Lecture 1

Stochastic Approximation

Vivek Borkar

IIT BOMBAY

January 2022

INTRODUCTION TO STOCHASTIC APPROXIMATION

Problem: Solve $h(x) = 0$ given noisy measurements of h, i.e., given access to a black box that, on input $x \in \mathcal{R}^d$, gives as output $h(x)+$ noise.

Robbins-Monro algorithm: Starting with $x_0 \in \mathcal{R}^d$, do:

$$x(n+1) = x(n) + a(n)[h(x(n)) + M(n+1)], \; n \geq 0.$$

Here the stepsize sequence (or 'learning parameter') $\{a(n)\}$ satisfies: $a(n) \geq 0$ and

$$\sum_n a(n) = \infty, \; \sum_N a(n)^2 < \infty.$$

($\Longrightarrow$ slow decrease to zero, e.g., $\frac{1}{n}$, $\frac{1}{n \log n}$, $\frac{1}{n^{2/3}}$ etc.).

1. $h : \mathcal{R}^d \mapsto \mathcal{R}^d$ is Lipschitz: $\|h(x) - h(y)\| \leq L\|x - y\|$ for $x, y \in \mathcal{R}^d$.

2. $\{M(n)\}$ a square-integrable martingale difference sequence, i.e., for

$$\mathcal{F}_n := \sigma(x_0, M_m, m \leq n), n \geq 0,$$

we have

$$E\left[\|M(n)\|^2\right] < \infty$$

and in addition, it is ‘uncorrelated with past’, i.e.,

$$E[M_i(n+1)|\mathcal{F}_n] = 0 \ \forall \ i.$$

. (Equivalently,

$$E[M_i(n+1)|x_0, M_m, m \leq n] = 0 \ \forall \ i.)$$

. Thus

$$\begin{aligned}
&E[M_i(n+1)f(x_0, M_1, \cdots, M_n)] \\
&\quad = \ E[E[M_i(n+1)f(x_0, M_1, \cdots, M_n)|x_0, M_m, m \leq n]] \\
&\quad = \ E[E[M_i(n+1)|x_0, M_m, m \leq n]f(x_0, M_1, \cdots, M_n)] \\
&\quad = \ 0.
\end{aligned}$$

Hence ‘uncorrelated with past’.

Furthermore, we assume that for some $K > 0$,

$$E\left[\|M(n+1)\|^2|\mathcal{F}_n\right] \leq K\left(1+\|x(n)\|^2\right) \ \forall\ n \geq 0.$$

(equivalently,

$$E\left[\|M(n+1)\|^2|x_0, M_m, m \leq n\right] \leq K\left(1+\|x(n)\|^2\right) \ \forall\ n \geq 0.)$$

In particular, if

$$\sup_m \|x(n)\| < \infty \text{ a.s.,}$$

we have

$$\sup_n E\left[\|M(n+1)\|^2|\mathcal{F}_n\right] < \infty \text{ a.s.}$$

This is more general than it appears. Suppose the algorithm is

$$x(n+1) = x(n) + a(n) f(x(n), \xi(n+1)), \ n \geq 0,$$

where $\{\xi(n)\}$ are IID. This is often how many recursive algorithms are stated.

This can be put in the above form by letting

$$h(x) = E[f(x, \xi(n)] = E\left[f(x(n), \xi(n+1)) | x(n) = x\right]$$

$$= E\left[f(x(n), \xi(n+1) | \mathcal{F}_n\right],$$

$$M(n+1) = f(x(n), \xi(n+1)) - h(x(n)).$$

Examples: Stochastic Gradient Descent ($h = -\nabla f$),

reinforcement learning algorithms (more later)

Highlights:

1. Typically small amount of computation and memory requirements per iterate

2. Incremental: makes a small change in the current iterate at each step

3. Slowly decreasing stepsize captures ‘exploration’ ($\approx$ large steps initially) vs ‘exploitation’ ($\approx$ small steps later) trade-off

4. Averages out the noise (can be thought of as a generalization of the Strong Law of Large Numbers)

1.-3. typical of adaptive behavior $\Longrightarrow$ extremely well suited for adaptive algorithms or models of adaptation

One of the two main workhorses of statistical computation, MCMC being the other.

Applications:

statistics, signal processing, machine learning, adaptive control and communications

Also for models of learning, bounded rationality, herding behavior, etc.

Classical approach for analysis: uses ‘almost supermartingales’ etc. (Robbins-Siegmund, $\cdots$)

Alternative approach: ODE (Ordinary Differential Equations) approach (Meerkov ’72, Derevetskii-Fradkov ‘74, Ljung ‘77)

ODE approach: Treat the iterates as a noisy discretization of the ODE

$$\dot{x}(t) = h(x(t)).$$

Recall the Euler scheme for this ODE:

$$x(n+1) = x(n) + ah(x(n)), \ n \geq 0,$$

where $a > 0$ is a small discrete time step.

Then SA can be viewed as an Euler scheme to approximate the ODE with slowly decreasing time steps $\{a(n)\}$ and measurement noise.

Robbins-Monro conditions:

$\Sigma_n a(n) = \infty \Longrightarrow$ the entire time axis is covered. This is essential because we want to track the asymptotic (as $t \uparrow \infty$) behavior of the ODE.

$\Sigma_n a(n)^2 < \infty \Longrightarrow$ the approximation of the ODE gets better with time:
$a(n) \to 0$ ensures that errors due to discretization are asymptotically zero
$\Sigma_n a(n)^2 < \infty$ ensures that errors due to the martingale difference noise are asymptotically zero, a.s. (multiplication by $a(n)$ reduces the (conditional) variance of noise)

Advantages:

1. Once you have mastered the approach, you can often write the limiting ODE by inspection and analyze it.

2. Designing algorithms: any convergent ODE is a template for an algorithm.

3. Finer dynamic phenomena lead to useful results, e.g., avoidance of unstable equilibria a.s. $\Longrightarrow$ avoidance of ‘traps’ (undesirable equilibria)

Analogy with SLLN suggests related results for fluctuations, e.g., central limit theorem, law of iterated logarithms, concentration inequalities

Further issues and variations:

stability tests, multiple timescales, distributed and asynchronous implementations, differential inclusion limits, constant stepsizes, other noise models, etc.

Example: Nonlinear urns

Consider an initially empty urn to which one ball, either red or blue, is added at each time.

Let $\xi(n+1) = 1$ if $(n+1)$st ball is red, $= 0$ if not (i.e., $\xi(n) = I\{n\text{th ball is red}\}$).

Let $S(n) := \sum_{m=1}^{n} \xi(m)$ the total number of red balls at time n and,

$x(n) := \frac{S(n)}{n}$ the fraction of red balls at time n.

Then

$$
\begin{aligned}
x(n+1) &= \frac{\sum_{m=1}^{n+1} \xi(m)}{n+1} \\
&= \frac{\sum_{m=1}^{n} \xi(m)}{n+1} + \frac{\xi(n+1)}{n+1} \\
&= \left(\frac{n}{n+1}\right) \frac{\sum_{m=1}^{n} \xi(m)}{n} + \frac{\xi(n+1)}{n+1} \\
&= \left(1 - \frac{1}{n+1}\right) x(n) + \frac{1}{n+1}\xi(n+1) \\
&= x(n) + a(n)(\xi(n+1) - x(n))
\end{aligned}
$$

for $a(n) := \frac{1}{n+1}$ which satisfies the Robbins-Monro conditions $\sum_n a(n) = \infty$ and $\sum_n a(n)^2 < \infty$.

Suppose

$$P(\xi(n+1) = 1|\xi(m), m \leq n) = p(x(n))$$

for some continuously differentiable $p(\cdot) : [0,1] \mapsto [0,1]$. Then

$$\begin{aligned} x(n+1) &= x(n) + a(n)(\xi(n+1) - x(n)) \\ &= x(n) + a(n)[(p(x(n)) - x(n)) + \\ &\quad (\xi(n+1) - p(x(n)))] \\ &= x(n) + a(n)[h(x(n)) + M(n+1)] \end{aligned}$$

for $h(x) := p(x) - x$ and $M(n+1) := \xi(n+1) - p(x(n))$.

Since $E[\xi(n+1)|\xi(m), m \leq n] = p(x(n))\ \forall n$, $\{M(n)\}$ is a martingale difference sequence.

Since $|M(n)| \leq 2$, the bound on conditional second moments is free.

The limiting ODE is

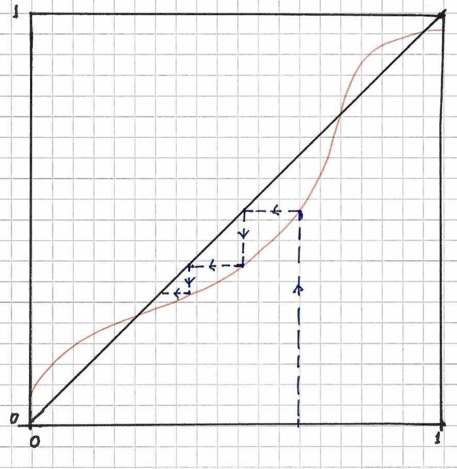
$$\dot{x}(t) = p(x(t)) - x(t).$$

Under our hypothesis of continuous differentiability of $p(\cdot)$, this has a unique solution for any initial condition. Set $x(0) = x_0 \in [0, 1]$.

