Stochastic Differential Equations, Discrete Approximations, and Connections to Optimization and Sampling Algorithms

K.C Zygalakis

School of Mathematics, University of Edinburgh

Maxwell Institute for Mathematical Sciences

Summer School on Applied Analysis
TU Chemnitz

Collaborators

Jesus M. Sanz-Serna (UC3M)

Paul Dobson (Heriot Watt)

- 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).
- P. Dobson, J. M. Sanz-Serna and K. C. Zygalakis, On the connections between optimization algorithms, Lyapunov functions, and differential equations: theory and insights, arXiv:2305.08658, (2023).
- J. M. Sanz-Serna and K. C. Zygalakis. Wasserstein distance estimates for the distributions of numerical approximations to ergodic stochastic differential equations. *Journal of Machine Learning Research*, 22, 1–37, (2021).

Engineering and Physical Sciences Research Council
The Leverhulme Trust

Overview

Statement of two (innocent looking) problems

Optimization

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

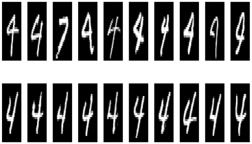
$$\min_{x \in \mathbb{R}^d} \pi(x)$$

Sampling

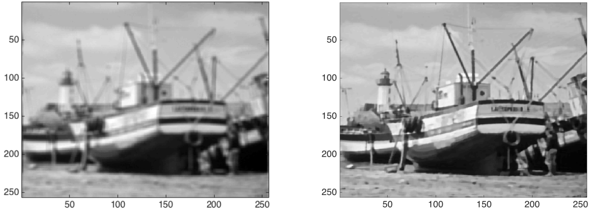
Let $x \in \mathcal{X}$, where $\mathcal{X} \subseteq \mathbb{R}^d$ and assume that we want to calculate an expectation with respect to a probability distribution with smooth density $\pi(x)$

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

Numerous applications



(a) Uncertainty quantification for classification methods



(b) Computational Imaging

THE UNIVERSITY of EDINBURGH

Computational Imaging I

- We are interested in an unknown image $x \in \mathbb{R}^d$.
- We measure y, related to x by a statistical model $p(y|x)$.
- The recovery of x from y is ill-posed or ill-conditioned, resulting in significant uncertainty about x.
- For example, in many imaging problems

$$y = Ax + w,$$

for some operator A that is rank-deficient, and additive noise w.

THE UNIVERSITY of EDINBURGH

Computational Imaging II

- We use priors to reduce uncertainty and deliver accurate results.
- Given the prior $p_r(x)$, the posterior distribution of x given y

$$\pi(x) := p(x|y) = p(y|x)p_r(x)/p_r(y)$$

  models our knowledge about x after observing y.
- Give a functional form to $p_r(x)$, we obtain

$$\pi(x) = \exp\{-\phi(x)\}/Z \ , \quad Z = \int \exp\{-\phi(x)\}\mathrm{d}x$$

Two approaches

1. MAP estimation:

$$\hat{x}_{MAP} = \underset{x\in\mathbb{R}^d}{\operatorname{argmax}}\, p(x|y) = \underset{x\in\mathbb{R}^d}{\operatorname{argmin}}\, \phi(x)$$

2. MMSE estimation:

$$\hat{x}_{MMSE} = \underset{x\in\mathbb{R}^d}{\operatorname{argmin}} \int \|u - x\|^2 \, p(x|y)dx = \mathbb{E}(x|y) = \mathbb{E}_{p(x|y)}(x)$$

THE UNIVERSITY of EDINBURGH

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.

