & r(1-r\delta) \\ r(1-r\delta) & r^2 \end{bmatrix}, \quad \alpha \leq \frac{1}{L}, \quad r \leq 1$$

as well as $\Xi = 0$ where

$$\Xi := \Xi_\delta(r,b) = (r+\delta)(1-\delta^2)b^2 - 2(1+r^2)(1-\delta^2)b + (r^3 - 3r^2\delta + 3r - \delta).$$

- Since $\delta = \sqrt{m\alpha}$ and $\alpha \leq L^{-1}$, this implies that

$$\rho^2 = 1 - \frac{r}{\sqrt{\kappa}}$$

hence the Nesterov algorithm maintains the acceleration of the original differential equation.

Convergence of the algorithm

Theorem

With the choices of parameters as in the previous slide the matrix T is negative semi-definite. As a result, for any x_{-1}, x_0, the sequence

$$\rho^{-2k}\left(f(x_k) - f(x_\star) + [d_k^{\mathcal{T}}, x_k^{\mathcal{T}} - x_\star^{\mathcal{T}}]\, P\, [d_k^{\mathcal{T}}, x_k^{\mathcal{T}} - x_\star^{\mathcal{T}}]^{\mathcal{T}}\right)$$

decreases monotonically, which, in particular, implies

$$f(x_k) - f(x_\star) \leq C\rho^{2k}$$

with

$$C = f(x_0) - f(x^\star) + \frac{m}{2}\left\|\frac{1 - r\delta}{\delta}(x_0 - x_{-1}) + r(x_0 - x^\star)\right\|^2.$$

Connection with the ODE

Convergence between discrete and continuous Lyapunov function

Fix the parameter $\bar{b} > 0$ and the initial conditions $x(0)$, $\dot{x}(0)$ for the momentum equations. For small $h > 0$, consider the Nesterov method with parameters $\alpha = h^2$ and $\beta = \beta_h = 1 - \bar{b}\sqrt{m}h + o(h)$. Assume that the initial points x_{-1}, x_0 are such that, as $h \downarrow 0$, $x_0 \to x(0)$ and $(1/h)(x_0 - x_{-1}) \to \dot{x}(0)$. Then, in the limit $kh \to t$,

1. $x_k \to x(t)$ and $(1/h)(x_{k+1} - x_k) \to \dot{x}(t)$.
2. The discrete Lyapunov function converges to the continuous Lyapunov function

Is consistency enough?

1. From an intuitive point of view the previous theorem is obvious, *i.e* you start with and ODE you discretise it and the numerical algorithm inherits its properties for some finite h
2. The key however is how large this h can be, while maintaining the negative definiteness of the matrix T.
3. From consistency in order to achieve acceleration one needs to be able to preserve the negative definiteness of T for time steps $h \leq cL^{-1/2}$
   - In the case of the Nesterov method one has that $h \leq L^{-1/2}$, which leads to acceleration.
   - This is however not true in general. In particular in the case of ▸ Heavy ball method one can show that the matrix T cannot be negative definite for $h \leq cL^{-1/2}$ for any $c > 0$, and hence the heavy ball method doesn't lead to acceleration.

[4] L. Lessard, B. Recht, A. Packard, *SIAM J. Optim.*, 26(1), 57–95. (2016)

Overview

1. Introduction
   - Candidate differential equations
   - Main approach

2. Momentum equation and optimization methods
   - Continuous time
   - Discrete time
   - Analysis of Nesterov method
   - Failure of the heavy ball method

3. Non Asymptotic bounds for numerical integrators of SDEs
   - Continuous time
   - Convergence to equilibrium
   - Discrete time
   - Application to underdamped Langevin dynamics

4. Conclusions

Continuous time formulation

$$
\begin{aligned}
d\xi(t) &= A\xi(t)dt + Bu(t)dt + \sigma dW(t),\\
x(t) &= C\xi(t),\\
u(t) &= \nabla f(x(t)).
\end{aligned}
$$