Since $p(x) - x$ is ≥ 0 for $x = 0$ and ≤ 0 for $x = 0$, $x(t) \in [0, 1]\ \forall t \geq 0$.

Then $x(t)$ must converge to a point in $[0, 1]$: At x_0, if $p(x_0) = x_0$, it is already at an equilibrium. If not, suppose, without loss of generality, that $p(x_0) > x_0$. Then $x(t)$ is increasing, but is bounded by 1, so it must converge. Similar argument works for $p(x_0) < x_0$.

But does an equilibrium exist? Yes, because $p(0) - 0 \geq 0$ and $p(1) - 1 \leq 0$, so by continuity, there is at least one x where $p(x) = x$. (Special case of Brouwer fixed point theorem.)


1
0
0
1

There can be more than one equilibria. An equilibrium x^* satisfies $p(x^*) = x^*$ and is stable if $p'(x^*) < 1$ and unstable if $p'(x^*) > 1$.

Under additional technicalities, can show that $x_n \to$ one of the stable equilibria a.s., and the probability of convergence to any stable equilibrium is strictly positive.

Model for increasing returns in economics, also for herding behavior, technology adoption, and some exotica, such as ‘counterclockwise clocks’, ‘how horses’ behinds dictated the design of space shuttle’, etc.

Suppose $\{\xi(n)\}$ are IID and square-integrable with mean μ. Then

$$x(n+1) = x(n) + a(n)[\mu - x(n) + M(n+1)], \; n \geq 0,$$

where $M(n) = \xi(n) - \mu$ are IID zero mean.

The limiting ODE is

$$\dot{x}(t) = \mu - x(t) \Longrightarrow x(t) = e^{-\mu t}x_0 + \left(1 - e^{-\mu t}\right)\mu \rightarrow \mu.$$

Hence $x(n) \rightarrow \mu$ a.s., which is the Strong Law of Large Numbers. This gives a ‘cheap’ proof of SLLN if you are familiar with this approach.

A TUTORIAL ON O.D.E.

We shall consider the ODE in $\mathcal{R}^d, d \geq 1$, given by

$$\dot{x}(t) = h(x(t)),\; x(0) = x_0.$$

1. Well-posedness (Jaques Hadamard)

(This part extends to non-autonomous systems: $\dot{x}(t) = h(x(t), t)$.)

2. Qualitative theory

A problem is said to be well-posed if

1. it has a solution,

2. the solution is unique, and,

3. the solution depends continuously on problem parameters.

For ODEs, this means that the ODE should have a unique solution *for all time*, that depends continuously on the initial condition.

Examples:

1. $\dot{x}(t) = 2\sqrt{x(t)}$, $x(0) = 0$. This has solutions $x(t) = t^2$ and $x(t) \equiv 0$.

2. $\dot{x}(t) = x^2(t)$, $x(0) = 1$. This has a unique solution $x(t) = \frac{1}{1-t}$ for $0 \le t < 1$ which blows up as $t \uparrow 1$.

In fact it is possible to have uncountably many solutions for every initial condition.

$h : \mathcal{R}^d \mapsto \mathcal{R}^d$ satisfies a (global) Lipschitz condition if for some $L > 0$,

$$\|h(x) - h(y)\| \leq L\|x - y\| \ \forall \ x, y \in \mathcal{R}^d.$$

It is locally Lipschitz if $\forall \ R > 0$, there exists an $L_R > 0$ such that

$$\|h(x)-h(y)\| \leq L_R\|x-y\| \ \forall \ x, y \in B_R := \{z \in \mathcal{R}^d : \|z\| \leq R\}$$

for some $L > 0$.

Theorem If h is Lipschitz, the ODE $\dot{x}(t) = h(x(t))$, $x(0) = \hat{x}$, is well-posed.

Existence: Fix $T \in (0, L)$ and a continuous function $x_0(\cdot) : [0, T] \mapsto \mathcal{R}^d$ with $x_0(0) = \hat{x}$. Recursively define

$$x_{n+1}(t) := \hat{x} + \int_0^t h(x_n(s))ds, \ \ t \in [0, T]. \qquad (\dagger)$$

(Picard iterations) Then for $n \geq 1$,

$$\begin{aligned} \|x_{n+1}(t) - x_n(t)\| &= \|\int_0^t (h(x_n(s)) - h(x_{n-1}(s)))ds\| \\ &\leq \int_0^t \|h(x_n(s)) - h(x_{n-1}(s))\|ds \\ &\leq L \int_0^t \|x_n(s) - x_{n-1}(s)\|ds \\ &\leq LT \max_{t \in [0,T]} \|x_n(t) - x_{n-1}(t)\|. \end{aligned}$$

Thus

$$\max_{t \in [0,T]} \|x_{n+1}(t) - x_n(t)\| \leq LT \max_{t \in [0,T]} \|x_n(t) - x_{n-1}(t)\|.$$

Hence

$$\max_{t\in[0,T]} \|x_{n+1}(t)-x_n(t)\| \leq (LT)^n \max_{t\in[0,T]} \|x_1(t)-x_0(t)\|, \ n \geq 0,$$

$$\Longrightarrow \sum_{n=0}^{\infty} \max_{t\in[0,T]} \|x_{n+1}(t) - x_n(t)\| < \infty.$$

Thus $x_n(t) = x_0(t) + \sum_{m=0}^{n-1}(x_{m+1}(t) - x_m(t))$ converges to some $x(t)$ uniformly in $t \in [0, T]$.

Passing to the limit as $n \uparrow \infty$ in $(\dagger)$,

$$x(t) = \hat{x} + \int_0^t h(x(s))ds, \ \ t \in [0, T].$$

Then $x(\cdot)$ satisfies the ODE with $x(0) = \hat{x}$. Repeat for $[T, 2T], [2T, 3T], \cdots$.

Uniqueness: This needs Gronwall inequality:

Suppose $0 \leq y(\cdot) : [0, T] \mapsto \mathcal{R}$ is differentiable and satisfies

$$y(t) \leq C + K \int_0^t y(s)ds, \ \ t \in [0, T],$$

for some $C, K > 0$. Then

$$y(t) \leq Ce^{KT}, \ \ t \in [0, T].$$

Proof Let $z(t) := \int_0^t y(s)ds, \ \ t \geq 0$. Then

$$\dot{z}(t) = y(t) \leq C + Kz(t)$$

$$\Longrightarrow e^{-Kt}(\dot{z}(t) - Kz(t)) \leq Ce^{-Kt}$$

$$\Longrightarrow \frac{d}{dt}\left(e^{-Kt}z(t)\right) \leq Ce^{-Kt},\ z(0) = 0.$$

Integrating both sides from 0 to t, $t \in [0, T]$,

$$e^{-Kt}z(t) \leq \frac{C}{K}\left(1 - e^{-Kt}\right)$$

$$\Longrightarrow z(t) \leq \frac{C}{K}\left(e^{Kt} - 1\right)$$

$$\Longrightarrow y(t) \leq C + Kz(t)$$

$$\leq C + C\left(e^{Kt} - 1\right) = Ce^{Kt}.$$

□

Consider $\dot{x}(t) = h(x(t))$, $\dot{y}(t) = h(y(t))$, $t \geq 0$, with $x(0) = y(0)$. Then

$$\|x(t)-y(t)\| \leq L\int_0^t \|x(s)-y(s)\|ds \Longrightarrow \|x(t)-y(t)\| = 0 \ \forall\, t \geq 0$$

by the Gronwall inequality, implying uniqueness.

In general, for $x(0) = x, y(0) = y$,

$$\|x(t) - y(t)\| \leq \|x - y\| + L\int_0^t \|x(s) - y(s)\|ds \Longrightarrow$$

$$\|x(t) - y(t)\| \leq e^{Lt}\|x - y\| \ \forall\ t \geq 0$$

by the Gronwall inequality, implying continuous dependence on initial condition.

Hence the ODE is well-posed.

Remarks: 1. Picard iteration is not a good computational scheme. Euler scheme is the most basic choice. Suppose h is bounded. Let $a := \frac{T}{N}, N >> 1$, let

$$X_N((n+1)a) := X_N(na) + ah(X_N(na)),\ 0 \leq n < N.$$

Interpolate linearly:

$$X_N(t) := X_N(na) + (t - na)h(X_N(na)),\ t \in [na, (n+1)a].$$

Then as $N \uparrow \infty$, $X_N(t), t \in [0, T]$, converges to a solution of the ODE uniformly on $[0, T]$. This too proves existence of a solution and needs only continuity of h. But uniqueness may fail. In numerical analysis, more sophisticated discretizations are available.

Discretization matters! Algorithms, deterministic or stochastic, can often be viewed as discretizations of ODEs.