THE UNIVERSITY of EDINBURGH

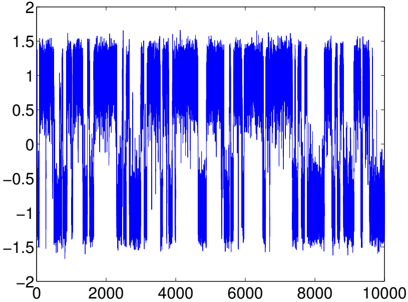
Langevin dynamics

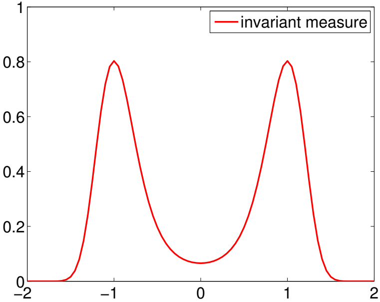
Consider the stochastic differential equation

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

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

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





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

Optimization

If we assume that x_* is the unique minimizer of $\pi(x)$ we want to understand

- Continuous time: How fast does $\pi(x(t)) - \pi(x_*)$ convergences to zero (equivalently how fast does $\|x(t) - x_*\|$ convergences to zero)
- Discrete time: How fast does $\pi(x_k) - \pi(x_*)$ convergences to zero (equivalently how fast does $\|x_k - x_*\|$ convergences to zero)

Thinking on space of probability measures

We can think that the solution of the gradient flow in terms of defining a solution on the space of probability measures as the time t evolves. Of course it is the trivial one since the probability measure induced is just $\delta_{x(t)}$.

Sampling

- The solution of Langevin equation is now defining a (more complicated) probability measure
- Similar to optimization we want to understand how fast does the corresponding probability measure induced convergences to the limit measure π_*
- A lot of different ways of measuring the convergence, we will do this using the Wasserstein distance

Wasserstein distance

We define

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

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

- Continuous time: How fast does $W_P(\Phi_t \pi_0, \pi^\star)$ converges to zero?
- Discrete time: How fast does $W_P(\Psi_h^n \pi_0, \pi^\star)$ converges?

An example

1. Consider the SDE $dX = -\gamma X dt + \sqrt{2} dW$,
   - $\pi_*(x)$ is the density associated with $\mathcal{N}(0, \gamma^{-1})$
2. We will solve this by $x_{n+1} = x_n - \gamma h x_n + \sqrt{2h}\xi_n$, where $\xi_n \simeq \mathcal{N}(0,1)$
   - Assume that x_0 is deterministic
   - It is not difficult to show that in this case $x_n \sim \mathcal{N}(m_n, \sigma_n^2))$ where

$$
\begin{aligned}
m_n &= (1-\gamma h)^n x_0 \to 0 \\
\sigma_n &= 2h \left[\frac{(1-\gamma h)^{2n} - 1}{(1-\gamma h)^2 - 1} \right] \to \frac{2}{2\gamma - \gamma^2 h} \neq \gamma^{-1}
\end{aligned}
$$

3. We hence see that $W_P(\Psi_h^n \pi_0, \pi^\star)$ doesn't convergence to zero. This is genuinely true in the case of SDEs something to keep in mind for later analysis.

Overview

Optimization

Equations and functions assumptions

Gradient flow:

$$\dot{x} + \nabla f(x) = 0$$

Momentum equation:

$$\ddot{x} + \bar{b}\sqrt{m}\dot{x} + \nabla f(x) = 0$$

Quadratic case: $f(x) = \frac{1}{2}x^T Qx, \quad \sigma(Q) \in [m, L]$

Nonlinear case: $f(x) \in \mathcal{F}(m, L)$

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

Continuous time formulation

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

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 $y^\star = x^\star$.

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

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: $\ddot{x} + \bar{b}\sqrt{m}\dot{x} + \nabla f(x) = 0$.