Here $\xi \in \mathbb{R}^N$ is the state, $u \in \mathbb{R}^d$ is the input, $x \in \mathbb{R}^d$ is the output that is mapped to u by the nonlinear map ∇f and W represents the standard M-dimensional Brownian motion. The real matrices A, B, C and σ are constant, with sizes $N \times N$, $N \times d$, $d \times N$ and $N \times M$ respectively. We define

$$D = (1/2)\sigma\sigma^T.$$

Two examples

1. The overdamped Langevin equation

$$dx = -c\nabla f(x)\,dt + \sqrt{2c}\,dW(t),$$

for which we have $N = d, \quad M = d$, $\xi = x$, and

$$A = 0_d \quad B = -cI_d \quad C = I_d \quad \sigma = \sqrt{2c}I_d.$$

2. The underdamped Langevin equation

$$\begin{aligned} dv &= -\gamma v\,dt - c\nabla f(x)\,dt + \sqrt{2\gamma c}\,dW(t), \\ dx &= v\,dt. \end{aligned}$$

for which we have $N = 2d$, $M = d$, $\xi = [v^T, x^T]^T$ and

$$A = \begin{bmatrix} -\gamma I_d & 0 \\ I_d & 0 \end{bmatrix}, \quad B = \begin{bmatrix} -cI_d \\ 0 \end{bmatrix}, \quad C = \begin{bmatrix} 0 & I_d \end{bmatrix}, \quad \sigma = \begin{bmatrix} \sqrt{2\gamma c}I_d \\ 0 \end{bmatrix}.$$

Equilibrium behaviour

Necessary conditions

Assume that S is an $N \times N$ positive semidefinite symmetric matrix.

- The relations

$$\begin{aligned} \mathrm{Tr}(A+DS) &= 0, \\ CB + CDC^T &= 0, \\ CA + B^T S + 2CDS &= 0, \\ SA + A^T S + 2SDS &= 0, \end{aligned}$$

  imply that the SDE has invariant probability distribution $\pi^\star$ with density

$$\propto \exp\big(-f(C\xi) - (1/2)\xi^T S\xi\big).$$

- If $SC^T = 0$, then the marginal of $\propto \exp\big(-f(C\xi) - (1/2)\xi^T S\xi\big)$ on $x = C\xi$ is the target $\propto \exp(-f(x))$.

Convergence to the invariant distribution I

- Similarly to the optimization case a natural question to ask is how fast does the true solution of the SDE converges to the invariant measure?
- We will do by bounding the error in terms of time of the following Wasserstein distance

  $$W_P(\Phi_t \pi, \pi^\star)$$

  where π denotes the probability distribution of the initial value $\xi(0)$, while

  $$W_P(\pi_1, \pi_2) = \left(\inf_{\zeta \in Z} \int_{\mathbb{R}^N} \|x - y\|_P^2 d\zeta(x, y) \right)^{1/2},$$

with P a positive definite matrix, and where Z is the set of all couplings between π_1 and π_2.

Convergence to the invariant distribution II

In order to estimate the quantity of interest, we consider

$$\begin{aligned} d\xi^{(1)}(t) &= A\xi^{(1)}(t)dt + B\nabla f(C\xi^{(1)}(t))dt + \sigma dW(t), \\ d\xi^{(2)}(t) &= A\xi^{(2)}(t)dt + B\nabla f(C\xi^{(2)}(t))dt + \sigma dW(t), \end{aligned}$$

Contractivity implies convergence

Assume that $P \succ 0$ and $\lambda > 0$ exist such almost surely,

$$\|\xi^{(2)}(t) - \xi^{(1)}(t)\|_P^2 \leq e^{-\lambda t}\|\xi^{(2)}(0) - \xi^{(1)}(0)\|_P^2, \qquad t > 0.$$

Then, for arbitrary distributions, π_1 and π_2,

$$W_P(\Phi_t\pi_1, \Phi_t\pi_2) \leq e^{-\lambda t/2} W_P(\pi_1, \pi_2), \qquad t > 0,$$

and, in particular, for arbitrary π,

$$W_P(\Phi_t\pi, \pi^\star) \leq e^{-\lambda t/2} W_P(\pi, \pi^\star), \qquad t > 0.$$

Convergence to equilibrium III

- On top of assuming that $f \in \mathcal{F}(m, L)$ we will assume that it is twice differentiable. This implies that the eigenvalues of $\nabla\nabla f$ are bounded between m and L

