Caution remark

① Ongoing work

$\Downarrow$

② Possible errors likely! !

Solutions for DNNs

dynamics and norms

1. Dynamical system approach : square loss

   Linear case one layer

   Multilayer nonlinear

(... Complexity control, minimum norm)

2. Dynamical system approach : exp losses
   - 1 layer linear, 1 layer nonlinear, Srebro, us, normalization, regularization, norm control min norm

Dynamical systems approach: square loss

We are interested in characterizing which solutions GD converges to. One reason is eventually to understand which kind of complexity control takes place in deep nets.

Notation

$$f(x) = W_K^{l,m} \cdots \left(\cdots \sigma\left(W_2^{qi}\left(\sigma\left(W_1^{ij} \; x_j\right)\right)\right)\right)$$

Jacobian of $f \rightsquigarrow \dfrac{\partial f}{\partial w_k^{ij}}$

Hessian of $f \rightsquigarrow \dfrac{\partial^2 f}{\partial w_k^{ij} \, \partial w_{k'}^{ab}}$

Why learning rate ?

An intuition for a non-discrete math guy is given by

$\dot{x} + \gamma(t) x = 0$

Solution is $x(t) = x_0 e^{-\int \gamma(t)\, dt}$.

For convergence to minimum (ϕ) we need $\int \gamma(t)\, dt \to \infty$

For robustness against noise such as discretization noise in

$\dot{x} + \gamma(t)\left(x + u(t)\right) = 0$

we also need $\gamma(t) \to 0$

Gradient dynamical systems and GD

The loss function is $\equiv$ potential (Liapunov) function $L(w) = \sum_i^n \ell_i(f_w(x_i), y_i)$

The gradient dynamical system induced by GD is

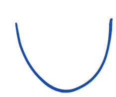
$$\dot{W} = -\gamma_t \nabla_w L(w) = -\gamma_t F(w)$$

Square loss $\Longrightarrow$ $\ell_i = (f_w(x_i) - y_i)^2$

We are interested in equilibrium points (convergence of GD) that is w^* s.t. $\underline{\dot{w} = 0}$

We are interested in stability of w^*. For this we look at Hessian of L that is

$$H = \frac{\partial^2 L}{\partial w_k^{ij} \partial w_{k'}^{ab}}$$

H of convex function is pos. def. 

Linear one-layer networks

In this case $K = 1$ and

$$f(x) = W^{1,J} x_J = W^T x$$

Thus

$$L_u(f) = \sum_n^N (y_n - W^{1J} x_J^n)^2 \qquad y \in \mathbb{R}(-1,1)$$
$$x \in \mathbb{R}^d$$

$$\dot{W}^{1,J} = 2\sum_n^N E_n \, x_J^n$$
$$\Downarrow$$
$$W^{1,J}_{min} \quad s.t \quad Y = WX, \quad W = YX^{-1}$$

Suppose overparametrization that is $N < d$
Then f fits data that is $E_n = 0 \quad \forall n = 1, .. N$
at which point $\dot{w}^{1J} = 0$

Notice that during dynamics only the weights that are in the span of the x^n <u>change</u>. The others in the null space of x^n do <u>not</u> change, the gradient is zero.

The Hessian of L is

$$H_{ij} = \sum_n \left(x_J^n \, x_i^n \right)$$

$$= \text{at min} = \sum x^n x^{nT}$$

that is the sum of autocovariance matrices, each one being semidefinite (positive). Thus there are zero eigenvalues.

Show slide

In this situation – degenerate H – there is no unique minimum and no explicit control of norm.

Regularization. Adding an arbitrarily small regularization term solves the problem here

Instead of

$$L = \sum_n^N (y_n - W^{IJ} x_J^n)^2$$

I use

$$L = \sum_n^N (y_n - W^{IJ} x_J^n)^2 + \frac{\lambda}{2} \sum_J \|W^{IJ}\|^2$$

$\|\underline{w}\|^2$

Thus

$$\dot{W}^{IJ} = \left(\sum_n^N E_n x_J^n + \lambda W^{IJ} \right)$$

which is called "weight decay".

