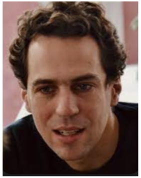
On Structural Stability and First-order Optimization with Time-dependent Adaptive Step Policy

Xiao Wang

Shanghai University of Finance and Economics

07-18-2023

Kimon Antonakopoulos
EPFL

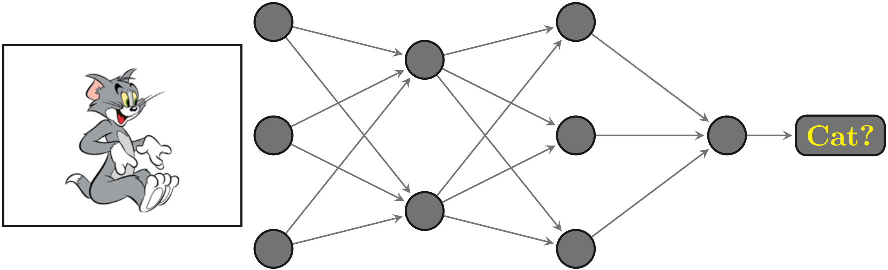


Panayotis Mertikopoulos
CNRS

Ioannis Panageas
UC Irvine

Georgios Piliouras
SUTD, DeepMind

Background



$$\min_{\mathbf{W}\in\mathbb{R}^*} \frac{1}{n}\sum_{i=1}^{n} \|f(\mathbf{W}, \mathbf{x}_i) - y_i\|^2$$

Algorithm

Gradient descent and its variants

The most popular algorithm in optimization might be the gradient descent:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha \nabla f(\mathbf{x}_t)$$

Some typical variants:

- SGD

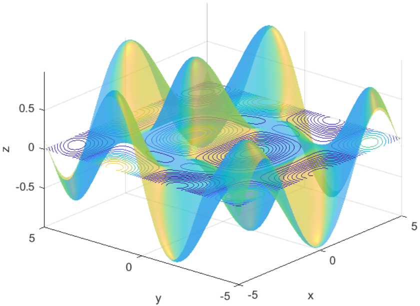
$$\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha g_t \quad \mathbb{E} g_t = \nabla f(\mathbf{x}_t)$$

- Adaptive gradient descent:

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \frac{1}{\sqrt{\delta_0^2 + \sum_{s=0}^{t} \|\nabla f(\mathbf{x}_s)\|^2}} \nabla f(\mathbf{x}_t)$$

Stages of understanding an algorithm

- Convergence: $\mathbf{x}_t$ converges to some fixed point $\mathbf{x}^*$ of the algorithm, i.e., critical point for gradient descent.
- Convergence to local optima: $\mathbf{x}^*$ is a local minimum (**our focus**).
- Convergence to global optima: $\mathbf{x}^*$ is a global minimum.



Our setup

$$\min_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}).$$

In the above, $f:\mathbb{R}^d \to \mathbb{R}$ is assumed to be lower bounded and continuously differentiable:

- $\inf_{\mathbf{x}\in\mathbb{R}^d} f(\mathbf{x}) > -\infty$
- There exists $L > 0$ such that

$$\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \le L\|\mathbf{x}-\mathbf{y}\| \quad \text{for all } \mathbf{x},\mathbf{y}\in\mathbb{R}^d.$$

- The undesirable critical points is the set of *strict saddle points*,

$$\|\nabla f(\mathbf{x}^*)\| = 0, \quad \text{and} \quad \lambda_{\min}(\nabla^2 f(\mathbf{x}^*)) < 0.$$

Underlying dynamics

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Gamma_t \nabla f(\mathbf{x}_t)$$

where Γ_t is a time-dependent step policy matrix.

Example

- $\Gamma_t = \alpha_t \cdot I$, for $\alpha_t \to 0$, e.g., $\alpha_t = \frac{1}{\sqrt{t}}$.
- AdaGrad $\Gamma_t = G_t^{-\frac{1}{2}}$,

$$G_t = \delta_0^2 I + \sum_{s=0}^{t} \nabla f(\mathbf{x}_s) \nabla f(\mathbf{x}_s)^\top.$$

Main Results

Convergence to local minimizers

The deterministic gradient descent (and its variants) with adaptive step policy only converges local minimizers. Or equivalently, non-convergence to spurious critical points, i.e., saddle points.