Another matrix formulation

Let $P \succ 0$ be an $N \times N$ symmetric matrix and $\lambda > 0$. Assume that, for each $y_1, y_2 \in \mathbb{R}^d$, the matrix

$$\mathcal{T}(\lambda, P, y_1, y_2) = \lambda P + P\big(A + B\bar{\mathcal{H}}(y_1, y_2)C\big) + \big(A + B\bar{\mathcal{H}}(y_1, y_2)C\big)^T P$$

is $\preceq 0$. Then the contractivity estimates hold. Here

$$\bar{\mathcal{H}}(y_2, y_1) = \int_0^1 \mathcal{H}\big(y_1 + z[y_2 - y_1]\big)\, dz$$

Dimension reduction

- The previous proposition is difficult to use in practice.
- The following structure though is typical in applications

$$A = \widehat{A} \otimes I_d, \qquad B = \widehat{B} \otimes I_d, \qquad C = \widehat{C} \otimes I_d,$$

Continuous generalized eigenvalue problem

Given the symmetric, positive definite $\widehat{P}$, and $\widehat{Z}(H)$ given by

$$\widehat{Z}(H) = -\widehat{P}\left(\widehat{A} + H\widehat{B}\widehat{C}\right) - \left(\widehat{A} + H\widehat{B}\widehat{C}\right)^T \widehat{P}.$$

Assume that, as H varies in $[m, L]$, the eigenvalues Λ of the generalized eigenvalue problem $\widehat{Z}(H)x = \Lambda \widehat{P}x$ are positive and bounded away from zero and let $\lambda > 0$ be the infimum of those eigenvalues. Then the contractivity bound with $P = \widehat{P} \otimes I_d$ holds almost surely.

Two examples

1. ▸ Overdamped Langevin equation: We have that $\widehat{P} = 1$, and that $\lambda = 2cm$.
2. ▸ Underdamped Langevin equation: For $c = 1/L$ we have $\lambda = 1/\kappa$ and

$$\widehat{P} = \begin{bmatrix} 1 & 1 \\ 1 & 2 \end{bmatrix}, \qquad \widehat{L} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix}.$$

- It is possible to show that the best possible rate corresponds to the choice of $c = 4/(L + m)$ yeilding $\lambda = 4/(\kappa + 1)$

Discrete time formulation

We will focus on algorithms with one function evalution

$$\begin{aligned} \xi_{n+1} &= A_h \xi_n + B_h u_n + \sigma_h^{\xi} \Omega_n, \\ y_n &= C_h \xi_n + \sigma_h^{y} \Omega_n, \\ u_n &= \nabla f(y_n). \end{aligned}$$

- Similarly to the continuous case we want to study the convergence to equilibrium
- Note that in general the numerical equilibrium will be different than the invariant measure of the continuous time SDE

Convergence to (discrete) equilibrium

In order to estimate the quantity of interest we will consider

$$\begin{aligned}\xi_{n+1}^{(1)} &= A_h\xi_n^{(1)} + B_h\nabla f(C_h\xi_n^{(1)} + \sigma_h^y\Omega_n) + \sigma_h^\xi\Omega_n,\\ \xi_{n+1}^{(2)} &= A_h\xi_n^{(2)} + B_h\nabla f(C_h\xi_n^{(2)} + \sigma_h^y\Omega_n) + \sigma_h^\xi\Omega_n,\end{aligned}$$

and denote by $\Psi_{h,n}\pi$ the probability distribution for ξ_n of the numerical solution when π is the distribution of ξ_0

Contractivity implies convergence

Assume that $P_h \succ 0$ and $\rho_h \in (0,1)$ exist such that almost surely,

$$\|\xi_{n+1}^{(2)} - \xi_{n+1}^{(1)}\|_{P_h}^2 \leq \rho_h\|\xi_n^{(2)} - \xi_n^{(1)}\|_{P_h}^2, \qquad n = 0, 1, \dots$$

Then, for arbitrary distributions, π_1 and π_2,

$$W_P(\Psi_{h,n}\pi_1, \Psi_{h,n}\pi_2) \leq \rho_h^{n/2} W_P(\pi_1, \pi_2), \qquad n = 0, 1, \dots$$