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

Quadratic case

- The continuous time formulation now becomes

$$\dot{\xi}(t) = (\bar{A} + \bar{B}Q\bar{C})\xi(t)$$

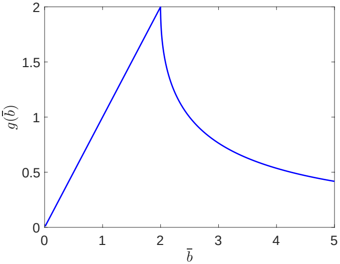
- Solution is given by

$$\xi(t) = e^{(\bar{A}+\bar{B}Q\bar{C})t}\xi(0)$$

- To deduce a convergence rate to the minimizer we need to understand the spectral properties of $e^{(\bar{A}+\bar{B}Q\bar{C})t}$

Quadratic case: Gradient flow vs momentum equations

- Gradient flow: rate of convergence e^{-2mt}
- Momentum equation: rate of convergence $e^{-g(\bar{b})\sqrt{m}t}$



- Clearly using the first order dynamics is suboptimal in terms of convergence

(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

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

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

1. $\langle x - y, \nabla f(x) - \nabla f(y) \rangle \geq m \|x - y\|^2$.
2. $\|\nabla f(x) - \nabla f(y)\|^2 \leq L^2 \|x - y\|^2$.
3. $\frac{mL}{m+L} \|x - y\|^2 + \frac{1}{m+L} \|\nabla f(x) - \nabla f(y)\|^2 \leq (\nabla f(x) - \nabla f(y))^T (x - y)$

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

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

(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

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

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

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

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

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

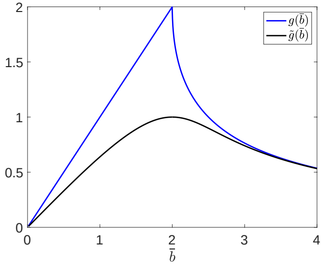
$$\bar{M}^{(3)} = \begin{bmatrix} \bar{C}^{\mathcal{T}} & 0 \\ 0 & I_d \end{bmatrix} \begin{bmatrix} -\frac{mL}{m+L} I_d & \frac{1}{2} I_d, \\ \frac{1}{2} I_d & -\frac{1}{m+L} I_d \end{bmatrix} \begin{bmatrix} \bar{C} & 0 \\ 0 & I_d \end{bmatrix}.$$

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

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

Nonlinear case: Gradient flow vs momentum equations

- Gradient flow: Again we have that $\lambda = 2m$.
- Momentum equations: We have that $\lambda = \tilde{g}(\bar{b})\sqrt{m}$



1. You lose some of the rate you can prove between the linear and the nonlinear case
2. Still the momentum dynamics accelerate the convergence to equilibrium ($\sqrt{m} \gg m$ when $m \ll 1$.)
3. One should discretise the momentum dynamics.

Sampling

Continuous time formulation

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

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 $barA$, $\bar{B}$, $\bar{C}$ and σ are constant, with sizes $N \times N$, $N \times d$, $d \times N$ and $N \times M$ respectively. We define

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

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

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

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

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

Equilibrium behaviour

Necessary conditions

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

- The relations

$$
\begin{aligned}
\mathrm{Tr}(\bar{A} + \bar{D}S) &= 0, \\
\bar{C}\bar{B} + \bar{C}\bar{D}\bar{C}^T &= 0, \\
\bar{C}\bar{A} + P\bar{B}^T S + 2\bar{C}\bar{D}S &= 0, \\
S\bar{A} + \bar{A}^T S + 2S\bar{D}S &= 0,
\end{aligned}
$$

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

Examples revisited

- Overdamped Langevin: Taking $S = 0$ we have that $\pi(x) \propto \exp(-f(x))$
- Underdamped Langevin: Taking

$$S = \frac{1}{c}\begin{bmatrix} I_d & 0 \\ 0 & 0 \end{bmatrix}$$

we have that

$$\pi(x, v) \propto \exp\left(-f(x) + \frac{1}{2c}\|v\|^2\right)$$

Convergence to the invariant distribution I

In order to estimate the quantity of interest, we consider

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

Contractivity implies convergence

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

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

Then, for arbitrary distributions, π_1 and π_2,

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

and, in particular, for arbitrary π,

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

Why contractivity implies convergence?

