Coursera Deep Learning Course 2-Week 2: Optimization Methods

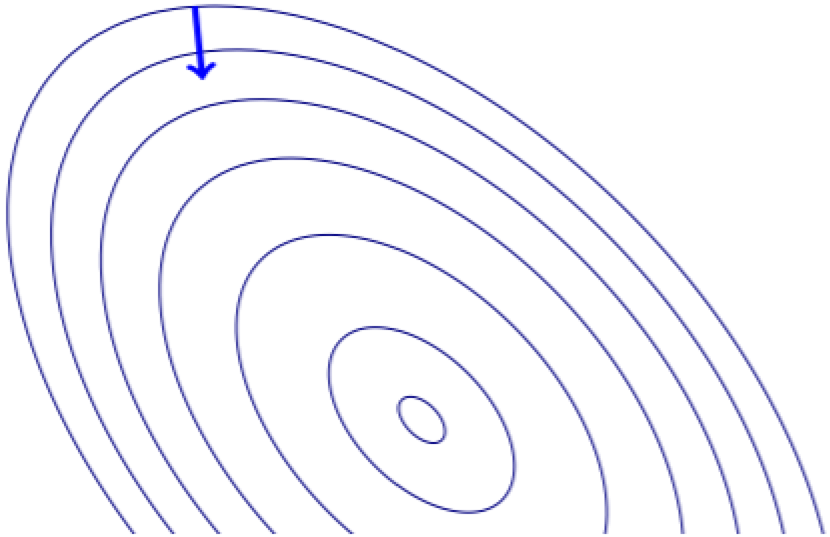
Kang Zhao

2018.09.06

Course Contents

- Mini-batch gradient descent
- Gradient Descent with momentum
- RMSprop
- Adam
- Learning Rate Decay