The equilibrium point shifts to

$$\lambda W^{IJ} = \sum^N E_n x_J^n - \sum (y - W^{IJ} x_J^n) x_J^n$$

that is

$$\lambda W^{i,j} = \sum_n \left[y_n x_n^j - W^{i,l} x_l^n x_j^n \right]$$

corresponding to

$$(-Y + WX)X^T + \lambda W = 0$$

$$-YX^T + W(\lambda I + XX^T) = 0$$

$$YX^T = W(XX^T + \lambda I)$$

$$W = YX^T(XX^T + \lambda I)^{-1}$$

The Hessian is now

$$H_{ij} = \sum_n^N x_i^n x_j^{n\,T} + \lambda I$$

always positive definite for $\forall \lambda > 0$

Implicit regularization

For square loss and linear networks GD converges to the same solution of regularization with $\lambda \to 0$ which is the pseudoinverse solution

$$W = YX^T(XX^T)^{-1} = YX^+$$ which is the minimum norm solution.

The condition for this to happen is that the initial condition for GD is $W^0 \sim 0$.

Then the degenerate components of the gradient $\frac{\partial L}{\partial W^{ij}} = 0$ do not change the weights W^{ij} which remain ~ 0 thus $\Rightarrow$ min norm

slide

Norm control, regularization, implicit regularization

① The regularization min is hyperbolic and independent of initial conditions. This also means one can perturb and get back to same w.

② The implicit regularization min is degenerate and depends on initial conditions. This means that perturbations during GD will change final w.

A similar situation holds for multilayer linear networks.

slides

Deep nets, square loss

For general networks $f(x) = W_K^{i,\ell}\,\sigma(\cdots \sigma(W_1^{ij} x_J))$

the gradient equations are

$$\dot{W}_K^{i,j} = -\frac{\partial}{\partial W_k^{ij}} L = -\sum_n E_n \frac{\partial f}{\partial W_k^{ij}} \left\{ -\lambda W_k^{ij} \right\}$$

$$H_{abk'}^{ijk} = \frac{\partial^2}{\partial W_k^{ij}\,\partial W_{k'}^{ab}} L =$$

$$= \sum_n^N \frac{\partial f}{\partial W_k^{ij}} \frac{\partial f}{\partial W_{k'}^{ab}} - E_n \frac{\partial^2 f}{\partial W_k^{ij}\,\partial W_{k'}^{ab}}$$

$$\left\{ -\lambda \underbrace{I}_{\delta_{ia}\,\delta_{jb}\,\delta_{kk'}} \right\}$$

As for linear networks there is no finite equilibrium point because H is in general degenerate. This is clear in the deep polynomial case because only the W_k^{ij} in the span of the $\sum_n \frac{\partial f}{\partial W_k^{ij}}$ change. If $N << \# W_K^{ij}$ there is degeneracy.

More formally Kenji Kawaguchi proved

Theorem 4 (theory III)

Under standard assumption the Hessian matrix of a NN which is overparametrized has at least one zero eigenvalue.

This is similar to linear case. However, unlike the linear case we cannot guarantee that starting with $W^0 \sim 0$ the solution will be min. norm. First, in general there will be many degenerate minima. Second, the degenerate directions near a minimum will not be degenerate during GD.
This issue is related to the conditions under which I can use linearization tools - such as the Hessian - around minima of a dynamical system to characterize its behaviour. It turns out that a general condition for validity of linearization is that the Hessian is non-degenerate. The Hartman Grobman theorem says

Hartman Grobman Theorem

Dynamical system $\dot{W} = -F(w) = -\nabla_w L$

If F has hyperbolic equilibrium W^* and Hessian of L H does not have zero eigenvalues there, then $\exists$ N of W^* and homeomorphism $h: N \to \mathbb{R}^d$ s.t. $h(W^*) = 0$ and in N the flow of $\dot{W} = -F(w)$ is topologically conjugate by the continuous map $U = h(w)$ to the flow of linearized system $\dot{u} = -Hu$