From the definition of the Wasserstein distance we have

$$
\begin{aligned}
W_P^2(\Phi_t\pi_1, \Phi_t\pi_2) &\leq \mathbb{E}\|\xi^{(2)}(t) - \xi^{(1)}(t)\|_P^2 \quad \text{using contractivity} \\
&\leq e^{-\lambda t}\mathbb{E}\|\xi^{(2)}(0) - \xi^{(1)}(0)\|_P^2 \\
&\leq e^{-\lambda t} W_P^2(\pi_1, \pi_2)
\end{aligned}
$$

Quadratic Lyapunov function

- We have an extended system of equations

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

- We are looking for a quadratic Lyapunov function

$$
V(t) = (\xi_1(t) - \xi_2(t))^T P(\xi_1(t) - \xi_2(t))
$$

Quadratic case

- In the quadratic case $f(x) = \frac{1}{2}x^T Qx$ we consider $Z = \xi_1(t) - \xi_2(t)$ then
    1. $\dot{Z}(t) = \left(\bar{A} + \bar{B}Q\bar{C}\right) Z(t)$
    2. So everything is back to optimization teritory

Non-linear case

- 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 $\lambda = 4/(\kappa+1)$

Overview

Optimization

Discrete time

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

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
$$x_{k+1} = x_k - \alpha\nabla f(x_k).$$
2. For $\gamma = \beta$ we recover the Nesterov method.
3. For $\gamma = 0$, $\beta \neq 0$ we recover the heavy ball method.

Quadratic case

- The continuous time formulation now becomes

$$\xi_{k+1} = (A + BQC)\xi_k$$

- Solution is given by

$$\xi_k = (A + BQC)^k \xi(0)$$

- To deduce a convergence rate to the minimizer we need to understand the spectral properties of $(A + BQC)$

Quadratic case: Convergence rates

$$\|\xi_k - \xi^*\|^2 \leq \rho^{2k} \|\xi_0 - \xi^*\|^2$$

1. Gradient descent: $\alpha = \frac{2}{m+L}$, and $\rho = \frac{\kappa-1}{\kappa+1}$
2. Nesterov method: $\alpha = \frac{4}{3L+m}$, $\beta = \frac{\sqrt{3\kappa+1}-2}{\sqrt{3\kappa+1}+2}$, and $\rho = 1 - \frac{2}{\sqrt{3\kappa+1}}$
3. Heavy ball: $\alpha = \frac{4}{(\sqrt{L}+\sqrt{m})^2}$, $\beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2$, and $\rho = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$

(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 $f(x^\star)$.

THE UNIVERSITY of EDINBURGH

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 $f(x_k) - f(x^\star) \leq \frac{a_0(f(x_0) - f(x^\star)) + (\xi_0 - \xi^\star)^{\mathcal{T}} P (\xi_0 - \xi^\star)}{a_0} \rho^{2k}$.

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

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

decreases monotonically, which, in particular, implies

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

with

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

Sampling

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,

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

Then, for arbitrary distributions, π_1 and π_2,

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

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. $W_{P_h}(\Psi_{h,n+1}\pi, \pi^*) \leq \underbrace{W_{P_h}(\Psi_h(\Psi_{h,n}\pi), \Psi_h\pi^*)}_{\text{numerical contraction}} + \underbrace{W_{P_h}(\Psi_h\pi^*, \Phi_h\pi^*)}_{\text{local error}}$
  2. $W_{P_h}(\Psi_{h,n+1}\pi, \pi^*) \leq \underbrace{W_{P_h}(\Phi_h(\Psi_{h,n}\pi), \Phi_h\pi^*)}_{\text{SDE contraction}} + \underbrace{W_{P_h}(\Psi_h(\Psi_{h,n}\pi), \Phi_h(\Psi_{h,n}\pi))}_{\text{local error}}$
- 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.
- There is potentially a third way of obtaining an error decomposition, by characterising directly the difference between the true and the numerical invariant measure