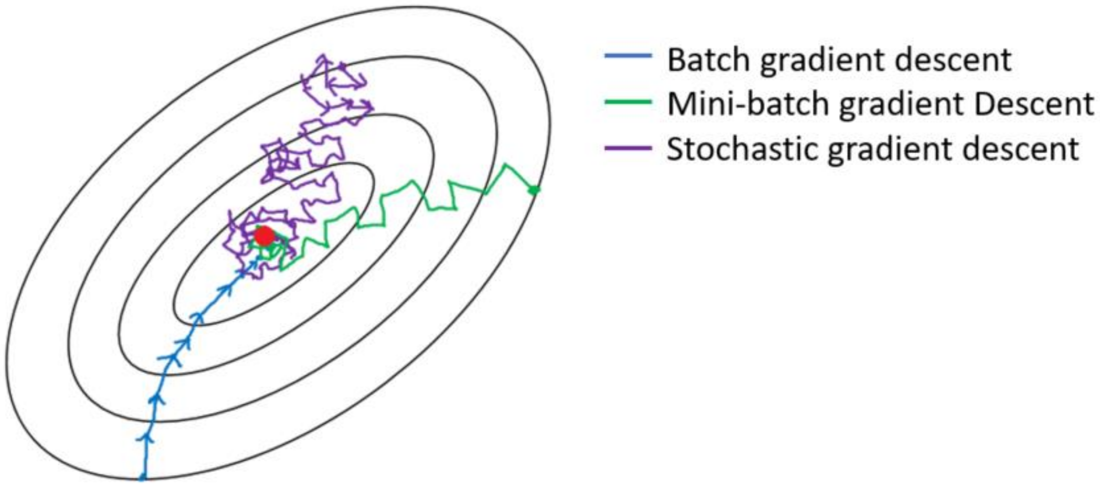
Optimization Overview



Gradient Descent

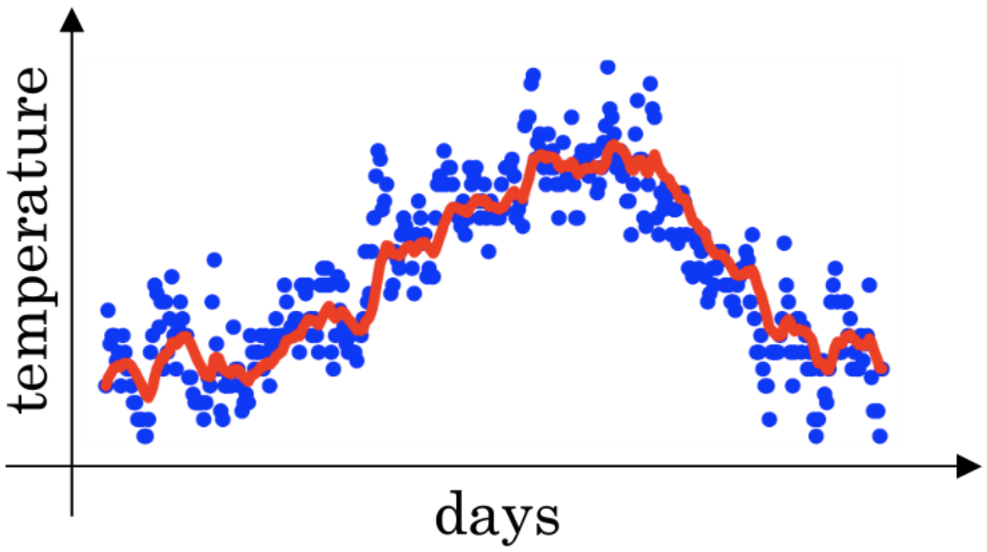
Mini-Batch & SGD

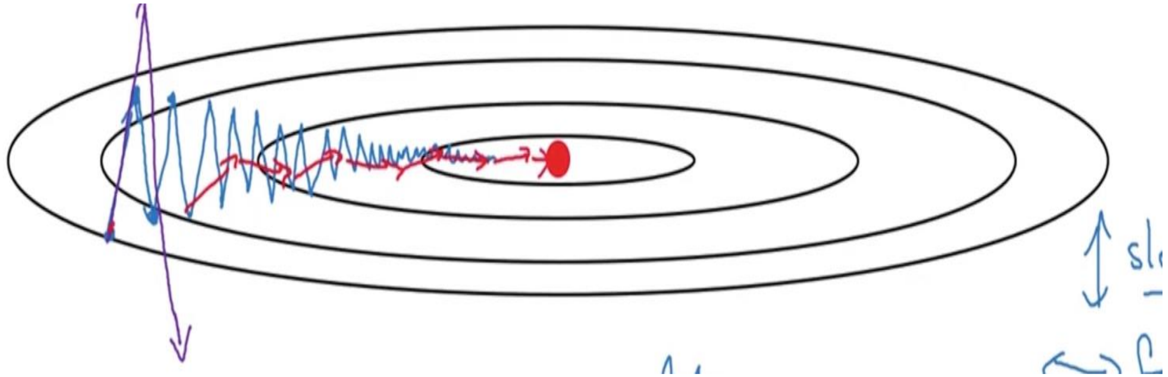
- Stochastic Gradient Descent
  - Train only one sample at one epoch
- Batch Gradient Descent
  - Train all the samples at one epoch
- Mini Batch Gradient Descent
  - Train k samples together at one epoch



Mini-batch size: K

Exponentially Weighted (Moving) Average





Momentum & RMSprop & Adam

$$\left.\begin{aligned} Vd_w &= \beta_1 * Vd_w + (1-\beta_1) * d_w \\ Vd_b &= \beta_1 * Vd_b + (1-\beta_1) * d_b \end{aligned}\right\} \text{“momentum”-like update}$$