Given ODE $\dot{x}(t) = h(x(t))$, I can always 'speed it up' by replacing it by $\dot{x}(t) = \alpha h(x(t))$ with $\alpha >> 1$. To see this, let

$$\tau = t/\alpha, y(t) = x(\tau(t)).$$

Then

$$\dot{y} = \frac{d}{d\tau}x(\tau)\frac{d\tau(t)}{dt} = \alpha h(x(\tau(t)))\frac{1}{\alpha} = h(y(\tau)).$$

Thus this amounts to a pure time scaling that does not change the trajectory, but changes the speed with which it is traversed.

However, this is not so with discretizations.

Discretizations that preserve the ‘physics’ of the differential equation (e.g., conserved quantities) matter in physics. Often they lead to better algorithms (cf. the work of Ashia Wilson).

2. Local Lipschitz condition gives local well-posedness (i.e., existence, uniqueness, continuous dependence on the initial condition) for a small time interval, but the solution may not exist for all time.

3. ‘Linear growth condition’ below suffices for a solution to exist for all time:

$$\|h(x)\| \leq K(1 + \|x\|)$$

for some $K >$. Then

$$\|x(t)\| \leq \|x(0)\| + \| \int_0^t h(x(s))ds\| \leq \|x(0)\| + K \int_0^t (1 + \|x(s)\|)ds$$

$$\Longrightarrow \|x(t)\| \leq (\|x(0)\| + KT)e^{KT}, \; t \in [0, T],$$

by Gronwall inequality. Note that Lipschitz condition implies linear growth, because for a fixed $x_0 \in \mathcal{R}^d$,

$$\|h(x) - h(y)\| \leq L\|x - y\| \Longrightarrow$$

$$\|h(x)\| \leq \|h(x_0)\| + \|h(x) - h(x_0)\|$$

$$\leq \|h(x_0)\| + L\|x - x_0\|$$

$$\leq (\|h(x_0)\| + L\|x_0\|) + L\|x\|.$$

4. A symmetric well-posedness theory can be developed for $t \leq 0$. Thus, e.g., for Lipschitz h, there is a unique solution for all $t \in \mathcal{R}$.

5. 'Discrete Gronwall inequality' also holds and is proved similarly: Let $x_n \geq 0, a_n > 0, n \geq 0$, and $C, K > 0$ such that

$$x_{n+1} \leq C + K \sum_{m=0}^{n} a_m x_m \ \forall\ n \geq 0.$$

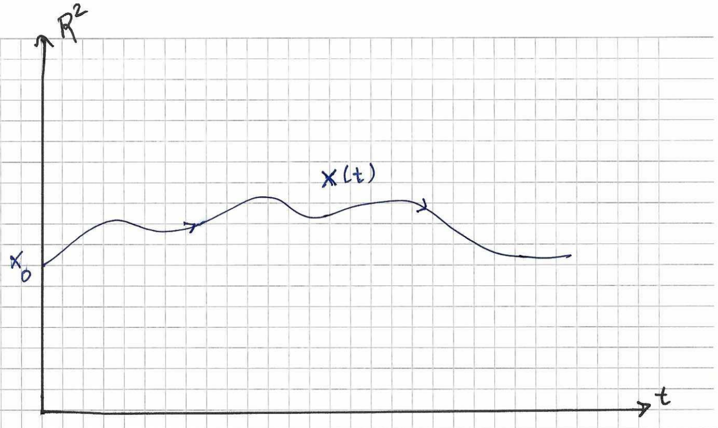
Then $x_{n+1} \leq Ce^{K \sum_{m=0}^{n} a_m}$.

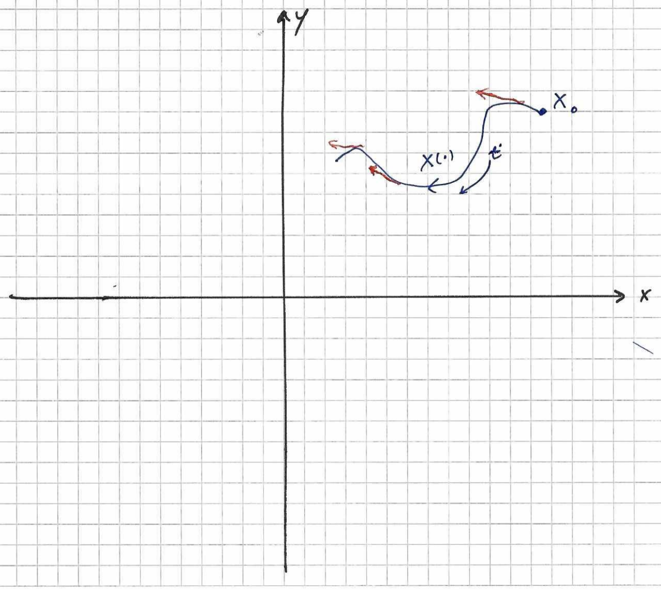
We shall be using this later.

Qualitative theory: Assume well-posedness. There are two ways of thinking of ODEs:

1. Graph of $t \mapsto x(t) \in \mathcal{R}^d$, i.e., $x(t)$ as a function of time t. The componentwise time derivative at t is $h(x(t))$.

2. Trajectories $x(\cdot)$ as curves in $\mathcal{R}^d$ with t as a running parameter (phase portrait). The tangent at point x on the curve is $h(x)$. Often one flips this picture around to imagine vectors $h(x)$ at each point x in space, i.e., a 'vector field', and think of the trajectories as curves that are drawn so as to be tangent to the vector field at all points, i.e., 'integral curves'.


R^2
$x(t)$
x_0
t


y
x_0
$x(\cdot)$
t
x

The ‘qualitative theory of differential equations’ studies what the phase portraits look like.

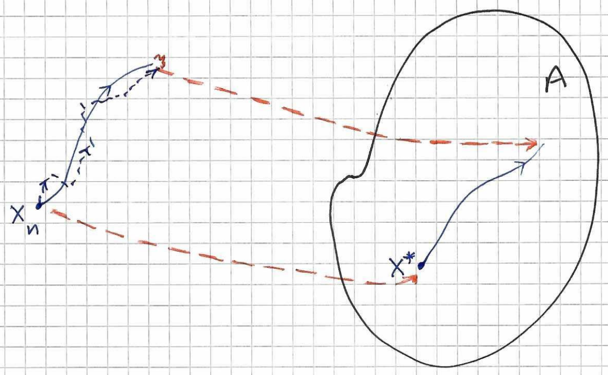
Some important concepts in this are:

The ω-limit set of a trajectory $x(\cdot)$ is the set of all points x such that $\exists\ t_n \uparrow \infty$ (depending on x) such that $x(t_n) \to x$, i.e., the set of limit points of $x(t)$ as $t \uparrow \infty$. One can show that this set is closed (because ‘limit point of limit points is a limit point’), but can be empty (e.g., for $\dot{x}(t) = 1$). The α-limit set is defined similarly for $t_n \downarrow -\infty$.

A set $A \subset \mathcal{R}^d$ is said to be positively invariant if $x(0) \in A \Longrightarrow x(t) \in A \; \forall \; t \geq 0$. Negative invariance is defined similarly. A set both positively and negatively invariant is said to be invariant.

If $h(x) = 0$, then $x(t) = x \; \forall t$ is the unique solution through x and x is then said to be an equilibrium. A periodic solution is said to be a limit cycle. Both form invariant sets.

The ω- and α-limit sets are invariant: if $x(t_n) \to x$ and $\tilde{x}(\cdot)$ is the unique trajectory through x, then $x(t_n + t) \to \tilde{x}(t)$ for $t \in \mathcal{R}$. $\mathcal{R}^d$ is trivially invariant.

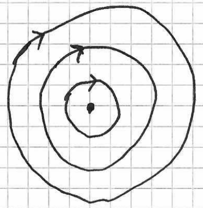

A
X
n
X*

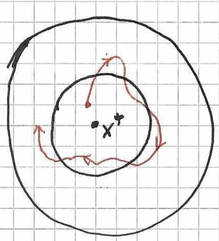
We shall primarily focus on equilibria. An equilibrium x^* is said to be Liapunov stable if given $\epsilon > 0$, there exists a $\delta > 0$ such that

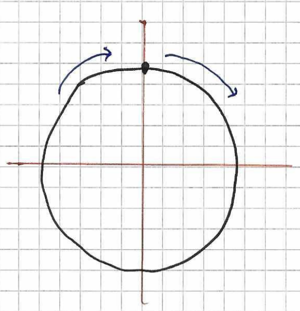
$$\|x(0) - x^*\| < \delta \implies \|x(t) - x^*\| < \epsilon \ \forall \ t \geq 0.$$

That is, ‘trajectories initiated sufficiently near x^* remain near x^*’ (e.g., the harmonic oscillator).

If x^* is Liapunov stable and there exists an open neighbourhood O of x^* such that $x(0) \in O \implies x(t) \to x^*$, then x^* is said to be asymptotically stable. The largest positively invariant open set D such that $x(0) \in D \implies x(t) \to x^*$ is called the domain of attraction of x^*.




x^+



One sufficient condition is that there exist a continuously differentiable $V : D \mapsto [0, \infty)$ such that

$$\lim_{x \to \partial D} V(x) = \infty \text{ and}$$

$$\langle \nabla V(x), h(x) \rangle < 0 \ \forall x \in D, x \neq x^*.$$

$$\Longrightarrow \frac{d}{dt} V(x(t)) = \langle \nabla V(x(t)), h(x(t)) \rangle < 0 \text{ when } x(t) \neq x^*,$$

i.e., $V(x(t))$ is decreasing along the trajectory.

