# Striking behavior of SGD.

consider

$$\min_{X \succeq 0} F(X) = \| A(X) - y \|_2^2$$

here $X$ is a matrix, $A$ a linear fn that accepts a matrix and makes a vector, $y$ a vector, $A_i$ symmetric

$$A_i(X) = \text{ith component of } A(X)$$
$$= \text{trace}(A^T X)$$
$$(= \text{sum}(\text{sum}(A .* X))).$$

$X$ is $n \times n$, and we are interested in the case where $\dim y = m \ll n^2$

(so your *really* can't recover $X$ without something interesting going on).

we apply some opt procedure to $F(X)$ to get $X^*$

1) (obvious) there are many $X^*$ that are optimal

2) (less so) the $X^*$ we get depends on the procedure we use

3) we are interested in

$$\|X - X^*\|^2 \leftarrow \text{generalization error}$$

It should be clear that this could be pretty much arbitrary (eg $A_i$ = matrix with one 1 in it

rather than work with $X$,
consider working with $U$

$$X = UU^T$$ (so p.s.d. is satisfied)

now if $U$ is tall + thin, we are imposing a rank constraint on $X$

- doing so has a long history
- it is often useful (recommender systems, feature seln, etc)

instead, for us $U$ is square.

$$\min_U f(U) = \| A(UU^T) - y \|_2^2$$

1) Gradient descent does odd things on this problem (Gunasekar, fig 1)

Why:

- very rough analysis.

Consider $$u^{(n+1)} = u^{(n)} - \eta \nabla_u f(u).$$

- this is a form of numerical integration of a dynamical system

$$u_{(t+\Delta t)} = u_{(t)} - \Delta t \nabla_u f(u)$$

or $$\dot{u}_t = \frac{du}{dt} = -\nabla_u f(u).$$

take the gradient! ⇓ to get.

$$\nabla_u f(u) = 4 \left[ \sum_i r_i A_i \right] u$$

where $r_i = A_i(UU^T) - y_i$

now $X_t = U_t U_t^T$

so $\dot{X}_t = \dot{U}_t U_t^T + U_t \dot{U}_t^T$

$= -\left(\sum_i r_i A_i\right) X_t - X_t \left(\sum_i r_i A_i\right)$

now we care about limit points of this flow.

→ assume $m = 1$, $A_1 = A$, $y_1 = y$.

$\dot{X} = -r_t (AX_t + X_t A)$

$X_t = \exp(S_t A)\, X_0 \exp(S_t A)$

↑ initial cond

$S_t = -\int_0^T r_t \, dt$

Now assume it's OK to pass to DS limit

IF $X_t$ has a limit point, we expect $|S_t|$ to be big. (G claim ∞?)

Now

$$\exp(sA) = I + sA + \frac{s^2A^2}{2} + \cdots$$

if $v$ is eigenvector, $\lambda$ eval

$$\exp(sA)v = \left(e^{s\lambda}\right)v.$$

→ this means that, as long as you start away from zero, one e.v. of $X$ should be a lot larger than others

# gradient descent for a linear predictor

- we have $\{\underline{x}_n, y_n\}_{1:N}$, $\underline{x}_i$ a vector

  $y_i$ a label $\in \{1, -1\}$.

- $$\min_w \sum_{i=1}^{N} \ell(y_i w^T x_i)$$

is our learning problem, for $\ell$ some loss.

Assume: - linear separability

$$w_*^T x_i \geq 0 \quad \text{all } i$$

- $\ell$ is: positive, differentiable

log-loss,

NOT hinge loss

monotonically decreasing to zero

derivative is Lipschitz

$\lim_{u \to \infty} \ell(u) \neq 0$ (so you can go downhill)

- now assume WLOG all labels are 1 (by flipping sign of $x_n$ as required)

- So we care about

$$\sum_i \ell(w^T x_i)$$

- GD gives

$$w(t+1) = w(t) - \eta \nabla L$$
$$= w(t) - \eta \sum_i \ell' \cdot x_i$$

- notice that

$$\lim_{t \to \infty} \|w(t)\| = \infty$$

because

$$w_*^T \nabla \mathcal{L} = \sum_i \ell'(w_*^T x_i)$$

but $w_*^T x_i > 0$ (lin sep)

and $\ell' < 0$

so $w_*^T \nabla \mathcal{L} < 0$ so $\nabla \mathcal{L} \neq 0$

BUT GD on smooth loss goes to critical pt.

so $\nabla \mathcal{L} \xrightarrow[t \to \infty]{} 0$

which means $w(t)^T x_i > 0$ for large enough $t$ ( ~~or else~~ we must have $\ell' \to 0$ )

So what is

$$\lim_{t \to \infty} \frac{w(t)}{\|w(t)\|} \quad ?$$

(if it exists).

work with $\ell(u) = e^{-u}$

assume $\frac{w(t)}{\|w(t)\|} \longrightarrow w_\infty$

$$w(t) = g(t)\, w_\infty + \rho(t).$$

where $g(t) \longrightarrow \infty$, $\frac{\rho(t)}{g(t)} \longrightarrow 0$

then

$$-\nabla \mathcal{L}(w) = \sum_i \exp(-w(t)^T x_i)\, x_i$$

$$= \sum_i e^{-g w_\infty^T x_i}\, e^{-\rho^T x_i}\, x_i$$

Now notice that this means, for big $t$, only samples w/ smallest value of $w_\infty^T x_i$ contribute to gradient.

so $w_\infty$ is dominated by these

$\Rightarrow w_\infty$ = linear combination of $x_i$ st $w_\infty^T x_i$ is min.

now consider $\hat{w} = \frac{w_\infty}{\min_i (w_\infty^T x_i)}$

( support vectors — more than 1

so $\hat{w} = \sum_i \alpha_i x_i$

and either $\alpha_i \geq 0$, $\hat{w}^T x_i = 1$

OR $\alpha_i = 0$, $\hat{w}^T x_i > 1$

(KKT for SVM!)

In each case, choice of optim procedure was not innocent from persp of learning.

Q: What about ADAM, Adagrad, Momentum, etc?

Q: What about SGD?

Q: What about convnets?