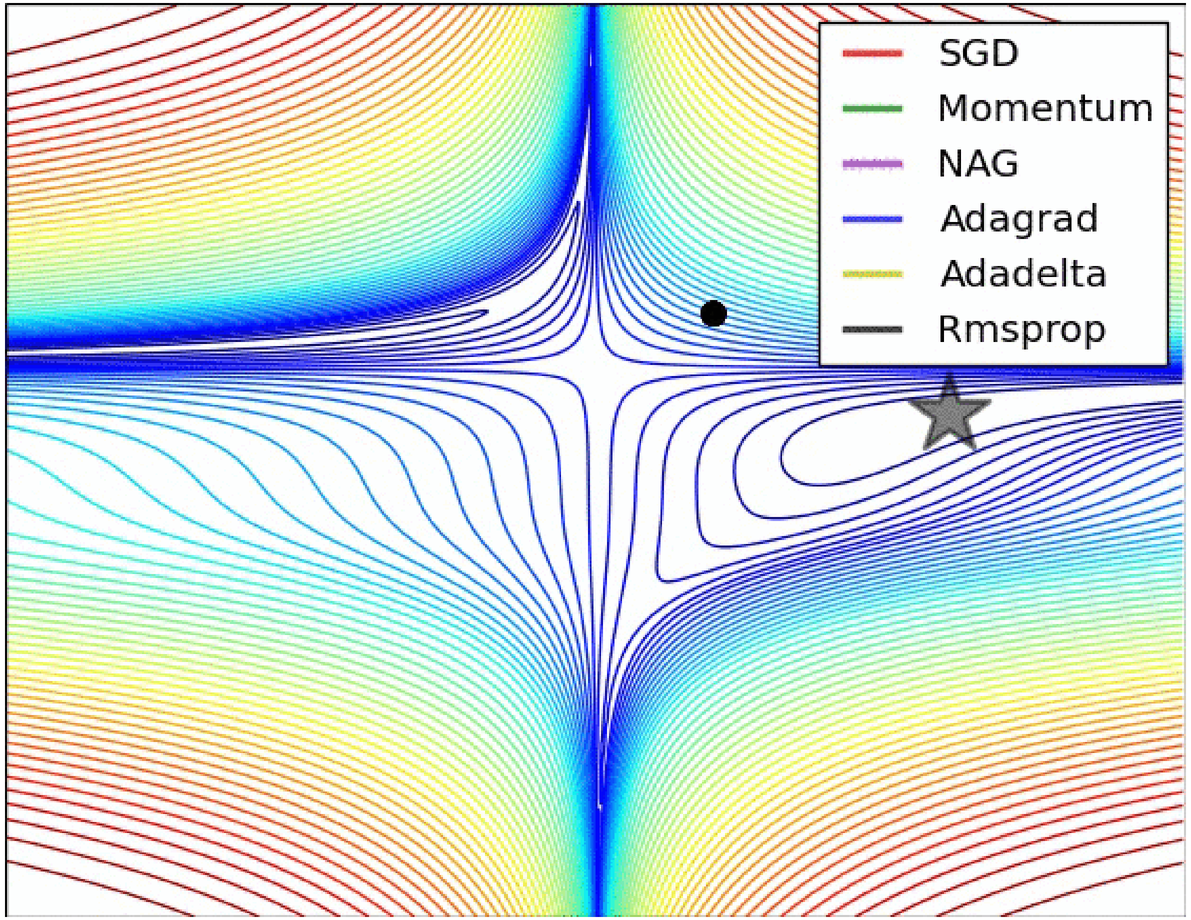
$$\left.\begin{aligned} Sd_w &= \beta_2 * Sd_w + (1-\beta_2) * {d_w}^2 \\ Sd_b &= \beta_2 * Sd_b + (1-\beta_2) * {d_b}^2 \end{aligned}\right\} \text{“RMSprop”}$$

$$V_{d_w}^{corrected} = Vd_w \,/\, (1-\beta_1^t)\ , \qquad V_{d_b}^{corrected} = Vd_b \,/\, (1-\beta_1^t)$$
$$S_{d_w}^{corrected} = Sd_w \,/\, (1-\beta_2^t)\ , \qquad S_{d_b}^{corrected} = Sd_b \,/\, (1-\beta_2^t)$$