Since $V(\cdot) \geq 0$, $x(t) \to x^*$.

Also, if we consider $B_c(x^*) := \{x : V(x) \leq c\} \subset D$ for a suitable $c > V(x^*)$, then

$$x(0) \in B_c(x^*) \Longrightarrow x(t) \in B_c(x^*) \ \forall t \geq 0.$$

Thus x^* is Liapunov stable. Hence it is asymptotically stable. V is then called a Liapunov function.

Conversely, if x^* is asymptotically stable, such a V exists and may be taken to satisfy $V(x) \to \infty$ as $x \to \partial D$ (Converse Liapunov theorem).

More generally, $x(t) \to$ the largest invariant set contained in $\{x : \langle \nabla V(x), h(x) \rangle = 0\}$. (Lasalle invariance principle)

If there exists a continuously differentiable $V : D \mapsto [0, \infty)$ such that

$$\lim_{\|x\| \uparrow \infty} V(x) = \infty \text{ and}$$

$$\langle \nabla V(x), h(x) \rangle < 0 \ \forall x \notin C$$

for some bounded set C, then

$$\frac{d}{dt} V(x(t)) = \langle \nabla V(x(t)), h(x(t)) \rangle < 0 \text{ when } x(t) \notin C$$

$$\implies x(t) \to C.$$

In particular, the trajectories remain bounded.

Thus, e.g., h locally Lipschitz + above condition holds $\implies$ well-posedness.

Consider the linear system $\dot{x}(t) = Ax(t)$ for some $A \in \mathcal{R}^{n\times n}$.

Then the origin, i.e., the zero vector θ, is an equilibrium.

If A is nonsingular, it is the only equilibrium.

It is asymptotically stable if all eigenvalues of A are in the left half plane.

If not, suppose there are no eigenvalues on the imaginary axis.

If there are $m < n$ eigenvalues in the left half plane, we can write $\mathcal{R}^d = S \oplus U$ where:

$S :=$ the m-dimensional ‘stable subspace’ corresponding to the eigenvectors for eigenvalues in the left half plane, and,

$U :=$ the $(d - m)$-dimensional ‘unstable subspace’ corresponding to the eigenvectors for eigenvalues in the right half plane.

Then $x(0) \in S$ implies $X(t) \to \theta$ and $x(0) \in U$ implies that $x(t)$ moves away from θ.

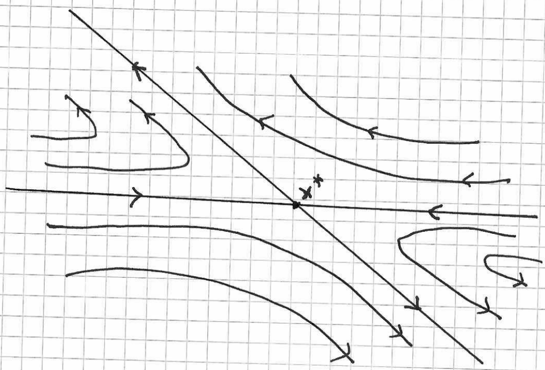
More importantly, if $x(0) \notin S$, $x(t)$ eventually moves away from θ.

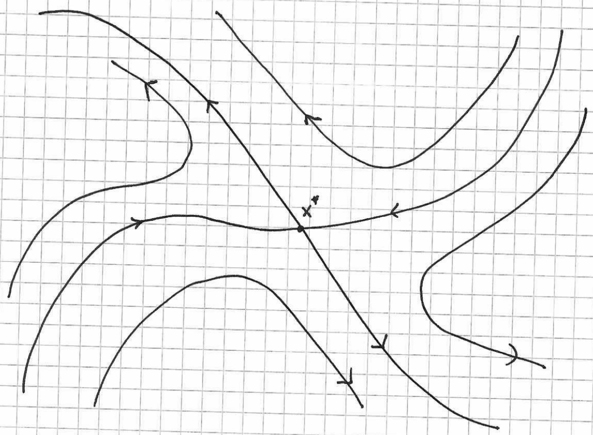
That is, $x(t) \to \theta$ if and only if $X(0) \in S$, which has zero volume in $\mathcal{R}^d$.

In other words, for a typical ('generic') initial condition, $x(t)$ moves away from θ.

The foregoing goes through for θ replaced by some $\widehat{x} \in \mathcal{R}^d$ if we replace the above linear ODE by the *affine* ODE

$$\dot{x}(t) = A(x(t) - \widehat{x}).$$


x*


x*

We now extend these ideas to the nonlinear case.

Assume that h is continuously differentiable and let $Dh(x)$ denote its Jacobian matrix at x, i.e., the matrix whose (i, j)th element is $\frac{\partial h_i}{\partial x_j}(x)$. By Taylor formula, for $x \approx x^*$,

$$h(x) \approx h(x^*) + Dh(x^*)(x - x^*) = Dh(x^*)(x - x^*).$$

Consider the affine ODE

$$\dot{z}(t) = Dh(x^*)(x(t) - x^*).$$

This is the *linearization* of the original ODE at x^*.

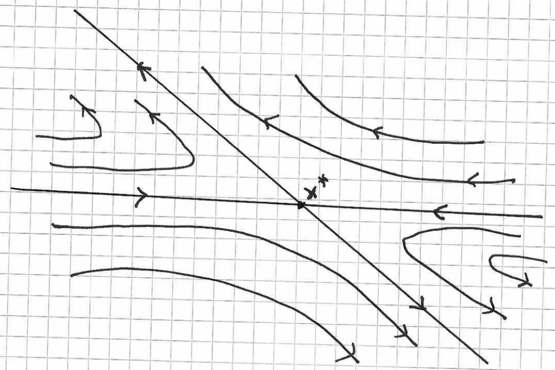
Let x^* be a hyperbolic equilibrium, i.e., $Dh(x^*)$ does not have any eigenvalues on the imaginary axis. Then we have:

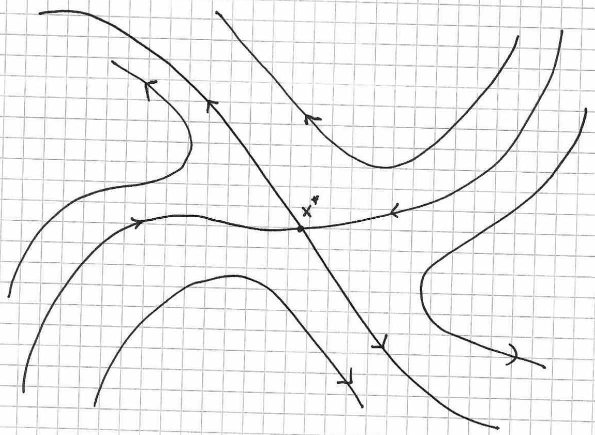
Hartman-Groβman theorem (in words): There exist open neighborhoods O_1, O_2 of x^* such that the phase portrait of the original ODE in O_1 and its linearization in O_2 can be mapped to each other by a continuous and continuously invertible transformation.

Thus ‘stable subspaces’ morph into ‘stable manifolds’, ‘unstable subspaces’ morph into ‘unstable manifolds’.

The important fact remains true: for unstable x^* (i.e., at least one eigenvalue of $Dh(x^*) \in$ the right half plane, 'generically' the trajectories move away from x^*.

'Generic' can be in the topological sense (i.e., $\forall\ x \in$ an open dense set) or measure theoretic sense (i.e., $\forall\ x \notin$ a zero measure set).


x^*


x^*

Hyperbolicity itself is generic if there is no other restriction, because a small perturbation of h will displace imaginary eigenvalues. But the physics of the problem may dictate otherwise, e.g., the harmonic oscillator.

For algorithms, it is mostly a reasonable assumption, but fails, e.g., when there is a continuum of equilibria (say, a line) because of overparametrization.

If it holds, more generally, if $Dh(x^*)$ is non-singular, the equilibria are isolated, because h is one-one in a neighborhood of x^* ('because $x \mapsto Dh(x^*)(x - x^*)$ is').

CONVERGENCE ANALYSIS

- The proof is pathwise, a.s.

- Thus all statements are a.s., even when not explicitly mentioned.

- Some of the constants used in the bounds are random, i.e., sample path dependent.

- Not all details are given.

