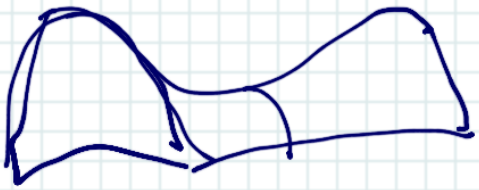
28/2/19

EE6337: Deep Learning

Today:
- Reviews
- Optimization for Training Deep Models
  - Traditional opt vs. opt for DL
  - Challenges: ill conditioning, local min, flat regions and saddle pts, exploding grads
  - Basic methods: SGD, momentum, Nesterov momentum



Stochastic Gradient Descent (SGD):

- Input: $\gamma(r)$; $r = 1$
- While stopping condition not met do
  - pick m training samples randomly from $\{(x_i, y_i)\}_{i=1}^{N}$
  - Compute gradient: $\nabla_\theta [R(\theta)] = \nabla_\theta \left[\sum_{i=1}^{m} R_i(\theta)\right]$
  - Update parameters: $\theta^{(r+1)} = \theta^{(r)} - \gamma^{(r)} \cdot \nabla_\theta$
  - $r = r+1$

Choice of $\gamma^{(r)}$ can be based on a heuristic rule that reduces $\gamma^{(r)}$ until it reaches some value $\gamma(k)$ and leaves it constant at $\gamma(k)$ for $r > k$.

Momentum: $\left[\begin{array}{l}\theta^{(r+1)} = \theta^{(r)} + v^{(r)} \\ v^{(r)} = \alpha v^{(r-1)} - \varepsilon \nabla_\theta^{(r)} R(\theta)\end{array}\right]$ $\alpha \in [0, 1)$

- Input ε, $r = 1$
- While stopping condition not met, do
  - Compute gradient
  - $v^{(r)} = \alpha \cdot v^{(r-1)} - \varepsilon \nabla_\theta^{(r)} R(\theta)$ ← momentum update
  - $\theta^{(r)} = \theta^{(r-1)} + v^{(r)}$.
  - $r = r + 1$