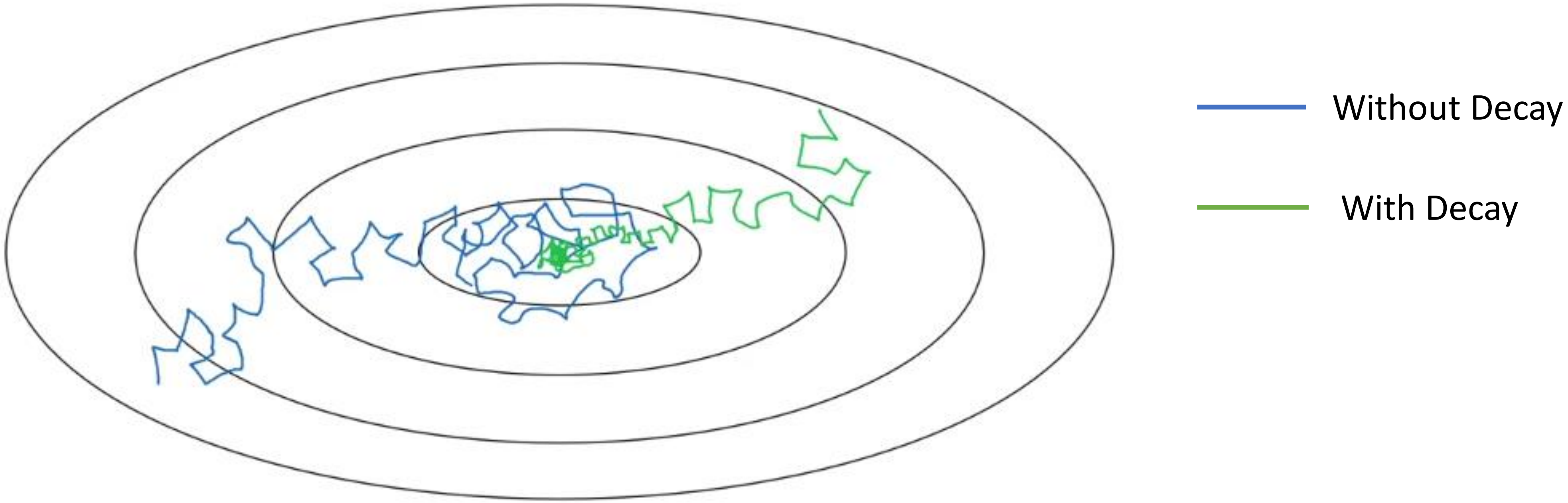
$$w := w - \alpha * \frac{V_{d_w}^{corrected}}{\sqrt{S_{d_w}^{corrected} + \varepsilon}}$$
$$b := b - \alpha * \frac{V_{d_b}^{corrected}}{\sqrt{S_{d_b}^{corrected} + \varepsilon}}$$