Example

Apart from gradient descent, many first-order algorithms can be proven convergent to local minimizers, e.g., mirror descent, proximal point method, AdaGrad on manifold and so on.

Mirror descent

Let D be a convex open subset of $\mathbb{R}^d$, and $M = D \cap A$ for some affine space A. Given a function $f : M \to \mathbb{R}$ and a mirror map Φ, the algorithm is

$$\mathbf{x}_{t+1} = h(\nabla\Phi(\mathbf{x}_t) - \alpha_t \nabla f(\mathbf{x}_t))$$

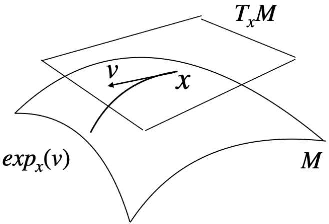
where

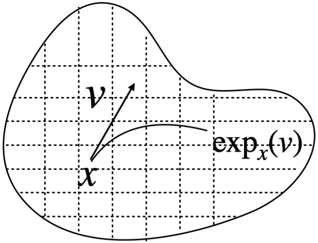
$$h(\mathbf{x}) = \mathrm{argmax}_{z \in M}\{\langle z, \mathbf{x}\rangle - \Phi(z)\}.$$

A special case of mirror descent is the Multiplicative Weights Update (MWU) that is used in game theory and multi-agent systems.

Exponential map

The exponential map $\mathrm{Exp}_{\mathbf{x}}(\mathbf{v})$ maps $\mathbf{v} \in T_{\mathbf{x}}M$ to $\mathbf{y} \in M$ such that there exists a geodesic $\gamma(t)$ with $\gamma(0) = \mathbf{x}$, $\gamma(1) = \mathbf{y}$ and $\gamma'(0) = \mathbf{v}$.





Riemannian Gradient Descent with step α_t

$$\mathbf{x}_{t+1} = \mathrm{Exp}_{\mathbf{x}_t}(-\alpha_t \mathrm{grad} f(\mathbf{x}_t))$$

Overview of the area

Robin Pemantle, 1990

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \alpha_t \mathbf{F}(\mathbf{x}_t) + \xi_t, \quad \alpha_t \to 0, \ |\xi_t| \to 0$$

The Annals of Probability
1990, Vol. 18, No. 2, 698–712

NONCONVERGENCE TO UNSTABLE POINTS IN URN MODELS AND STOCHASTIC APPROXIMATIONS[1]

BY ROBIN PEMANTLE

Cornell University

A particle in $\mathbf{R}^d$ moves in discrete time. The size of the nth step is of order $1/n$ and when the particle is at a position $\mathbf{v}$ the expectation of the next step is in the direction $\mathbf{F}(\mathbf{v})$ for some fixed vector function $\mathbf{F}$ of class C^2. It is well known that the only possible points $\mathbf{p}$ where $\mathbf{v}(n)$ may converge are those satisfying $\mathbf{F}(\mathbf{p}) = \mathbf{0}$. This paper proves that convergence

J.D. Lee, M. Simchowitz, M.I. Jordan, B. Recht, *COLT 2016*

$$\mathbf{x}_{k+1} = \mathbf{x}_k - \alpha_{\cancel{k}} \nabla f(\mathbf{x}_k) + \cancel{\xi_k}$$

The probability that gradient descent converges to saddle point is zero.

J.D. Lee, I. Panageas, G. Piliouras, M. Simchowitz, M.I. Jordan, B. Recht, *Math. programming, 2019*

Deterministic (without noise) gradient descent, mirror descent, coordinate descent, proximal point method, and manifold gradient descent with *constant step-size* avoid saddle points.

M.I. Jordan, *International Congress of Mathematicians, 2018*

"Dynamical, symplectic and stochastic perspectives on gradient-based optimization", surveyed on the topic of "escaping saddle points".