[4] A. Durmus, S. Majewski, B. Miasojedow, *J. Mach. Learn. Res.*, 20, 1–46. (2019)
[5] A. Durmus, A. Eberle, *arXiv:2108.00682* , (2021)

Strong local error expansion

Assumption

There is a decomposition

$$\phi_h(\widehat{\xi}_n, t_n) - \psi_h(\widehat{\xi}_n, t_n) = \alpha_h(\widehat{\xi}_n, t_n) + \beta_h(\widehat{\xi}_n, t_n),$$

and positive constants p, h_0, C_0, C_1, C_2 such that for $n \geq 0$ and $h \leq h_0$:

$$\left| \left\langle \psi_h(\widehat{\xi}_n, t_n) - \psi_h(\xi_n, t_n), \alpha_h(\widehat{\xi}_n, t_n) \right\rangle_{L^2, P_h} \right| \leq C_0 h \, \|\widehat{\xi}_n - \xi_n\|_{L^2, P_h} \, \|\alpha_h(\widehat{\xi}_n, t_n)\|_{L^2, P_h}$$

and

$$\|\alpha_h(\widehat{\xi}_n, t_n)\|_{L^2, P_h} \leq C_1 h^{p+1/2}, \qquad \|\beta_h(\widehat{\xi}_n, t_n)\|_{L^2, P_h} \leq C_2 h^{p+1}.$$

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 $W_{P_h}(\Psi_{h,n}\pi, \pi^*) < \epsilon$

[6] A. S. Dalalyan, COLT2017
[7] A. S. Dalalyan and A. Karagulyan, *Stoch. Proc. Appl,* 129(12):5278–5311, (2019).
[8] 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 $\mathcal{O}(\epsilon^{-1}\kappa^2 d^{1/2})$

[9] 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

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

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

Overview

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

Optimization algorithms as integrators

$$\frac{d}{dt}z = g^{[1]}(z)+g^{[2]}(z)+g^{[3]}(z) := \begin{bmatrix} -\bar{b}\sqrt{m}v \\ 0 \end{bmatrix} + \begin{bmatrix} -\frac{1}{\sqrt{m}}\nabla f(x) \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \sqrt{m}v \end{bmatrix};$$

Nesterov method can be expressed as

$$\begin{aligned}
Z_{k,1} &= z_k, \\
Z_{k,2} &= z_k + hg^{[1]}(Z_{k,1}), \\
Z_{k,3} &= z_k + hg^{[1]}(Z_{k,1}) + hg^{[3]}(Z_{k,2}), \\
Z_{k,4} &= z_k + hg^{[1]}(Z_{k,1}) + hg^{[3]}(Z_{k,2}) + hg^{[2]}(Z_{k,3}), \\
z_{k+1} &= z_k + hg^{[1]}(Z_{k,1}) + hg^{[2]}(Z_{k,3}) + hg^{[3]}(Z_{k,4}).
\end{aligned}$$

THE UNIVERSITY of EDINBURGH

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 $h \leq cL^{-1/2}$
4. What is special about Nesterov?

Structural conditions of integrators

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

- Key quantity $c := t_{11}/(m\delta)$, when $\gamma = 0$, $c = \cdots + \delta(\kappa - 1)\beta^2/2$.
- For acceleration, δ has to be $\mathcal{O}(1/\sqrt{\kappa})$ which makes it impossible for c to be ≤ 0.
- Presence of κ in t_{11} relates to the appearance of L in the matrix $N^{(1)}$
- This can be indeed eliminated if $EA - C = 0$
- In words: the point $y_k = C\xi_k$ where the gradient is evaluated has to coincide with the point $x_{k+1} = EA\xi_k$ that the algorithm would yield if $u_k = \nabla f(y_k)$ happened to vanish

