Gradient checking

- $General\ principle:$
  - $\frac{\partial f(\theta_1,\ldots,\theta_i,\ldots,\theta_n)}{\partial\theta_i} = \lim_{\epsilon\to 0}\frac{f(\theta_1,\ldots,\theta_i+\epsilon,\ldots,\theta_n)-f(\theta_1,\ldots,\theta_i-\epsilon,\ldots,\theta_n)}{2\epsilon}$
- **concatenate** $W^{[1]}, b^{[1]}\ldots, W^{[L]}, b^{[L]}$ and reshape into a vector $\Theta = \theta_1,\ldots,\theta_L$
  - $\mathscr{J}(W^{[1]}, b^{[1]},\ldots, W^{[L]}, b^{[L]}) = \mathscr{J}(\Theta) = \mathscr{J}(\theta_1,\ldots,\theta_L)$
- reshape $dW^{[1]}, db^{[1]}\ldots, dW^{[L]}, db^{[L]}$ into a vector $d\Theta$
- Is $d\Theta$ the gradient of the cost function $\mathscr{J}(\Theta)$

Implement grad check

- $for\ each\ i:$
  - $d\Theta^{[i]}_{approx} = \frac{\mathscr{J}(\theta_1,\ldots,\theta_i+\epsilon,\ldots,\theta_L)-\mathscr{J}(\theta_1,\ldots,\theta_i-\epsilon,\ldots,\theta_L)}{2\epsilon} \approx d\Theta^{[i]} = \frac{\partial\mathscr{J}}{\partial\theta_i}$
- $d\Theta_{approx} \approx^{?} d\Theta$
  - check $if\ \frac{\|d\Theta_{approx}-d\Theta\|_2}{\|d\Theta_{approx}\|_2+\|d\Theta\|_2} \approx \begin{cases} 10^{-7}, & good \\ 10^{-5} & \\ 10^{-3}, & worry \end{cases}, when\ \epsilon = 10^{-7}$

Notes

- **Don't use in training, only to debug**
- If algorithm fails, look at components $d\Theta^{[i]}_{approx}$ and $d\Theta^{[i]}$ to find out differences
- Remember regularization
  - $\mathscr{J}(w^{[1]}, b^{[1]},\ldots, w^{[L]}, b^{[L]}) = \frac{1}{m}\sum_{i=1}^{m}\mathscr{L}(\hat{y}^{(i)}, y^{(i)}) + \frac{\lambda}{2m}\sum_{l=1}^{L}\|w^{[l]}\|^2$
  - $d\Theta = grads\ of\ \mathscr{J}$
- Don't work with dropout
- When $w, b \approx 0$ grads check is fine, but fails after some iteration
  - Run grad check at random initialization
  - perhaps run again after some training