- Avoid saddle points

Lee et al. 16'
Gradient descent

Lee et al, '19
Many 1st order methods

AMPW '22
AdaGrad Family

PPW, '19
Classic MWU

PPW, '19
Vanishing step &
many 1st order methods

FPW, '22
Accelerated Riemannian GD
Accelerated MWU

- Escape saddle points

Ge et al. '15
PGD escapes saddle effectively

Jin et al
Accelerated PGD

Criscitiello & Boumal '20
Accelerated
Riemannian PGD

Jin et al '19
PGD

Criscitiello & Boumal '19
Riemannian PGD

Sun et al '19
Riemannian PGD

Li et al. '23
0'th order methods

Why saddle avoidance important?

Understanding why deep learning works (data, training, generalization), and this area of research partially answers "why training works":

- Provable convergence to local minimizer is crucial in understanding an algorithm;
- *Matrix completion has no spurious local minimum* (Ge et al. 2016), *Gradient descent finds global minima of deep neural networks* (Du et al. 2019);
- Heuristically implies that stochastic variants converges to local minimizers.

Basics of Dynamical System

Discrete-time dynamical system

A smooth dynamical system on a manifold M is a continuous differential function $g : \mathbb{Z} \times M \to M$, where $g(t, \mathbf{x}) = g_t(\mathbf{x})$ satisfies

- g_0 is the identity function.
- $g_t \circ g_s = g_{t+s}$ for all $t, s \in \mathbb{Z}$.

Fixed points and Stable set

Given a dynamical system $\mathbf{x}_{k+1} = g(k, \mathbf{x}_k)$, the set of fixed points is denoted by $\mathcal{X}^*$. The **global stable set** $W^s(\mathcal{X}^*)$ of $\mathcal{X}^*$ is the set of initial conditions where the sequence $\mathbf{x}_k$ converges to $\mathcal{X}^*$. Formally

$$W^s(\mathcal{X}^*) = \{\mathbf{x}_0 : \lim_{k\to\infty} \mathbf{x}_k \in \mathcal{X}^*\}.$$

Stable and unstable fixed points

Focusing on smooth dynamical system $g : M \to M$, we say

- $\mathbf{x}^* \in M$ is stable if eigenvalues of the differential $Dg(\mathbf{x}^*)$ have magnitude less than 1.
- $\mathbf{x}^* \in M$ is unstable if at least one eigenvalue of $Dg(\mathbf{x}^*)$ has magnitude greater than 1.

Gradient descent

Let $M = \mathbb{R}^n$ and $g(k, \mathbf{x}) = \mathbf{x} - \alpha_k \nabla f(\mathbf{x})$. Then strict saddle point $\mathbf{x}^*$, i.e., $\|\nabla f(\mathbf{x}^*)\| = 0$ and $\lambda_{\min}(\nabla^2 f(\mathbf{x}^*)) < 0$, is an unstable fixed point of g.

Technical Overview for Constant Step

Structural Stability of Dynamical Systems (Lyapunov, Hadamard, Smale...)

For a diffeomorphism $g : M \to M$, the iteration $g^n(\mathbf{x}_0)$ converges to an *unstable* fixed point $\mathbf{x}^*$ only if the initial point $\mathbf{x}_0$ is taken from the *stable manifold* of $\mathbf{x}^*$, $\mathbf{x}_0 \in W^s(\mathbf{x}^*)$.

- *Unstable*: the Jacobian $Dg(x^*)$ has an eigenvalue whose norm is greater or equal to one.
- *Stable manifold*:
  The graph of some function from stable to unstable space.
- *Proof of saddle avoidance*: Reduction to the cases where the theorem can be used.

Adaptive Steps

Take the AdaGrad as the main example

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Gamma_t \nabla f(\mathbf{x}_t)$$

- Local argument: in a neighborhood of a saddle point, the points that can be moved to saddle points by AdaGrad lie on a lower dimensional space (zero measure set);
- Global argument: carry the local zero measure set by inverse of AdaGrad, union of countable zero measure set is till of measure zero.