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

Revisiting the Lyapunov function

$$V(\xi,t) = e^{\lambda t}\left(f(y(t)) - f(y^\star) + (\xi(t)-\xi^\star)^{\mathcal{T}}\bar{P}(\xi(t)-\xi^\star)\right)$$

- We can try to relax the condition $\bar{P} \succeq 0$
- Through strong convexity we know that

$$f(y(t)) - f(y^\star) \geq \frac{m}{2}\|y(t)-y^\star\|^2.$$

- Hence

$$V(\xi,t) \geq e^{\lambda t}\left[(\xi(t)-\xi^\star)^{\mathcal{T}}\left(\frac{m}{2}\bar{C}^T\bar{C} + \bar{P}\right)(\xi(t)-\xi^\star)\right]$$

- If we can still establish that $V(\xi,t)$ is non-increasing we are good as long $\bar{C}^T\bar{C} + \bar{P} \succeq 0$

THE UNIVERSITY of EDINBURGH

Continuous case revisited

Improved (continuous) convergence to minimizer

Suppose that there exist $\lambda > 0$, $\sigma \geq 0$ and a symmetric matrix $\bar{P}$ with $\widetilde{P} := \bar{P} + (m/2)\bar{C}^T\bar{C} \succ 0$, that satisfy

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

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

$$\|y(t) - y_*\|^2 \leq \max\sigma(\bar{C}^T\bar{C})\,\|\xi(t) - \xi^*\|_{\widetilde{P}} \leq \frac{\max\sigma(C^T C)}{\min\sigma(\widetilde{P})} e^{-\lambda t} V(\xi(0), 0).$$

Discrete case revisited

Improved (discrete) convergence to minimizer

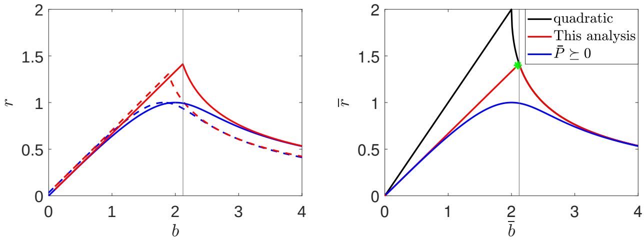
Suppose that there exist $a_0 > 0$, $\rho \in (0, 1)$, $\ell > 0$, and a symmetric matrix P, with $\widetilde{P} := P + (a_0 m/2)E^T E \succ 0$, such that

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

Then, for $f \in \mathcal{F}_{m,L}$, the sequence $\{x_k\}$ satisfies

$$\|x_k - x_\star\|^2 \leq \max\sigma(E^T E)\,\|\xi_k - \xi^\star\|_{\widetilde{P}} \leq \frac{\max\sigma(E^T E)}{\min\sigma(\widetilde{P})} V(\xi_0, 0)\rho^{2k}.$$

What do we gain?



- We can show that in continuous time for $\bar{b} = 3\sqrt{2}/2$ we can improve the convergence rate to $\lambda = \sqrt{2}\sqrt{m}$
- In the discrete setting for appropriate choice of the coefficients we can prove a convergence rate $\rho^2 = 1 - \frac{\sqrt{2}}{\sqrt{\kappa}} + \mathcal{O}(\kappa^{-1}), \quad \kappa \to \infty.$
- The convergence rate of Nesterov with the standard parameter choices $\alpha = L^{-1}, \beta = (\sqrt{k} - 1)/(\sqrt{k} + 1)$ is better that what previously proven.

Overview

THE UNIVERSITY of EDINBURGH

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 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.

4 A. Durmus, S. Majewski and B. Miasojedow. Analysis of Langevin Monte Carlo via Convex Optimization, *J. Mach. Learn. Res.*, 20, 1–46, 2019.

5 A. Durmus and A. Eberle. Asymptotic bias of inexact Markov Chain Monte Carlo methods in high dimension, arXiv:2108.00682, 2021

6 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.

7 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.

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

9 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.

10 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.