Exponential Weight: $\beta 1, \beta 2, \varepsilon$

Momentum & RMSprop & Adam



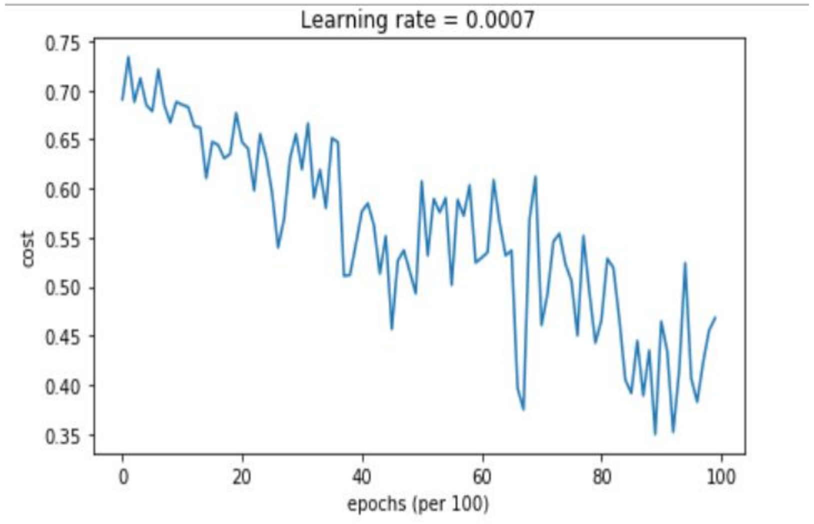
Learning Rate Decay

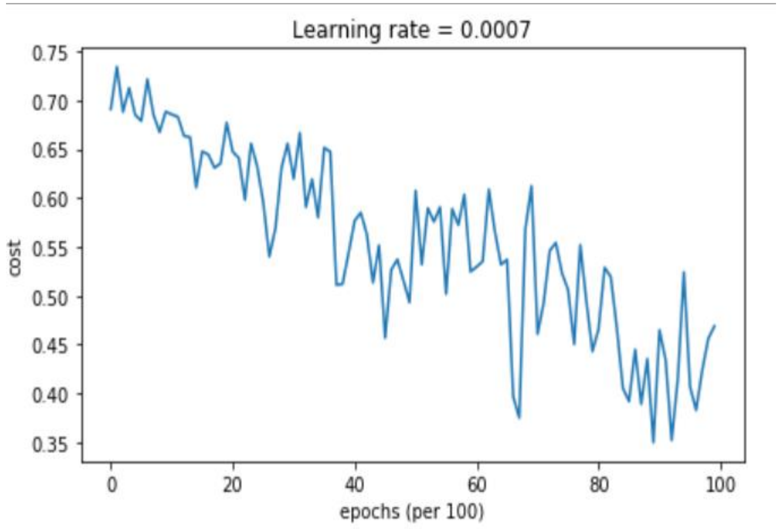


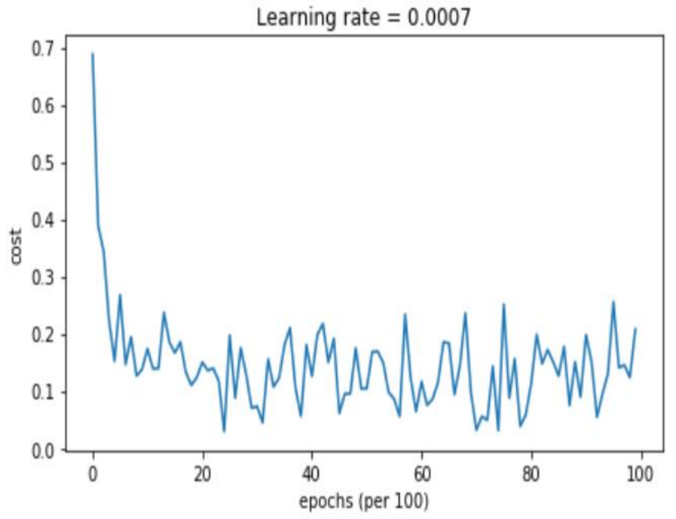
Learning Rate Decay: $\alpha = f(\alpha, epoch)$

Assignments

- Gradient Descent
- Mini-Batch Gradient Descent
- Momentum
- Adam
- Model with different optimization algorithms