Review on Lyapunov-Perron for ODEs

- For gradient descent with adaptive step size, the previous stable manifold theorem does not valid. It is necessary to derive a new version of stable manifold theorem for the underlying dynamical system of those first-order methods.
- The spirit comes from the classic stability theory of ODEs. Consider the ordinary differential equation

$$\frac{d\mathbf{x}}{dt} = A\mathbf{x} + \eta(\mathbf{x})$$

and the integral operator T defined as follows

$$T\mathbf{x}(t, \mathbf{a}) = U(t)\mathbf{a} + \int_0^t U(t-s)\eta(\mathbf{x}(s, \mathbf{a}))ds - \int_t^{\infty} V(t-s)\eta(\mathbf{x}(s, \mathbf{a}))ds$$

Stable Manifolds for ODEs

Example (L. Perko)

Consider the nonlinear ODE

$$\begin{aligned}\frac{dx_1}{dt} &= -x_1\\ \frac{dx_2}{dt} &= -x_2 + x_1^2\\ \frac{dx_3}{dt} &= x_3 + x_1^2\end{aligned}$$

The only equilibrium is the origin 0.

The stable-unstable decomposition can be obtained from the Jacobian matrix of the mapping on the right hand side

$$Df(0) = \begin{bmatrix} -1 & 0 & 0 \\ 0 & -1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

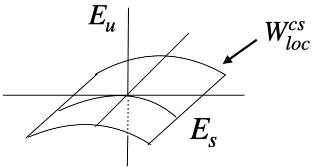
The solution is given by

$$\begin{aligned} x_1(t) &= c_1 e^{-t} \\ x_2(t) &= c_2 e^{-t} + c_1^2 (e^{-t} - e^{-2t}) \\ x_3(t) &= c_3 e^{t} + \frac{c_1^2}{3}(e^{t} - e^{-2t}) \end{aligned}$$

where $c = x(0)$ is the initial condition.

Denote $\phi_t(c)$ the flow defined by the solution, letting $\lim_{t\to\infty}\phi(c) = 0$ and $\lim_{t\to-\infty}$ respectively, we can solve that $c = (c_1, c_2, c_3)$ has to satisfty

$$c_3 = -\frac{c_1^2}{3}$$
$$c_1 = c_2 = 0$$



- The solution of $\frac{d\mathbf{x}}{dt} = A\mathbf{x} + \eta(\mathbf{x})$ can be written in terms of integral:

$$\mathbf{x}(t, \mathbf{a}) = e^{tA}\mathbf{a} + \int_0^t e^{(t-s)A}\eta(\mathbf{x}(s, \mathbf{a}))ds$$

- Denote by P^+ and P^- the projectors onto the stable, unstable subspaces E^s, E^u of e^A. Moreover, abbreviate

$$\mathbf{a}^+ = P^+\mathbf{a}, \quad \mathbf{a}^- = P^-\mathbf{a}$$

and

$$\eta^+ = P^+\eta, \quad \eta^- = P^-\eta.$$

- We can express $\mathbf{a}^-$ as a function of $\mathbf{a}^+$ by assuming $\mathbf{x}(t, \mathbf{a})$ is a solution. This can be done by multiplying e^{-tA} on integral solution of $\mathbf{x}(t, \mathbf{a})$,

$$e^{-tA}\mathbf{x}(t, \mathbf{a}) = \mathbf{a} + e^{-tA} \int_0^t e^{(t-s)A} \eta(\mathbf{x}(s, \mathbf{a}))ds$$

  projecting out the unstable part and rearranging, we have

$$\mathbf{a}^- = e^{-tA}\mathbf{x}^-(t, \mathbf{a}) - \int_0^t e^{-sA} \eta^-(\mathbf{x}(s, \mathbf{a}))ds$$

- Letting $t \to \infty$, we have

$$\mathbf{a}^- = -\int_0^\infty e^{-sA} \eta^-(\mathbf{x}(s, \mathbf{a}))ds$$

- In the end, plugging the expression of $\mathbf{a}^-$ back to the following integral solution,

$$\mathbf{x}(t,\mathbf{a}) = e^{tA}\mathbf{a} + \int_0^t e^{(t-s)A}\eta(\mathbf{x}(s,\mathbf{a}))ds = e^{tA}(\mathbf{a}^+,\mathbf{a}^-) + \int_0^t e^{(t-s)A}\eta(\mathbf{x}(s,\mathbf{a}))ds$$

we have that

$$\mathbf{x}(t,\mathbf{a}) = e^{tA}\mathbf{a}^+ + \int_0^t e^{(t-s)A}\eta^+(\mathbf{x}(s,\mathbf{a}))ds - \int_t^\infty e^{(t-s)A}\eta^-(\mathbf{x}(s,\mathbf{a}))ds$$

