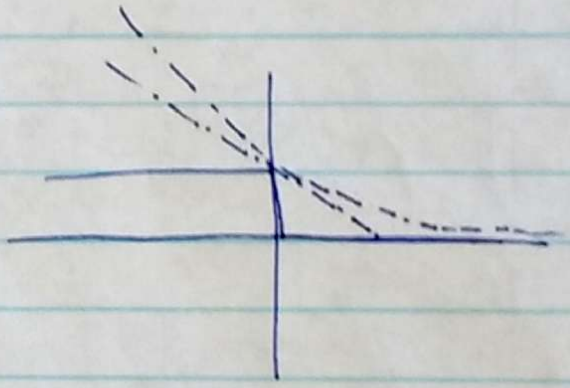
$$J(\theta) = \mathop{\mathbb{E}}_{(x,y)\sim P_{data}} \left[ L(f(x,\theta), y) \right]$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} L\left(f(x^{(i)}, \theta), y^{(i)}\right)$$

Examples:

$$L(y,\hat{y}) = -\text{Sign}(y\hat{y})$$



Surrogate:

$$L(y,\hat{y}) = (1 - y\hat{y})_+$$

$$L(y,\hat{y}) = (1 - y\hat{y})_+^2$$

$$(x)_+ = \begin{cases} x & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$

the surrogate is differentiable and more robust.

— — — —

$$\nabla_\theta J = \mathop{\mathbb{E}}_{(x,y)\sim P_{data}} \left[ \nabla_\theta L(f(x,\theta), y) \right]$$

$$\approx \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta L(f(x,\theta), y)$$

Why we care about "Stochastic Gradient Descend"?

— Time consuming to compute $\nabla_\theta J$ exactly (full pass)

- Variance of estimate $\frac{\sigma}{\sqrt{n}}$ but

  cost increases linearly

- How to choose batch size
  - → learning rate for small batch
  - → large batch may not fit
  - → small batch can function as regularizer

---

Ill-conditioning

$$f(x) \Big|_{x \to x^0} \approx f(x^0) + (x - x^0)^T g + \frac{1}{2}(x - x^0)^T H (x - x^0)$$

$$f(x^0 - \varepsilon g) \approx \underbrace{f(x^0)}_{\text{current value}} - \varepsilon g^T g + \frac{1}{2}\varepsilon^2 g^T H g$$

if $\frac{1}{2}\varepsilon^2 g^T H g$ dominates $\varepsilon g g^T$ it is ill-cond.

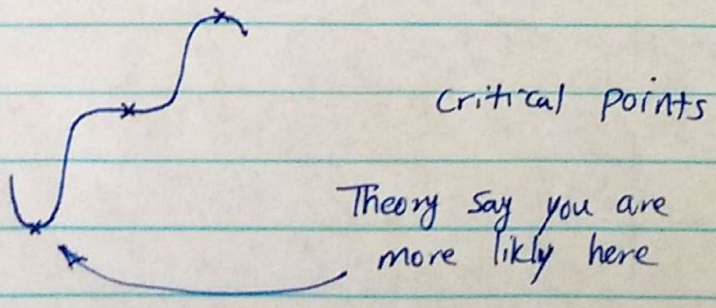
In practice, it might be fine. (Fig 8.1)

In Convex opt., we can handle this with Newton method but not in deep learning → b/c Newton method finds critical point which can be saddle point

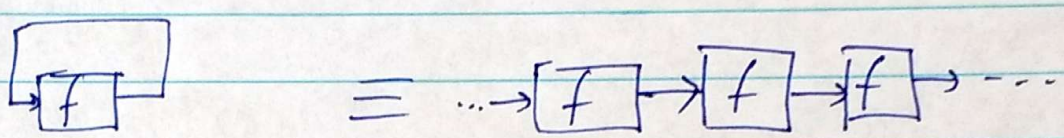
(3)

Local Min.

- Because of symmetry not all local min are bad
- They are bad if they are very different than global -- but we cannot tell
- Take home message: Don't blame everything on local min
- Plot $\|g\|_2^2$
- In high dim they are rare any way b/c we have saddle points more often



Deep recurrent network:



$$h^{T}(\theta) = f(f(f(\cdots f(x)\cdots))$$

$$\nabla h = \nabla f(x) \nabla f(f(x)) \cdots$$

Let's assume at the beging $f(x) = x$

$$\nabla f(x) = \nabla f(f(x)) = \cdots = W$$

$$\nabla h = W^t \qquad W = V \operatorname{diag}(\lambda) V^{-1}$$

$$W^t = V \operatorname{diag}(\lambda^t) V^{-1}$$

if $|\lambda_i| < 1$ $\lambda = \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_n \end{bmatrix}$

$\rightarrow |\lambda_i|^t \rightarrow 0$ vanishing gradient

if $|\lambda_i| > 1$

$\rightarrow |\lambda_i|^t \rightarrow \infty$ exploding gradient

SGD :

$$\theta^{new} \leftarrow \theta^{old} - \varepsilon_k \hat{g}$$

$\hat{g}$ ← gradient

$\varepsilon_k$ ↑ Step length

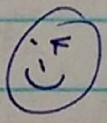
Sufficient Cond. for Conv.

$$\sum_{k=1}^{\infty} \varepsilon_k = \infty \longrightarrow O(k^{-1})$$

$$\sum_{k=1}^{\infty} \varepsilon_k^2 < \infty \longrightarrow O(k^{-\frac{1}{2}})$$

$$O(k^{-\frac{1}{2}}) < \varepsilon_k < O(k^{-1})$$

$$\varepsilon_k = (1 - \tfrac{\tau}{k})\varepsilon_0 + \tfrac{\tau}{k}\varepsilon_\tau$$

Choose $\tau$, $\varepsilon_0$, $\varepsilon_\tau$ Good luck! 

How should we think about SGD

Regret $J(\theta^k) - \min_\theta J(\theta)$ ⟸ How fast it decays

$O(\frac{1}{k})$ for strong convex

$O(k^{-\frac{1}{2}})$ Convex

non-convex ??