Our iteration is

$$x(n+1) = x(n) + a(n)\left[h(x(n)) + M(n+1)\right], \;\; n \geq 0.$$

Since we view $a(n)$ as a discrete time step, define the 'algorithmic time scale' by

$$t(0) := 0, \;\; t(n) := \sum_{m=0}^{n-1} a(m), \;\; n \geq 0.$$

Define $\bar{x}(t), t \in [0, \infty)$, by: $\bar{x}(t(n)) = x(n)\ \forall n \geq 0$, and

$$\bar{x}(t) = x(t(n)) + \left(\frac{t - t(n)}{t(n+1) - t(n)}\right)(x(t(n+1)) - x(t(n))),$$

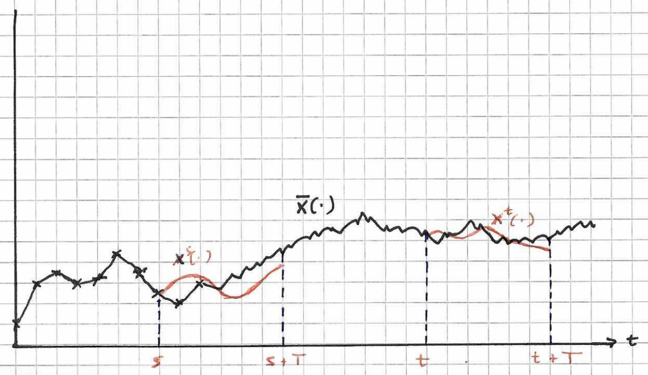
for $t \in [t(n), t(n+1)], n \geq 0$, i.e., a linear interpolation on $[t(n), t(n+1)]$. Then $\bar{x}(\cdot)$ is continuous and piecewise linear.

Assume stability, i.e., that $\sup_{n \geq 0} \|x(n)\| < \infty$ a.s.

Fix $T > 0$. We shall compare $\bar{x}(\cdot)$ on a sliding window $[t, t+T]$ as $t \uparrow \infty$, with the ODE trajectory on the same interval that matches with it at the beginning of the interval, i.e., with $x^t(s), s \in [t, t+T]$, satisfying

$$\dot{x}^t(s) = h(x(s)), \ s \in [t, t+T], \ x^t(t) = \bar{x}(t).$$

It suffices to consider $t = t(n)$ for some $n \geq 0$, which simplifies matters. For the general case, there is a negligible additional error which can be easily handled.


$\bar{x}(\cdot)$
$x^s(\cdot)$
$x^t(\cdot)$
t
s
s + T
t
t + T

Let $m(n) := \min\{k \geq n : t(k) \geq t(n){+}T\}$. Then $t(m(n)) \approx t(n) + T$. We shall compare $\bar{x}(\cdot)$ and $x^{t(n)}(\cdot)$ on the interval $[t(n), t(m(n))]$. Define

$$\zeta(n) = \sum_{m=0}^{n-1} a(m)M(m+1),\ \ n \geq 1.$$

This is a martingale, i.e., $E\left[\zeta(n+1)|\mathcal{F}_n\right] = \zeta(n)\ \forall n$. Also,

$$\sum_n E\left[\|\zeta(n+1) - \zeta(n)\|^2|\mathcal{F}_n\right] = \sum_n a(n)^2 E\left[\|M(n+1)\|^2|\mathcal{F}_n\right]$$

$$\leq \sum_n a(n)^2 K(1{+}\|x(n)\|^2) \leq K(1{+}\sup_n \|x(n)\|^2) \sum_m a(m)^2 < \infty$$

which is $< \infty$ a.s. By convergence theorem for square-integrable martingales, $\zeta(n)$ converges a.s. as $n \uparrow \infty$.

Convergence theorem for square-integrable martingales: Let $(Z_n, \mathcal{F}_n)$ be a square-integrable martingale. Then its quadratic variation process $\langle Z \rangle_n, n \geq 0$, is given by

$$\langle Z \rangle_n := \sum_{m=0}^{n} E\left[(Z_{m+1} - Z_m)^2 | \mathcal{F}_m\right].$$

Theorem Almost surely,

$\lim_{n \uparrow \infty} \langle Z \rangle_n < \infty \Longrightarrow Z_n$ converges.

Generalizes Kolmogorov's result: If $\{X_n\}$ are independent zero mean with bounded variances, then:

$\sum_n X_n$ converges $\Longleftrightarrow \sum_n E\left[X_n^2\right] < \infty$.

For $0 \leq k \leq m(n) - n$,

$$\bar{x}(t(n+k)) = \bar{x}(t(n)) + \sum_{i=0}^{k-1} a(n+i)h(\bar{x}(t(n+i)))$$

$$+ \sum_{\ell=0}^{k-1} a(n+\ell)M(n+\ell+1)$$

$$= \bar{x}(t(n)) + \sum_{i=0}^{k-1} a(n+i)h(\bar{x}(t(n+i)))$$

$$+ \ (\zeta(n+k) - \zeta(n))$$

Also,

$$x^{t(n)}(t(n+k)) = x^{t(n)}(t(n)) + \sum_{i=0}^{k-1} a(n+i)h(x^{t(n)}(t(n+i)))$$

$$+ \sum_{\ell=n}^{n+k-1} \int_{t(\ell)}^{t(\ell+1)} (h(x^{t(n)}(t)) - h(x^{t(n)}(t(\ell))))dt$$

$$= x^{t(n)}(t(n)) + \sum_{i=0}^{k-1} a(n+i)h(x^{t(n)}(t(n+i)))$$

$$+ \int_{t(n)}^{t(n+k)} (h(x^{t(n)}(t)) - h(x^{t(n)}([t])))dt$$

where $[t] := \max\{t(m) : t(m) \leq t < t(m+1)\}$.

Note that $\bar{x}(t(n)) = x^{t(n)}(t(n)) = x(n)$.

Hence

$$\|\bar{x}(t(n+k) - x^{t(n)}(t(n+k))\| \quad \leq$$

$$\sum_{i=0}^{k-1} a(n+i)\|h(\bar{x}(t(n+i))) - h(x^{t(n)}(t(n+i)))\| + \ I + \ II,$$

where $I :=$ the discretization error, $II :=$ the error due to noise. Thus

$$\|\bar{x}(t(n+k) - x^{t(n)}(t(n+k))\| \quad \leq$$

$$L\sum_{i=0}^{k-1} a(n+i)\|\bar{x}(t(n+i))) - x^{t(n)}(t(n+i))\| + \ I + \ II,$$

By the discrete Gronwall inequality, $\exists\ C(T) > 0$ such that

$$\sup_{n \leq m \leq m(n)} \|\bar{x}(t(m)) - x^{t(n)}(t(m))\| \leq C(T)(I + II).$$

For $\infty > K \geq \sup_{t\in[t(n),t(m(n))]} \|h(x^{t(n)}(t))\| > 0$,

$$
\begin{aligned}
& \| \int_{t(m)}^{t(m+1)} (h(x^{t(n)}(t)) - h(x^{t(n)}([t]))dt\| \\
= & \ \| \int_{t(m)}^{t(m+1)} (h(x^{t(n)}(t)) - h(x^{t(n)}(t(m)))dt\| \\
\leq & \ L \int_{t(m)}^{t(m+1)} \|x^{t(n)}(t) - x^{t(n)}(t(m))\| dt \\
= & \ L \int_{t(m)}^{t(m+1)} \left\| \int_{t(m)}^{t} h(x^{t(n)}(s))ds \right\| dt \\
\leq & \ \frac{LK}{2}(t(m+1) - t(m))^2 \\
= & \ L'a(m)^2.
\end{aligned}
$$

Hence

$$I \;=\; \left\| \int_{t(n)}^{t(m(n))} (h(x^{t(n)}(t)) - \bar{h}(x^{t(n)}([t])))dt \right\|$$

$$\leq\; L' \sum_{m \geq n} a(m)^2 \downarrow 0 \;\; \text{as} \;\; n \uparrow \infty.$$

Also, $II \leq \sup_{m \geq 0} \|\zeta_{n+m} - \zeta_n\| \rightarrow 0$ a.s. as $n \uparrow 0$.

Thus a.s., as n$\uparrow \infty$,

$$\max_{n \leq m \leq m(n)} \|\bar{x}(t(m)) - x^{t(n)}(t(m))\| \rightarrow 0 \text{ a.s.} \;\; \Longrightarrow$$

$$\lim_{t \uparrow \infty} \max_{s \in [0,T]} \|\bar{x}(t+s) - x^{t(n)}(t+s)\| \rightarrow 0 \text{ a.s.} \;\; \forall\; T > 0.$$

Let $D := \{x \in \mathcal{R}^d : \exists\; 0 < s(n) \uparrow \infty$ such that $\bar{x}(s(n)) \rightarrow x\}$

$= \{x \in \mathcal{R}^d : \exists\; 0 < t_k \uparrow \infty$ such that $x(t_k) \rightarrow x\}$.

Claim: D is an invariant set for the ODE.