- The meaning of above expression is following: the right hand side can be considered an operator T acting on mappings $\mathbf{x}(t,\mathbf{a})$, transforming $\mathbf{x}(t,\mathbf{a})$ to a new mapping. And $\mathbf{x}(t,\mathbf{a})$ converges to the saddle point if and only if $\mathbf{x}(t,\mathbf{a})$ is a fixed point of the operator T.

- The previous integral equation can be solved by the method of successive approximation. Let $\mathbf{x}^{(0)}(t, \mathbf{a})$ be the initial mapping, and the $(j+1)$-th iterate provided $\mathbf{x}^{(j+1)}(t, \mathbf{a})$ is based on the integral operator:

$$\mathbf{x}^{(j+1)}(t, \mathbf{a}) = e^{tA}\mathbf{a}^{+} + \int_0^t e^{(t-s)A}\eta^{+}(\mathbf{x}^{(j)}(s, \mathbf{a}))ds - \int_t^{\infty} e^{(t-s)A}\eta^{-}(\mathbf{x}^{(j)}(s, \mathbf{a}))ds$$

- With proper topology on the space of mappings $\mathbf{x}(t, \mathbf{a})$, this successive approximation ensures that

$$\lim_{j\to\infty} \mathbf{x}^{j}(t, \mathbf{a}) = \mathbf{x}(t, \mathbf{a})$$

uniformly for all $t \geq 0$ and $\|\mathbf{a}\|$ small enough.

- An interesting observation: in the successive approximation process, the unstable components of the initial condition **a** is not involved in computation,

$$\mathbf{x}(t,\mathbf{a}) = e^{tA}\mathbf{a}^{+} + \int_0^t e^{(t-s)A}\eta^{+}(\mathbf{x}(s,\mathbf{a}))ds - \int_t^{\infty} e^{(t-s)A}\eta^{-}(\mathbf{x}(s,\mathbf{a}))ds$$

  So it is convenience to let them to be all 0's.
- Let $t = 0$, we have

$$\mathbf{x}(0,\mathbf{a}) = \mathbf{a}^{+} + 0 - \int_0^{\infty} e^{-sA}\eta^{-}(\mathbf{x}(s,a_1,...,a_k,0,...,0))ds$$

  which clearly implies that if **a** is the initial condition of the solution converging to 0, then the unstable components of **a** is a function of the stable components.

Local theory of stability for AdaGrad

Proposition

Let Γ_t be one of adaptive step-size policies of AdaGrads. Then the limit of $\{\Gamma_t\}_{t\in\mathbb{N}}$ exists and in particular, the limit is positive definite,

$$\lim_{t\to\infty} \Gamma_t = \Gamma \quad \text{with} \quad \Gamma > 0.$$

Local structure of AdaGrad around critical points

$$\mathbf{x}_{t+1} = (I - \Gamma\nabla^2 f(0))\mathbf{x}_t - \Gamma\theta(x_t) - (\Gamma_t - \Gamma)\nabla f(\mathbf{x}_t)$$

where 0 is a critical point.

- Trivial:

$$\Gamma_t = \Gamma + \Gamma_t - \Gamma$$

-

$$\mathbf{x}_{t+1} = \mathbf{x}_t - \Gamma_t\nabla f(\mathbf{x}_t) = x_t - \Gamma\nabla f(\mathbf{x}_t) - (\Gamma_t - \Gamma)\nabla f(\mathbf{x}_t)$$

- Taylor expansion of $\nabla f(\mathbf{x})$ at critical point 0:

$$\nabla f(\mathbf{x}) = \nabla^2 f(0)\mathbf{x} + \theta(\mathbf{x}).$$

Denote

$$\eta(t,x) = -\Gamma\theta(x) - (\Gamma_t - \Gamma)\nabla f(x),$$

and then the local form of AdaGrad is the following:

$$x_{t+1} = \left(I - \Gamma\nabla^2 f(0)\right) x_t + \eta(t, x_t).$$

Proposition (not used in the talk)