Homeomorphism cont. map with inverse between topological spaces

Extension to deep nets via HG for square loss

Regularization can transform degenerate Hessian at a min into hyperbolic min

Consider

$$\dot{W}_k^{i,j} = -\frac{\partial}{\partial W_k^{i,j}} L = -\sum_n E_n \frac{\partial f}{\partial W_k^{i,j}} - \lambda W_k^{i,j}$$

$$W^* \text{ s.t. } \dot{W} = 0 \Longleftarrow E_n = \epsilon \;\; \forall n, \quad f(x_n; W^*) - y_n = \epsilon$$

Then

$$H = \sum_n^N \frac{\partial f}{\partial W_k^{ij}} \frac{\partial f}{\partial W_{k'}^{ab}} - E_n \frac{\partial^2 f}{\partial W_k^{ij} \partial W_{k'}^{ab}} - \lambda I$$

$\delta_{ia} \delta_{jb} \delta_{kk'}$

can be pos. def.

at a minimum

If H = non degenerate ^ then Hartman Grobman

applies and minimum is hyperbolic.

This may happen when $E_n \simeq \mathcal{E}$;
then Hessian may be pos def. for
small.

min. norm and complexity control. Otherwise not, as shown
in slide (fig 4 Theory III).

But ... it depends on H if $E_n \neq 0$

Dynamical systems approach :

exponential type loss

Exponential loss $\Longrightarrow$ $L = \sum_n^N \ell_i$

$$\ell_i = e^{-y_i f(x_i)}$$

$y_i = \pm 1$

Linear net $\Longrightarrow$ $f(x) = W^{1,J} x_J$

Assuming separable data $y_n f(x_n) > 0 \to f(x_n)$

GD generate dynamical system

$$\dot{W}^{1,J} = -\frac{\partial}{\partial W^{1,J}} \sum_n^N e^{-W^{1,J} x_J^n}$$

$$= + \sum_n^N e^{-W^{1,J} x_J^n} x_J^n$$

with vector notation $W^{1,J} \longrightarrow \underline{w}$

$$\dot{\underline{w}} = \sum_n^N \underline{x}^n e^{-w^T x^n}$$

Thus the components of w grow positive or negative to infinity slower and slower

Srebro result (2017)

Lemma 1 GD with $\ell_i = e^{-y_i f(x_i)}$ and f linear has $w(t) = \tilde{w} \log t + \rho(t)$ s.t.
$\lim_{t\to\infty} \frac{w(t)}{\|w(t)\|} = \frac{\tilde{w}}{\|\tilde{w}\|}$ where $\tilde{w}$ is solution of

hardmargin SVM, that is

$$\tilde{w} = \operatorname{argmin}_{w \in \mathbb{R}^d} \|w\|^2 \text{ s.t } \forall n \quad w^T x_n \geq 1$$

The convergence is independent of initial conditions

Remarks Notice that as for square loss the degenerate w components are not changed during GD. However normalization sets them to zero asymptotically...

We will show that

1) regularization restores hyperbolicity for linear nets

2) a normalization version of GD (similar to weight normalization) also restores pos def Hessian but only under very special data

Because of ① + ② it seems we have similar situation as for square loss: regularization (and perhaps normalization) can guarantee hyperbolicity, thus validity of Hartman Grobman and thus extension to DNNs, including control of norm.

Next we study the effect of regularization and normalization

Regularization for linear nets with exp loss

$$\dot{W} = -\frac{\partial}{\partial W}\left(L + \lambda w^2\right) = \sum_n x_n e^{-w^T x_n} - \lambda w$$

The minimum w^* is given by

$$\frac{1}{\lambda}\sum_n x_n e^{-w^{*T} x_n} = w^*$$

This is not linear but we linearize and use HG

$$X = \frac{1}{\lambda}\sum e^{-w^{*T} x_n} x_n \; -$$

$$X = \underline{w^*} \longrightarrow$$

more in general $w^* = w^\infty$ is a linear combination of support vectors thus $w^T x_n = 1$

$$H = -\sum_n^N x_n x_n^T e^{-w_*^T x_n} - \lambda I$$

which is a stable min. ($x_n x_n^T$ has zero eigenvalues).