Checking discrete contractivity

- In a similar way as in the continuous case one can reduce the high dimensional matrix inequality to a low dimensional generalized eigenvalue problem

Discrete generalized eigenvalue problem

Given the symmetric, positive definite $\widehat{P}_h$, set

$$\widehat{Z}_h(H) = \left(\widehat{A}_h + H\widehat{B}_h\widehat{C}_h\right)^T \widehat{P}_h\left(\widehat{A}_h + H\widehat{B}_h\widehat{C}_h\right).$$

Assume that, as H varies in $[m, L]$, the supremum ρ_h of the eigenvalues R of the generalized eigenvalue problems $\widehat{Z}_h(H)x = R\widehat{P}x$ is < 1. Then the contractivity bound with $P_h = \widehat{P}_h \otimes I_d$ holds almost surely.

A general error decomposition

- We are interested in characterising the following error

$$W_{P_h}(\Psi_{h,n+1}\pi, \pi^*)$$

- There are two different ways to decompose it
    1. $W_{P_h}(\Psi_{h,n+1}\pi, \pi^*) \le \underbrace{W_{P_h}(\Psi_h(\Psi_{h,n}\pi), \Psi_h\pi^*)}_{\text{numerical contraction}} + \underbrace{W_{P_h}(\Psi_h\pi^*, \Phi_h\pi^*)}_{\text{local error}}$
    2. $W_{P_h}(\Psi_{h,n+1}\pi, \pi^*) \le \underbrace{W_{P_h}(\Phi_h(\Psi_{h,n}\pi), \Phi_h\pi^*)}_{\text{SDE contraction}} + \underbrace{W_{P_h}(\Psi_h(\Psi_{h,n}\pi), \Phi_h(\Psi_{h,n}\pi))}_{\text{local error}}$
- We will follow the first decomposition, the first term is controlled by the numerical contractivity of the numerical scheme, while the second term relates to the local strong order of convergence of the numerical scheme.

Bringing everything together

A general theorem

Assume that there are constants $h_0 > 0$, $r > 0$ such that for $h \leq h_0$ the contractivity estimate holds with $\rho_h \leq (1 - rh)^2$. Then, for any initial distribution π, stepsize $h \leq h_0$, and $n = 0, 1, \ldots$,

$$W_{P_h}(\pi^\star, \Psi_{h,n}\pi) \leq (1 - hR_h)^n W_{P_h}(\pi^\star, \pi) + \left(\frac{\sqrt{2} C_1}{\sqrt{R_h}} + \frac{C_2}{R_h} \right) h^p,$$

with

$$R_h = \frac{1}{h}\left(1 - \sqrt{(1 - rh)^2 + C_0 h^2}\right) = r + o(1), \quad \text{as} \quad h \downarrow 0.$$

Non-asymptotic estimates

The theorem allows us to study arbitrary one step integrators in terms of their non-asymptotic properties, namely how many steps n one should make in order to ensure that $W_{P_h}(\Psi_{h,n}\pi, \pi^*) < \epsilon$

[5] A. S. Dalalyan, COLT2017
[6] A. S. Dalalyan and A. Karagulyan, *Stoch. Proc. Appl,* 129(12):5278–5311, (2019).
[7] A. Durmus and E. Moulines, *Ann. Appl. Probab.*27(3):1551–1587, (2017)

Exponential Euler

$$
\begin{aligned}
v_{n+1} &= \mathcal{E}(h)v_n - \mathcal{F}(h)c\nabla f(x_n) + \sqrt{2\gamma c}\int_{t_n}^{t_{n+1}} \mathcal{E}(t_{n+1}-s)dW(s),\\
x_{n+1} &= x_n + \mathcal{F}(h)v_n - \mathcal{G}(h)c\nabla f(x_n) + \sqrt{2\gamma c}\int_{t_n}^{t_{n+1}} \mathcal{F}(t_{n+1}-s)dW(s).
\end{aligned}
$$

where

$$
\mathcal{E}(t) = \exp(-\gamma t), \qquad \mathcal{F}(t) = \int_0^t \mathcal{E}(s)\,ds = \frac{1-\exp(-\gamma t)}{\gamma},
$$

and

$$
\mathcal{G}(t) = \int_0^t \mathcal{F}(s)\,ds = \frac{\gamma t + \exp(-\gamma t) - 1}{\gamma^2}.
$$

- Analysing this integrator using the tools developed the number of steps needed to achieve the desired accuracy scales as

$$
(m^{1/2}\epsilon)^{-1}\kappa^{3/2}d^{1/2}.
$$

- This is an improvement of the previous available estimate $\mathcal{O}(\epsilon^{-1}\kappa^2 d^{1/2})$

[8] X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan, *COLT 2018*.