Proof: Suppose $s(n) \uparrow \infty$ and $\bar{x}(s(n)) \to x$. Then $x \in D$. Let $\dot{\tilde{x}}(t) = h(\tilde{x}(t)), \tilde{x}(0) = x$. By the above, for $T > 0$,

$$\bar{x}(s(n) + T) - x^{s(n)}(s(n) + T) \to 0.$$

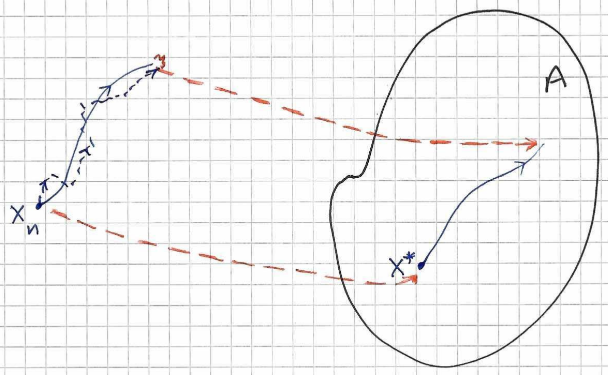
By continuous dependence of ODE on initial condition,

$$x^{s(n)}(s(n)) = \bar{x}(s(n)) \to x \Longrightarrow x^{s(n)}(s(n)+T) - \tilde{x}(T) \to 0.$$

Thus

$$\bar{x}(s(n) + T) - \tilde{x}(T) \to 0,$$

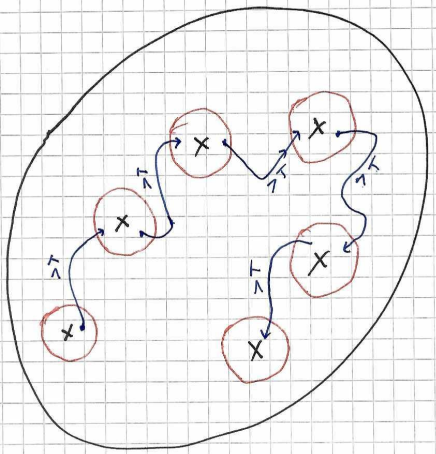
implying $\tilde{x}(T) \in D$. Similar argument works for $T < 0$. Hence D is invariant.


A
X_n
X^*

Say that D is an internally chain transitive invariant set if given any $\epsilon, T > 0$ and points $x, y \in D$, we can find $n \geq 1$ and a chain of points $x_0 = x, x_1, \cdots, x_n = y$, in D such that for $0 \leq i < n$, there exists a trajectory segment of the ODE in D of duration at least T which starts in the ϵ-neighborhood of x_i and ends in the ϵ-neighborhood of x_{i+1}.

Benaim's theorem: Almost surely, $x(n) \to$ an internally chain transitive invariant set of the ODE.

Our interest will be in the case when the only such sets are equilibria.


> T
> T
> T
> T
> T

Minor extensions:

Can add to M_{n+1} an adapted bounded ‘measurement error’ $\varepsilon_n \to 0$ without affecting the claims.

$\|\varepsilon\| < \epsilon \implies x_n \to$ a small neighborhood of the desired limit under reasonable conditions.

‘Adapted’ stepsizes possible, i.e., random $\{a(n)\}$ such that $E[M(n+1)|x(m), M(m), a(m), m \leq n] =$ the zero vector $\forall\ n \geq 0$.

VARIANTS

Two time scale schemes I : Consider the coupled iterations

$$x_{n+1} = x_n + a(n)[h(x_n, y_n) + M_{n+1}],$$

$$y_{n+1} = y_n + b(n)[g(x_n, y_n) + M'(n+1)],$$

with

$$\sum_n a(n) = \sum_n b(n) = \infty, \sum_n (a(n)^2 + b(n)^2) < \infty, \frac{b(n)}{a(n)} \to 0.$$

$\frac{b(n)}{a(n)} \to 0 \implies \{y_n\}$ updated on a slower timescale than $\{x_n\}$.

Consider the ‘algorithmic time scale’ corresponding to time steps $\{a(n)\}$. Then, writing the second iteration as

$$y_{n+1} = y_n + a(n) \left(\frac{b(n)}{a(n)}\right) [g(x_n, y_n) + M'(n+1)],$$

the limiting o.d.e. is

$$\dot{x}(t) = h(x(t), y(t)), \ \dot{y}(t) = 0,$$

i.e., $\{x_n\}$ sees $\{y_n\}$ as ‘quasi-static’ or (nearly) a constant. Hence it tracks the o.d.e.

$$\dot{x}(t) = h(x(t), y)$$

for $y \approx y_n$. Suppose $x(t) \to \lambda(y)$. Then $x_n - \lambda(y_n) \to 0$ a.s. (i.e., $\{y_n\}$ sees $\{x_n\}$ as ‘quasi-equilibrated’).

Rewrite the iteration for $\{y_n\}$ as

$$\begin{aligned} y_{n+1} \;=\; & y_n + b(n)[g(\lambda(y_n), y_n) + \\ & (g(x_n, y_n) - g(\lambda(y_n), y_n)) + M'(n+1)]. \end{aligned}$$

Using the algorithmic time scale defined in terms of $\{b(n)\}$, $\{y_n\}$ tracks

$$\dot{y}(t) = g(\lambda(y(t)), y(t)).$$

If $y(t) \to y^*$, then $(x_n, y_n) \to (\lambda(y^*), y^*)$.

Analogous to ‘singularly perturbed differential equations’:

$$\begin{aligned}\dot{x}(t) &= h(x(t), y(t)),\\ \dot{y}(t) &= \epsilon g(x(t), y(t)),\end{aligned}$$

in the $\epsilon \downarrow 0$ limit.

Emulates nested iterations

(fast iteration ≈ a subroutine).

Two time scale schemes II : 'Markov noise' $\{Y_n\}$ on natural clock ($n = 0, 1, 2, \cdots$):

$$x_{n+1} = x_n + a(n)[h(x_n, Y_n) + M_{n+1}],$$

where

$$P(Y_{n+1} \in A|\mathcal{F}_n) = p_{x_n}(A|Y_n).$$

Let $p_x(\cdot|\cdot)$ be irreducible with unique stationary distribution π_x. Then $\{x_n\}$ tracks the o.d.e.

$$\dot{x}(t) = \int h(x(t), y)\pi_{x(t)}(dy).$$

Intuition: $\{x_n\}$ updated on a slow time scale

$\Longrightarrow Y_n \approx$ a Markov chain governed by $p_{x_n}(\cdot|\cdot)$

$\{x_n\}$ sees $\{Y_n\}$ as 'quasi-equilibrated'

$\Longrightarrow$ the distribution of $Y_n \approx \pi_{x_n}$

$\Longrightarrow$ averaging property of stochastic approximation leads to averaging of $h(x_n, \cdot)$ w.r.t. π_{x_n}.

Two time scale schemes III : Gossip + Learning

(Tsitsiklis-Bertsekas-Athans) Processor i performs the vector iteration

$$x^i_{n+1} = \sum_j p(j|i)x^j_n + a(n)[h^i(x_n) + M^i_{n+1}]$$

where $P := [[p(\cdot|\cdot)]]$ an irreducible stochastic matrix with stationary distribution π.

Important special case: P doubly stochastic

$\Longleftrightarrow \pi$ uniform.

This tracks the o.d.e.

$$\dot{x}^i(t) = \sum_j \pi(j) h^j(x(t)).$$

Also, $\|x_n^i - x_n^j\| \to 0$ s')for $i \neq j$
$\Longrightarrow$ convergence to common limit.

Intuition: Consider scalar iterations for simplicity. The iteration $x^{\cdot}_{n+1} = P x^{\cdot}_n$ is a marginally stable linear system that converges to a point in its invariant subspace of constant vectors.

The point ($\sum_j \pi_j x_0^j$ to be precise) depends on the initial condition.

This effect operating on the fast time scale ends up confining the slower dynamics to the invariant subspace, hence the averaged dynamics.

Extensions: $p(\cdot|\cdot) \to p_x(\cdot|\cdot) \Longrightarrow \pi \to \pi_{x(t)}$.

More generally, replace $\sum_j p(j|i)x_n^j$ by $f_j(x(n)) \Longrightarrow$ confines the o.d.e. to the set of fixed points of $f(\cdot) = [f_1(\cdot), \cdots, f_d(\cdot)]$. Useful for distributed projected stochastic approximation.

Stability test: Suppose $h_\infty(x) := \lim_{c\uparrow\infty} \frac{h(cx)}{c}$ exists and the o.d.e.

$$\dot{x}(t) = h_\infty(x(t))$$

has the origin as its unique globally asymptotically stable equilibrium. Thus $\sup_n \|x(n)\| < \infty$ a.s.

Other tests available.

'Stabilization' possible thorugh projection, resets, tweaking stepsizes, etc.

Intuition: If $\|x_{n(k)}\| \uparrow \infty$, then the rescaled interpolations on $[t(n(k)), t(m(n(k)))]$ given by

$$\check{x}(t) := \frac{\bar{x}(t)}{\|\bar{x}(t(n(k)))\|}, \ t \in [t(n(k)).t(m(n(k)))],$$

track $\dot{x}(t) = h_{\infty}(x(t))$ and hence tend to the origin.

Since $\bar{x}(\cdot)$ and $\check{x}(\cdot)$ differ only by a scale factor, the same holds true for $\bar{x}(\cdot)$.

Hence the iterates cannot blow up.