$\eta(t,\cdot)$ satisfies the Lipschitz type condition: for any $\epsilon > 0$, there exists a neighborhood $\mathbb{B}$ of 0 and some large t_0, so that for any $x, y \in \mathbb{B}$ and $t > t_0$, it holds that

$$\left\|\eta(t,x) - \eta(t,y)\right\| \leq \epsilon \left\|x - y\right\|.$$

Since the Hessian $\nabla^2 f(0)$ is diagonalizable, under a change of coordinates, we only have to consider the diagonalized dynamics:

$$x_{t+1} = (I - H)x_t + \eta(t, x_t).$$

Furthermore, if 0 is a saddle point, H has positive and negative eigenvalues on the diagonal, denote $H = H^+ \oplus H^-$.

- For a specific saddle point, the dimensions of positive and negative eigen-spaces are fixed.
- According to the eigen-space decomposition w.r.t. H, $\eta(t, \cdot)$ can be decomposed to $\eta(t, \cdot)^+$ and $\eta(t, \cdot)^-$, i.e.
$$\eta(t, \cdot) = \eta(t, \cdot)^+ \oplus \eta(t, \cdot)^-.$$

- Recursively use $x_{t+1} = (I - H)x_t + \eta(t, x_t)$ from x_0, we can obtain the "integral" form of x_{t+1}:

$$x_{t+1} = A(t)x_0 + \sum_{i=0}^{t} A(t-i-1)\eta(i, x_i)$$

  where $A(t)$ is t-product of $(I - H)$.
- Continuous time counterpart:

$$\mathbf{x}(t, \mathbf{a}) = e^{tA}\mathbf{a} + \int_0^t e^{(t-s)A}\eta(\mathbf{x}(s, \mathbf{a}))ds$$

- Note that the negative eigenvalue of H corresponds to the eigenvalue less than 1 in $(I - H)$, and then $A(t) = B(t) \oplus C(t)$, so the expression of x_{t+1} can be further decomposed to

$$x_{t+1}^{+} = B(t)x_0^{+} + \sum_{i=0}^{t} B(t-i-1)\eta^{+}(i, x_i)$$

$$x_{t+1}^{-} = C(t)x_0^{-} + \sum_{i=0}^{t} C(t-i-1)\eta^{-}(i, x_i)$$

- Multiplying $C(t)^{-1}$ to both sides of x^-_{t+1}:

$$C(t)^{-1}x^-_{t+1} = C(t)^{-1}C(t)x^-_0 + C(t)^{-1}\sum_{i=0}^{t} C(t-i-1)\eta^-(i, x_i),$$

or equivalently:

$$x^-_0 = C(t)^{-1}x^-_{t+1} - C(t)^{-1}\sum_{i=0}^{t} C(t-i-1)\eta^-(i, x_i)$$

- Now suppose $\{x_t\}_{t\in\mathbb{N}}$ is a sequence generated by AdaGrad with initial condition x_0 and $x_t \to 0$, then of course $x^-_{t+1} \to 0$. Since $C(t)$ is diagonal and has eigenvalues > 1, the inverse $C(t)^{-1} \to 0$ as $t \to 0$. So the term $C(t)^{-1}x^-_{t+1}$ vanishes.

- Let $t \to 0$, we have the identity:

$$x_0^- = \lim_{t\to\infty} C(t)^{-1} \sum_{i=0}^{t} C(t-i-1)\eta^-(i, x_i)$$

whose ODE counterpart is

$$\mathbf{a}^- = -\int_0^\infty e^{-sA}\eta^-(\mathbf{x}(s,\mathbf{a}))ds$$

- Consider x_i as function of $x_0 = (x_0^+, x_0^-)$ since it is generated from x_0, i.e.,

$$x_i = x_i(x_0^+, x_0^-).$$

- The identity is nothing but an implicit function (proof needed) of x_0^+ and x_0^-:

$$x_0^- = F(x_0^+, x_0^-).$$

Theorem (informal)

If dynamical system

$$x_{t+1} = (I - H)x_t + \eta(t, x_t)$$

converges to a saddle point with initial condition x_0, then the components x_0^+ and x_0^- lie on the hypersurface defined by the equation

$$x_0^- = F(x_0^+, x_0^-)$$

If the function F is good (differentiable), then the hypersurface is a lower dimensional space, so of measure 0.

