Optimization

why does it work ?

minima ? How many ?

Do they control norm = complexity ?

Plan

Background on SGD and NN

- cross-entropy, traj

Minima

- Number, Bezout, square loss
- Degeneracy Bezout
- SGD and Langevin
- SGD finds global minima

Stability

- equil. points ?
- Hessian ?
- Variations

Loss functions

<u>L^2 loss</u> $$L(w) = \sum_i^N \left(f(x_i; w) - y_i \right)^2$$

<u>Exp. loss</u> $$L(w) = \sum_i^N e^{-y_i f(x_i; w)}$$

logistic is $$\sum_i^N \ln\left(1 + e^{-y_i f(x_i; w)}\right)$$

crossentropy is multi-label version logistic

Gradient Descent optimization

$$\mathring{W}_k^{ij} = -\gamma \frac{\partial}{\partial w_k^{ij}} L$$

$$\updownarrow$$

$$W_k^{ij\,t+1} - W_k^{ij\,t} = -\gamma \frac{\partial L}{\partial w_k^{ij}}$$

Instead of $L = \sum_i^N l_i$

use <u>minibatches</u> selected at random for each iteration $\Longrightarrow$ SGD

Minima

Can we say how many which kind, independent of GD?

Key fact: DNN are usually overparametrized with $M >> N$

M is $\#\ W^{i\,T}_{H}$

N is $\#$ eqts.

Bezout theorem

Instead of

$$\min_{w} \sum_{i}^{N} \left(f(x_i; w) - y_i \right)^2$$

consider

$$f(x_i; w_k^{ij}) = y_i$$

$$i = 1, \dots, N \qquad i, j, k \dots$$

- because it is easy to find zero error

- because overparametrization

Bezout theorem

If $\sigma(z)$ were $P(z)$ then

$f(\underline{x}; w)$ is multivariate polynomial in the $W^{(i)}_{k}$ (and in the $\underline{x}$).

Then $\boxed{f(x_i) - y_i = 0}$ $i = 1, \dots N$

is a system of N polynomial equations in M variables. $N \sim 60k$ in CIFAR, $M \sim 300k$

Bezout theorem

A set of N polynomial eqts. in M variables of degree K has $\leq K^N$ isolated solutions if $M = N+1$.

$M \gg N$ then the solutions

are degenerate.

Remarks

(i) This is similar to systems of linear equations.

(ii) For the size of NN to-day # isolated solutions is very high ~ # protons universe and furthermore degenerate because of $M >> N$.

(iii) The point on M and N and degeneracy is what we use next.

Because $M > N$ the global minima corresponding to zero error for all x_i $\left(f(x_i; W) - y_i = 0 \quad i = 1, \dots N \right)$ are degenerate.

What about all minima?

The stationary points of the gradient are

$$-\nabla_W L = 0 \qquad \text{which means}$$

$$-\frac{\partial}{\partial w_H^{ij}} \sum_i^N \left(f(x_i) - y_i \right)^2 = 0$$

These are M equations in M unknowns. If f polynomial the equations are polynomial equations, Bezout theorem applies, the solutions are in general <u>not</u> degenerate.

This $\Rightarrow$

1. global solutions are degenerate
2. local minima are not degenerate

SGD, SGDL

For the next step I need to establish similarity between SGD and Langevin equations.

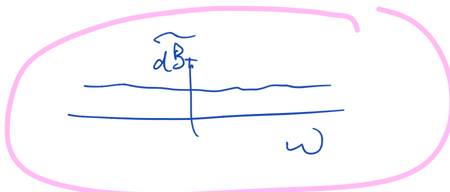
$$L = \sum_i^N (f(x_i) - y_i)^2 = \sum_i V(f, z_i)$$

- GD $\Delta w_t = W_{t+1} - W_t = -\gamma_t \nabla_w L \sim \dot{W}_n^{ij} = -\frac{\partial}{\partial w_n^{ij}} L$

dynamical gradient system

- SGD $\Delta w_t = -\gamma_t \nabla_w V(f, z)$ with z_i chosen at random

– GDL



$$\dot{W}_{n}^{ij} = -\frac{\partial L}{\partial W_{k}^{ij}} + dB_t$$

SDE

where dB_t is the derivative of the Brownian motion, that is zero mean white noise with Gaussian statistics

SGD is similar to GDL in simulations and also if I write SGD as

$$\dot{W} = -\nabla_W (L - L + V) = -\nabla_W L + \xi_t$$

where $\xi_t = \nabla_W (L - V)$ is a pseudonoise

s.t. $\mathbb{E}(\xi_t) = 0$. ξ_t is defined in terms of minibatches (where CLT applies, giving ξ_t some Gauss like properties)

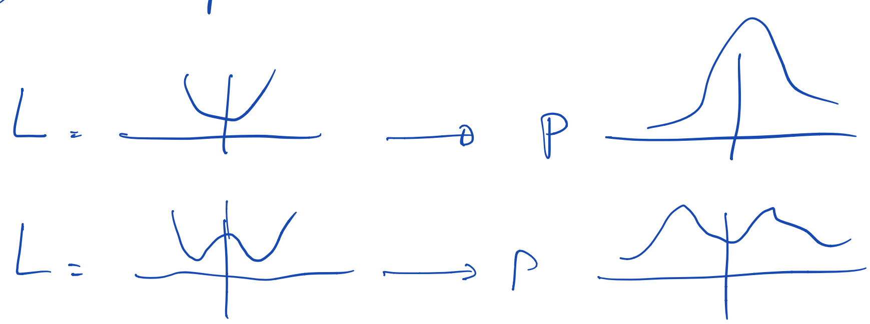
Let us speak about GDL which is a SDE.

$$\dot{w} = -\nabla_w L(w) + d B_t$$

Its "solution" for stationary prob. distrib is

$$p(w) = \frac{1}{Z} e^{-\frac{L}{T}}$$

This means that if



Important: p shows concentration of probabilities with large d: most of probability mass is in large volume minima of $\mathcal{U}$: see slides

The conclusion is that the prob solution of GDL ~~prefers~~ with high probability degenerate minima. Together with Bezout conclusions this implies that GDL prefers global minima (vs local ones). Because of GDL $\approx$ SGD this is valid also for SGD.

The last point in this class — which is also a harbinger of next class — is about the structure of the solutions of GD with square loss in the overparametrized case.

The dynamical system is

$$\dot{W}_k^{ij} = -\nabla_w L = 2 \sum_i^N \underbrace{\left(y_i - f(x_i)\right)}_{E_i} \frac{\partial f(x_i)}{\partial W_k^{ij}}$$

may be zero too!

if $E_i = 0 \quad \forall i$ then

$$\dot{W}_k^{ij} = 0$$

Are these solutions stable? Unique?

Let us look at Hessian of L

$$\frac{\partial^2 L}{\partial W_k^{ij} \partial W_{k'}^{ab}} = 2 \sum_i^N \left\{ -\frac{\partial f}{\partial W_{k'}^{ab}} \frac{\partial f}{\partial W_k^{ij}} + \left(y_i - f(x_i)\right) \frac{\partial^2 f}{\partial W_k^{ij} \partial W_{k'}^{ab}} \right\}$$

$$= \text{if } E = 0 \quad \text{to } \alpha \sum_i^N - \frac{\partial f}{\partial W_{h'}^{ab}} \frac{\partial f}{\partial W_h^{ij}}$$

if $-H$ is p d then stability. But

$\frac{\partial f}{\partial W_{h'}^{ab}} \frac{\partial f}{\partial W_h^{ij}}$ is often $\emptyset$ ... degenerate directions

valleys ... as expected from Bezout analysis

...