Constant stepsize: $a(n) \equiv a > 0$

Weaker claims: ‘high probability concentration’ instead of ‘a.s. convergence’: $\limsup_{n\uparrow\infty} E\left[\|x_n - x^*\|^2\right] \leq Ka$.

Useful for tracking slowly varying environments because for decreasing stepsize, the algorithmic time scale eventually becomes slower than the environment and therefore cannot track it. Also useful when the algorithm is hardwired..

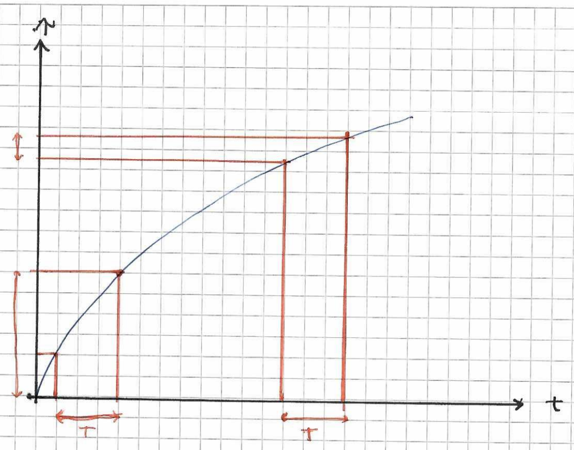
Distributed and asynchronous iterates: Consider the iteration

$$x_{n+1}(i) = x_n(i) + a(n)I\{i \in S_n\}\times$$

$$\Big[h_i(x_{n-\tau_{1i}(n)}(1), \cdots, x_{n-\tau_{di}(n)}(d)) + M_{n+1}(i)\Big].$$

Here $S_n \subset \{1, 2, \cdots, d\}$ is the set of the indices of the components updated at time n, $\tau_{ji}(n) := n -$ the time stamp of the most recent value received by i from j.

For bounded delays (or with a conditional moment bound), the delays do not matter asymptotically because the time scale ($n \to t(n) := \sum_{m=0}^{n} a(m)$) is getting shrunk.


v
t
T
T

Tracks the o.d.e.

$$\dot{x}(t) = \Lambda(t)h(x(t)).$$

$\Lambda(t)$ is a diagonal matrix with non-negative diagonal entries $\lambda_i(t)$ reflecting relative frequencies of update of different components,

e.g., for $S_n = \{X_n\}$ where $\{X_n\}$ is an ergodic Markov chain on $\{1, 2, \cdots, d\}$ with stationary distribution π, the ith diagonal entry is $\pi(i)$. (Example: TD(λ))

Replace $a(n)$ by $a(\nu(i,n))$ where

$$\nu(i,n) = \sum_{m=0}^{n} I\{X_m = i\}$$

is the 'local clock', then under additional conditions on $\{a(n)\}$, $\Lambda(t) \equiv \frac{1}{d} I$. (Need $t(n)$ growing 'logarithmically'.)

For example, the algorithmic time scales for components i and j are resp.

$$\sum_{m=0}^{n} a(\nu(i,m)), \quad \sum_{m=0}^{n} a(\nu(j,m)).$$

For $a(n) = \frac{1}{n+1}$, $\sum_{m=0}^{n} \frac{1}{m+1} \approx \log n$.

Hence, if $\liminf_{n\uparrow\infty}\frac{\nu(i,n)}{n} > 0$ a.s., then

$$\lim_{n\uparrow\infty}\frac{\sum_{m=0}^{n} a(\nu(i,m))}{\sum_{m=0}^{n} a(\nu(j,m))} \;=\; \lim_{n\uparrow\infty}\frac{\log\nu(i,n)}{\log\nu(j,n)}$$

$$=\; \lim_{n\uparrow\infty}\frac{\log\frac{\nu(i,n)}{n}+\log n}{\log\frac{\nu(j,n)}{n}+\log n} \;=\; 1.$$

For special algorithms (SGD, $\|\cdot\|_\infty$-contractions),

$$\liminf_{n\uparrow\infty}\frac{\nu(i,n)}{n} > 0 \text{ a.s. } (\Longrightarrow \lambda_i(t) > 0 \quad \text{a.s.})$$

ensures correct convergence.

Stochastic recursive inclusions: Consider the iteration

$$x_{n+1} = x_n + a(n)[y_n + M_{n+1}]$$

where $y_n \in F(x_n)$ for some Marchaud map $x \in \mathcal{R}^d \mapsto F(x) \subset \mathcal{R}^d$, i.e., satisfying:

1. $\forall x, F(x)$ is closed, bounded, convex,

2. F has a closed graph: the set $\{(x,y) : y \in F(x)\}$ is closed, i.e.,
$(x_n, y_n) \to (x, y),\ y_n \in F(x_n)\ \forall n \Longrightarrow y \in F(x)$,

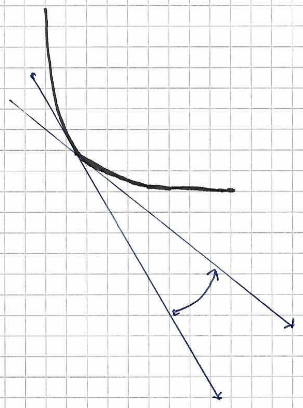
3. F has at most linear growth: $\exists K > 0$ such that $y \in F(x) \Longrightarrow \|y\| \leq K(1 + \|x\|)$.

The iteration tracks the differential inclusion

$$\dot{x}(t) \in F(x(t)).$$

e.g., stochastic subgradient descent, where $F = \partial f$, the subgradient of a convex function f at x defined as the cone

$$\partial f(x) := \{y \in \mathcal{R}^d : f(z) \geq f(x) + \langle y, z - x \rangle \ \forall z\}.$$



APPLICATIONS

Stochastic gradient descent: Here $h(x) = -\nabla f(x)$, tracks

$$\dot{x}(t) = -\nabla f(x(t)).$$

With ‘rich noise’, a.s. convergence to a local minimum if equilibria are isolated (in general, point convergence not obvious (Absil-Kurdyka), but okay for real analytic f).

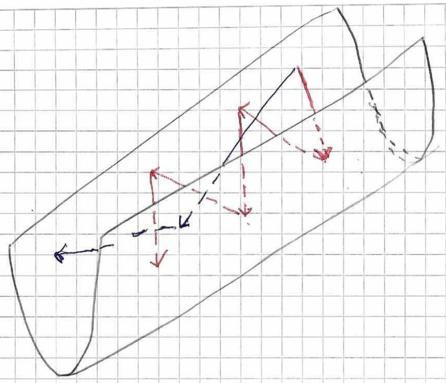
But *any* local minimum is a possible equilibrium with positive probability (follows from reachability with positive probability $+$ ‘trapping’).

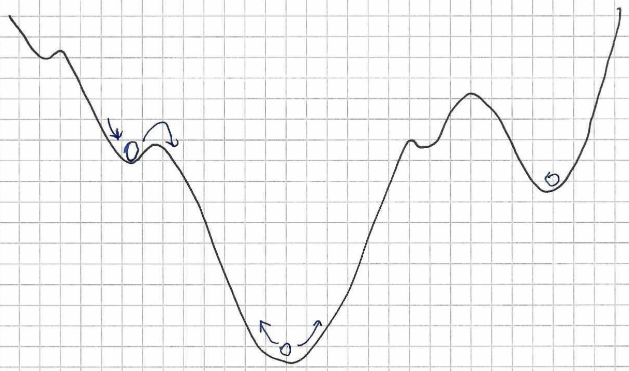
The iteration can slow down near saddle points etc. $\Longrightarrow$ can use momentum to accelerate:

$$x_{n+1} = x_n + a(n)[-\nabla f(x_n) + M_{n+1}] + b(n)[x_n - x_{n-1}].$$

($\approx$ Newton's law with friction in a potential well: speed up near unstable critical points and on flat landscapes, escapes shallow valleys, avoids zigzagging in 'pinched' landscapes, etc., but can oscillate close to equilibrium)

More sophisticated variants: Nesterov's method.





Approximate gradients:

1. Kiefer-Wolfowitz: Finite difference approximations

$$\frac{\partial f}{\partial x_i}(x) \approx \frac{f(x+\delta e_i) - f(x)}{\delta},$$

$$\frac{\partial f}{\partial x_i}(x) \approx \frac{f(x+\delta e_i) - f(x-\delta e_i)}{2\delta}.$$

Resp. $d+1$ and $2d$ function evaluations,

better discretization error in the latter.

2. Simultaneous perturbation (Spall): Take $\Delta_n(i), n \geq 0, 1 \leq i \leq d$, i.i.d. ± 1 with equal probability. Set $\Delta_n := [\Delta_n(1), \cdots, \Delta_n(d)]$.

$$\frac{\partial f}{\partial x_i}(x) \approx \frac{f(x + \delta\Delta_n) - f(x)}{\delta\Delta_n(i)}$$

$$\approx \frac{\partial f}{\partial x_i}(x) + \sum_{j \neq i} \frac{\partial f}{\partial x_j} \frac{\Delta_n(j)}{\Delta_n(i)}.$$

The second term can be absorbed into M_{n+1}.

Requires only two evaluations of f.