Global stability of AdaGrad

Diffeomorphism

An invertible map $\varphi : \mathbb{R}^d \to \mathbb{R}^d$ is a diffeomorphism if both φ and φ^{-1} are differentiable.

Property we actually use:

The image and pre-image of a zero measure set under diffeomorphism is of measure zero.

- Recall that we assume δ_0 is large enough so that $\delta_0 > L$.
- AdaNorm:

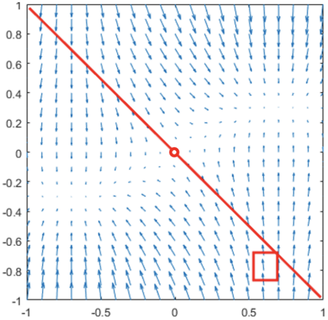
$$\frac{1}{\sqrt{\delta_0^2 + \sum_{s=0}^{t} \|\nabla f(x_s)\|^2}} \to 0 \quad \text{as} \quad \delta_0 \to \infty.$$

Limitations: saddle avoidance with inequality constraints

Compared to unconstrained settings, much less is known.

- Maher Nouiehed, Jason Lee and Meisam Razaviyayn. *Convergence to second-order stationarity for constrained non-convex optimization*, 2018.
- They provide an counter-example where the projected gradient descent might converge to strict saddle points with positive probability.

- $\min f(x,y) = -xye^{-x^2-y^2} + \frac{1}{2}y^2$, s.t. $x + y \leq 0$.



(b) Negative Gradient Flow for $f(\cdot)$

- If projected gradient descent has at least two repelling directions at saddle point, does it avoid saddle points?

From function approximation to PDE solving

- Using neural network to approximate some solution of a PDE follows the same steps as function approximation.
- Suppose the neural network has the form of

$$f(\mathbf{W}, \mathbf{a}, \mathbf{x}) = \frac{1}{\sqrt{m}} \sum_{r=1}^{m} a_r \sigma(\mathbf{w}_r \mathbf{x})$$

  where the activation can be chosen so that it has differentiability, e.g.

$$\sigma(x) = \frac{1}{\ell!} x^\ell \;\text{ if }\; x \geq 0 \;\text{ and }\; \sigma(x) = 0 \;\text{ if }\; x < 0.$$

- We compute the partial derivatives of the neural network.

$$\frac{\partial f}{\partial x_i} = \frac{\partial}{\partial x_i}\left(\frac{1}{\sqrt{m}}\sum_{k=1}^{m} a_k \sigma(\mathbf{w}_k \mathbf{x})\right) = \frac{1}{\sqrt{m}}\sum_{k=1}^{m} a_k \frac{\partial \sigma(\mathbf{w}_k \mathbf{x})}{\partial x_i}$$
$$= \frac{1}{\sqrt{m}}\sum_{k=1}^{m} a_k \sigma'(\mathbf{w}_k \mathbf{x}) w_{ki},$$

- Consider the very simple first-order PDE given as follows,

$$\frac{\partial f}{\partial x_1} + \ldots + \frac{\partial f}{\partial x_d} = h(\mathbf{x})$$

- The approximation problem regarding the linear PDE:

$$\sum \frac{\partial f}{\partial x_i} = \frac{1}{\sqrt{m}} \sum_{k=1}^{m} a_k \left(\sum_{s=1}^{d} w_{ks} \right) \sigma'(\mathbf{w}_k \mathbf{x}) \approx h(\mathbf{x})$$

- Denote

$$b_k = a_k \left(\sum_{s=1}^{d} w_{ks} \right)$$

  the above approximation problem is actually an classic neural network approximation with constraints.
- The complexity of these constraints depends on the non-linearity and order of the PDE.

Take-home msg

- We were happy with many first-order learning algorithms, and we can stay happy (even for algorithms to be discovered);
- Challenges come from inequality constraints. Almost everything is open on saddle avoidance/escaping.

Cám o'n !
Thank you !