This recovers the gist of Srebro result for $\lambda \to 0$.

RELU and homogeneity properties of DNNs

Definition of RELU implies

$$\sigma(z) = \frac{\partial \sigma(z)}{\partial z} \, z$$

and

$$f(W; x) = \Pi_{k=1}^{K} \rho_k \quad \tilde{f}(\tilde{W}_k^{ij}; x)$$

where $\rho_k \tilde{W}_k^{ij} = W_k^{ij}$

and $\rho_k = \| W_k \|$

Furthermore (Sasha et al.)

$$W_k^{ij} \frac{\partial f(x)}{\partial W_k^{ij}} = f(x) \qquad \forall k$$

Normalization

$$\tilde{w} = \frac{w}{\|w\|} \qquad \|\tilde{w}\| = 1$$

1) $\dfrac{\partial \|w\|}{\partial w} = \tilde{w}$

2) $\dfrac{\partial \tilde{w}}{\partial w} = \dfrac{I - \tilde{w}\tilde{w}^T}{\|w\|} = S$

3) $S w = S \tilde{w} = 0$

Let us reparametrize GD in terms of $\tilde{w}, \rho$ (1 layer) starting from

$$\dot{w}^T = \sum_n^N \frac{\partial f(x_n)}{\partial w^T} e^{-f(x_n)}$$

This gives a new dynamical system

$$\|\dot{w}\| = \dot{\rho} = \frac{\partial \|w\|}{} \dot{w} = \tilde{w}^T \dot{w}$$

$$\dot{\tilde{w}} = \frac{\partial \tilde{w}}{\partial w} \dot{w} = S \dot{w}$$

Thus

$$\dot{\rho} = \frac{\cancel{w^T}}{\|w\|} \sum_n^N e^{-\rho \tilde{f}} f = \sum_n^N e^{-\rho \tilde{f}} \tilde{f}(x_n)$$

always growing

$$\dot{\tilde{w}} = \frac{\sum_n e^{-\rho \tilde{f}}}{\|w\|} \left(\frac{\partial f}{\partial \tilde{w}} - \tilde{w} f(x_n) \right)$$

For linear networks

$$\dot{\rho} = \sum_n^N e^{-\rho \tilde{w}^T x_n} \tilde{w}^T x_n$$

$\rho = \|w\|$

$$\dot{\tilde{w}} = \frac{\sum_n e^{-\rho \tilde{w}^T x_n}}{\rho} \left(x_n - \tilde{w} \tilde{w}^T x_n \right)$$

Define $$X = \sum_n \frac{e^{-\rho \tilde{w}^{*T} x_n}}{\rho} x_n$$

Then stationary point in w satisfies

$$X - \tilde{W}^* \tilde{W}^{*T} X = 0$$

$$\Rightarrow \quad \tilde{w}^* = \frac{X}{\|X\|}$$

If $\sum_n e^{-\rho \tilde{w}^{*T} x_n} \simeq \delta(x - x^*)$ then

at minimum

$$H = \left(I - \tilde{W}\tilde{W}^T\right) X X^T + \frac{1}{\rho}\left(\tilde{W}^T X + \tilde{W} X^T\right)$$

which is pos. def. unlike

However the argument depends on max effect in $\sum_n e^{-\rho(x_n)}$ which is not true in general!

Summary

- Square loss

  Hyperbolic behaviour is guaranted for 1 layer;
  ? can happen for multiple layers by
  small λ regularization. In this case
  WG guarantees extension of linear analysis to
  nonlinear deep nets. Then solutions are
  min norm solutions for $\lambda \to 0$.

- Exponential loss

GD converges to hyperb solution - independent of initial conditions - in terms of normalized weights for linear networks (frebro)
We prove the same is true for regularized GD, not true for normalized GD and true for Early stopping GD.