UBU algorithm

$$
\begin{aligned}
v_{n+1} &= \mathcal{E}(h)v_n - h\mathcal{E}(h/2)c\nabla f(y_n) + \sqrt{2\gamma c}\int_{t_n}^{t_{n+1}} \mathcal{E}(t_{n+1}-s)dW(s), \\
x_{n+1} &= x_n + \mathcal{F}(h)v_n - h\mathcal{F}(h/2)c\nabla f(y_n) + \sqrt{2\gamma c}\int_{t_n}^{t_{n+1}} \mathcal{F}(t_{n+1}-s)dW(s), \\
y_n &= x_n + \mathcal{F}(h/2)v_n + \sqrt{2\gamma c}\int_{t_n}^{t_{n+1/2}} \mathcal{F}(t_{n+1/2}-s)dW(s).
\end{aligned}
$$

1. This is a second order strong integrator
2. Under further smoothness assumptions on the third derivative, the number of steps n to achieve the desired accuracy scales as

$$(m^{1/2}\epsilon)^{-1/2}\kappa^{5/4}(1 + L^{-3/2}L_1)^{1/2}d^{1/4}.$$

[9] A. Alamo and J. M. Sanz-Serna, SIAM J. Numer. Anal., 54(6):3239–3257, (2016)

Overview

1. Introduction
   - Candidate differential equations
   - Main approach
2. Momentum equation and optimization methods
   - Continuous time
   - Discrete time
   - Analysis of Nesterov method
   - Failure of the heavy ball method
3. Non Asymptotic bounds for numerical integrators of SDEs
   - Continuous time
   - Convergence to equilibrium
   - Discrete time
   - Application to underdamped Langevin dynamics
4. Conclusions

Conclusions

- (Stochastic) differential equations are excellent starting point in terms of designing (sampling) optimization algorithms.
- However for optimization algorithms stability is crucial in terms of being able to utilize the favourable convergence rates of the continuous system.
- In terms of designing sampling methods one needs to be paying attention to
  1. the contractivity properties of the numerical scheme.
  2. the strong order of convergence of the numerical scheme.

Bibliography

1 M. Fazlyab, A. Ribeiro, M. Morari, and V. M. Preciado. Analysis of optimization algorithms via integral quadratic constraints: nonstrongly convex problems. *SIAM Journal on Optimization*, 28(3):2654–2689, 2018

2 W. Su, S. Boyd, and E. J. Candès. A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43, 2016.

3 A. C. Wilson, B. Recht, and M. I. Jordan. A Lyapunov analysis of momentum methods in optimization. *Journal of Machine Learning Research*, 22(113): 1–34, 2021.

4 L. Lessard, B. Recht, and A. Packard. Analysis and design of optimization algorithms via integral quadratic constraints. *SIAM Journal on Optimization*, 26(1):57–95, 2016.

5 A. Dalalyan. Theoretical guarantees for approximate sampling from smooth and log-concave densities. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):651–676, 2017.

6 A. Dalalyan and A. Karagulyan. User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient. *Stochastic Processes and their Applications*, 129(12):5278–5311, 2019.

7 A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted Langevin algorithm. *The Annals of Applied Probability*, 27(3):1551 – 1587, 2017.

8 X. Cheng, N. S. Chatterji, P. L. Bartlett, and M. I. Jordan. Underdamped Langevin MCMC: A non-asymptotic analysis. In *Proceedings of the 2018 Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pages 300–323, 2018.

9 A. Alamo and J. M. Sanz-Serna. A technique for studying strong and weak local errors of splitting stochastic integrators. *SIAM Journal on Numerical Analysis*, 54(6):3239–3257, 2016.