Curious behavior of stochastic gradient descent

Setup data x_i y_i

x_i: feats; y_i: labels $1, -1$

linear predictor w

loss ℓ.

so we minimize $\sum_i \ell(w^T [x_i\ y_i]) = L(w)$

by choice of w

Assume:

- data is separable
- loss is e^{-u} (this can be relaxed!)
- for simplicity of notation all labels are 1 (so if x_i has label -1, use $-x_i$ instead.)

Notice in this case, there is no finite w that minimizes the loss but

$$\lim_{w \to \infty} L(w) = 0$$

(in the right dirn)

So there is no finite critical point of the loss

Q: Where does SGD go?

SGD is

$$w_{t+1} = w_t - \eta \nabla L(w_t)$$
$$= w_t - \eta \sum \ell'(w^T x_i) x_i$$

1: (Soudry et al)
with assumptions,

1) $\lim_{t \to \infty} L(w(t)) = 0$

2) $\lim_{t \to \infty} \|w_t\| = \infty$

3) for all i, $\lim_{t \to \infty} w_t^T x_i = \infty$

Sketch of proof:

- Data is separable, so that there is some w^* s.t. $w^{*T} x_i > 0$ for all i.
- This means

$$w^{*T} \nabla \mathcal{L} = \sum_i \ell'(w^T x_i)(w^{*T} x_i) < 0$$

(cause $\ell'(w) = -e^{-w}$, so neg;)

- so there aren't any finite critical points.
- But GD goes to a critical point — so $\|w_t\| \to \infty$

$$\mathcal{L}(w_t) \to \infty$$

Q: GD goes to ∞; but in what dirn?

we have

$$w_t^T x_i \longrightarrow \infty$$

now consider

$$\frac{w_t}{\|w_t\|}$$

if it converges to some limit w_∞,

we have

$$w_t = g(t)\, w_\infty + \underline{\rho}(t) \quad \leftarrow \text{vector!}$$

where $g(t) \longrightarrow \infty$

$$\lim_{t \to \infty} \frac{\underline{\rho}(t)}{g(t)} = 0$$

$$-\nabla L(w) = \sum_i \exp\left[-w_t^T x_i\right] x_i$$

$$= \sum_i \exp\left[-g(t)\, w_\infty^T x_i\right] \exp\left[-\underline{\rho}(t)^T x_i\right] x_i$$

Now

as $t \to \infty$, consider weight of ith example

$e^{-g(t)\, w_\infty^T x_i}$ $\quad$ $e^{-\rho(t)^T x_i}$

so if $w_\infty^T x_i < w_\infty^T x_j$

$g(t) \to \infty$, the i'th example will have much higher weight.

$\Rightarrow$ eventually, the only significant weights are those associated with examples s, st

$w_\infty^T x_s \leq w_\infty^T x_i$ for all i

There are several of these (how many?)

$\Rightarrow$ w_∞ is a non-neg weighted comb of x_s

~~Yields~~ Consider

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

← scaled so it isn't infinite

We have

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

AND either

$$(\alpha_i \geq 0 \,, \quad \hat{w}^T x_i = 1)$$

OR

$$(\alpha_i = 0 \,, \quad \hat{w}^T x_i > 1)$$

But

- These are KKT for an SVD ($\alpha_i > 0$ for support vectors)
- This should be unexpected — our choice of minimizer has profoundly affected the choice of solution
- Why? - all sol'n's at ∞
  - more than one. (Check!)