Single function evaluation:

$$\frac{\partial f}{\partial x_i}(x) \approx \frac{f(x+\delta\Delta_n)}{\Delta_n(i)}$$

$$= \frac{f(x)}{\delta\Delta_n(i)} + \frac{\partial f}{\partial x_i}(x) + \sum_{j\neq i}\frac{\partial f}{\partial x_j}\frac{\Delta_n(j)}{\Delta_n(i)}.$$

Both the first and the third terms can be absorbed in M_{n+1}.

Numerical issues for small δ ('small divisor' problem).

Alternative: (Flaxman et al) Pick $\xi_n \in \mathcal{R}^d$ i.i.d. zero mean, covariance matrix = identity.

$$f_i(x+\delta\xi_n)\xi_n \approx f_i(x)\xi_n(i) + \delta\frac{\partial f}{\partial x_i}(x)\xi_n(i)^2$$

$$+ \ \delta\sum_{j\neq i}\frac{\partial f}{\partial x_j}(x)\xi_n(j)\xi_n(i)$$

$\approx \delta\frac{\partial f}{\partial x_i}(x)$ + martingale difference terms.

Additional averaging helps.

Other variants: Katkovnik-Kulchitskii algorithm:

Consider a gaussian density $g_\sigma(\cdot)$ with zero mean and variance σ^2. Then for small σ, $f(x) \approx \int g_\sigma(x-y)f(y)dy$, hence

$$\nabla f(x) \approx \int \nabla g_\sigma(x-y)f(y)dy,$$

which becomes another gaussian expectation and can be estimated by Monte Carlo simulation. Has the 'small divisor' problem for small σ, which can be ameliorated by additional averaging.

Blum/Fabian scheme: Use componentwise sign of $\nabla f(x)$. Needs only one bit of information per component, but can slow down the progress away from the minima and oscillate near the minima due to the discontinuity.

Resurrected as ‘signSGD’.

More generally, can consider multiple quantization steps (Gerencser)

Simulation-based optimization (IPA, SPA etc.)

Want to minimize $E_\theta[f(X)]$ (X real valued) over θ where θ parametrizes the probability distribution of X.

Generate IID $X_n = \Phi(U_n, \theta)$ with the law corresponding to θ with $\{U_n\}$ IID uniform on $[0, 1]$. Suppose Φ is continuously differentiable.Then do:

$$\theta_{n+1} = \theta_n - a(n)\frac{\partial}{\partial\theta} f(\Phi(U_{n+1}, \theta))|_{\theta=\theta_n}.$$

Much more sophisticated versions are available.

Likelihood ratio method (Glynn et al):

Suppose the distributions P_θ corresponding to θ have densities w.r.t. a base distribution P_{θ_0}. Denote by Λ_θ the likelihood ratio of P_θ w.r.t. P_{θ_0}. Then $E_\theta[f(X)] = E_{\theta_0}[f(X)\Lambda_\theta(X)]$. The algorithm then is

$$\theta_{n+1} = \theta_n - a(n) f(X_{n+1}) \nabla_\theta \Lambda_\theta(X_{n+1})|_{\theta=\theta_n},$$

where $\{X_n\}$ have law P_{θ_0}.

For global minimization, ‘simulated annealing’:

$$x_{n+1} + a(n)[-\nabla f(x_n) + M_{n+1}] + b(n)W_{n+1}$$

where $b(n) > 0$ is chosen appropriately given $\{a(n)\}$, and $\{W_n\}$ are i.i.d. $N(0,1)$.

Tracks the stationary distribution of the stochastic differential equation

$$dX(t) = -\nabla f(X(t))dt + \frac{C}{\log t}dB(t)$$

which asymptotically concentrates on the global minima of f. For $a(n) = \frac{1}{n}$, $b(n) = \frac{C}{\sqrt{n \log\log n}}$.

SGD for machine learning:

Data: Samples of input-output pairs $(X_n, Y_n), n \geq 1$.

Objective: Minimize 'empirical risk'

$$\frac{1}{N}\sum_{m=1}^{N} L(X_m, Y_m, \theta) \approx E_\theta[L(X_n, Y_n, \theta)].$$

Example: $L(X_n, Y_n, \theta) := E[\|Y_n - f_\theta(X_n)\|^2]$.

(Need 'uniform strong law of large numbers'. Sufficient conditions given by the Vapnik-Chervonenkis theory and its extensions.)

Target: $E[L(X_n, Y_n, \theta_n)] \leq \text{min}_\theta E[L(X_n, y_n, \theta)] + \epsilon$ for $0 < \epsilon << 1$.

Many special purpose variants: 'mini-batches', adaptive stepsizes (ADAGRAD, ADAM) etc.

Typical issues: extremely large size of the ambient space as well as the state space, number of samples.

Fixes: block coordinate descent etc.

Slowdown at flat patches and saddle points etc.

Fixes: randomization, momentum

Decay / blowing up of gradients

‘Gradient-like’ or ‘Liapunov’ systems:

Suffices to have a Liapunov function $V : \mathcal{R}^d \mapsto \mathcal{R}^+$ such that

$$\langle \nabla V(x), h(x) \rangle < 0$$

for $x \notin$ some desired point or set H. Then

$$\frac{d}{dt} V(x(t)) < 0 \text{ for } x(t) \notin H,$$

so $x(t) \to H$.

For $h(x) = -\nabla f(x)$, $V = f$ works.

Gradient ascent-descent: For $f(\cdot, y)$ convex and $f(x, \cdot)$ concave (at least one of them should be strict),

$$\dot{x}(t) = -\nabla f(x(t), y(t)),$$

$$\dot{y}(t) = \nabla f(x(t), y(t)).$$

Take $V(x, y) = \|x - x^*\|^2 + \|y - y^*\|^2$ where (x^*, y^*) is a saddle point. Useful for primal-dual methods for the problem: Min. $f(x)$ subject to $g(x) \leq C$. Then do

$$x_{n+1} = x_n - a(n)[(\nabla f(x_n) + \lambda_n \nabla g(x_n)) + M_{n+1}],$$

$$\lambda_{n+1} = \lambda_n + a(n)(g(x_n) - C).$$

More generally, one of the iterations need not be gradient-based.

In this case, one can use two time-scales, invoking envelope (Danskin's) theorem:

$$\left(\nabla \min_{\theta} f(x,\theta) = \nabla_x f(x,\theta)|_{\theta=\text{argmin}(f(x,\cdot))}\right).$$

'Non-smooth' case can be handled using sub-gradients.

Fixed point schemes: $F : \mathcal{R}^d \mapsto \mathcal{R}^d$ Lipschitz, want $x^* : F(x^*) = x^*$. The o.d.e. is

$$\dot{x}(t) = F(x(t)) - x(t).$$

Works for:

pseudo-contractions: $\|F(x) - F(x^*)\|_p \leq \alpha\|x - x^*\|_p$, $0 < \alpha < 1, 1 < p \leq \infty$.

anti-monotone f: $\langle F(x) - F(y), y - x\rangle < 0$ for $x \neq y$.

Special cases of ‘non-expansive’ maps: $\|F(x) - F(y)\| \leq \|x - y\|$.

Iterates in probability simplex: 'Replicator dynamics'

$$\dot{p}_i(t) = p_i(t)(f_i(p(t)) - \sum_j p_j(t) f_j(p(t))). \quad (*)$$

The probability simplex

$$S := \{x = [x_1, \cdots, x_d] \in \mathcal{R}^d : x_i \geq 0 \ \forall i, \sum_i x_i = 1\}$$

is invariant under $(*)$ (because $\frac{d}{dt}(\sum_i p_i(t)) = \sum_i \dot{p}_i(t) = 0$), as are its faces corresponding to one or more $x_i = 0$ (because $p_i(0) = 0 \Longrightarrow p_i(t) = 0 \ \forall t$).

Converges for, e.g., $f_i = \frac{\partial F}{\partial x_i}$ ('potential game' : $V = -F$) and anti-monotone payoffs ($V =$ relative entropy).

The actual iterates need projection. Better to iterate $x_n(i), 1 \leq i < d$, so that you need only occasional (instead of every time) projection to the simplex

$$S_0 := \left\{ x \in \mathcal{R}^{d-1} : x_i \geq 0, 1 \leq i < d, \ \sum_{i=1}^{d-1} x_i \leq 1 \right\}.$$

Spurious boundary equilibria are often unstable and are avoided a.s. if the noise is rich enough.

Possible application to ‘mean-field dynamics’.

Non-Liapunov systems: Two important dynamics with 'generic convergence':

1. Newton flow (Smale): $\dot{x}(t) = -(Dh(x(t)))^{-1}h(x(t))$. (for $h(x) = \nabla f(x)$, $\dot{x}(t) = -(\nabla^2 f(x(t)))^{-1}\nabla f(x(t))$).

2. Cooperative dynamics (Hirsch): $Df(x)$ 'irreducible' and

$$\frac{\partial f_i}{\partial x_j} \geq 0 \ \forall \ i \neq j.$$

$$\frac{\nabla f(x)}{\|\nabla f(x)\|} = \text{constant}$$