Learning rate = 0.0007

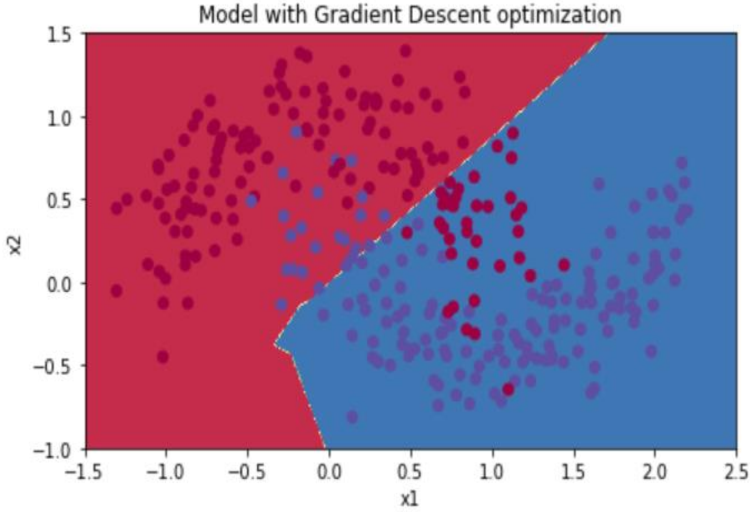

Learning rate = 0.0007

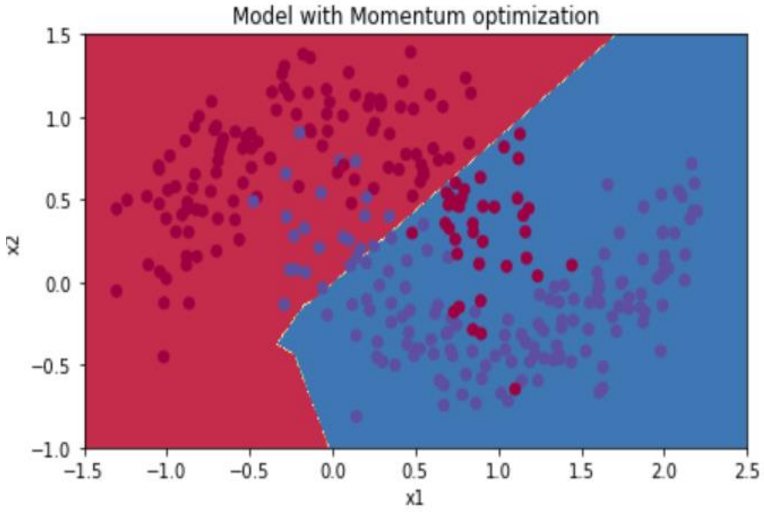

Learning rate = 0.0007

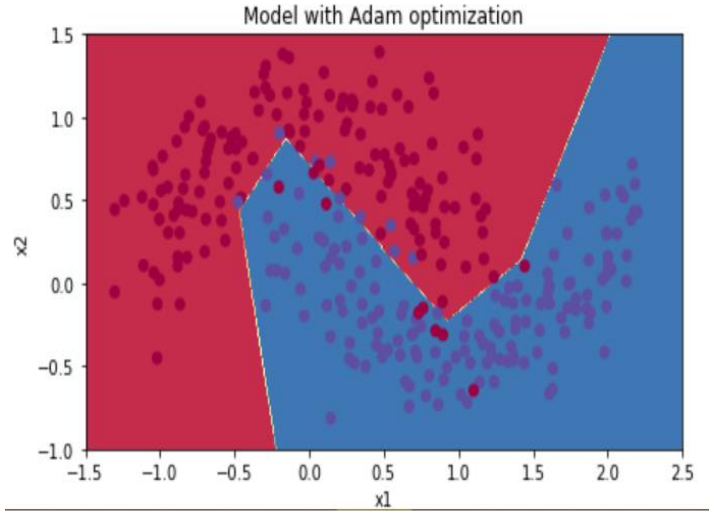
```
Accuracy: 0.796666666667
```

```
ccuracy: 0.796666666667
```

```
Accuracy: 0.94
```


Model with Gradient Descent optimization


Model with Momentum optimization


Model with Adam optimization