# Optimization Technique in Training Deep Models

Chapter 8

# Outline

- Problem definition
- First order and second order methods
- SGD and momentum techniques in deep models
- Function properties in optimization
- Discussions

# Background

- Machine learning problems

loss function

$$J(\boldsymbol{\theta}) = \mathbb{E}_{(\boldsymbol{x},\mathrm{y})\sim\hat{p}_{\text{data}}} L(f(\boldsymbol{x};\boldsymbol{\theta}), y),$$

feature label mapping model parameter

- Empirical risk minimization with independence

$$\min \mathbb{E}_{\boldsymbol{x},\mathrm{y}\sim\hat{p}_{\text{data}}(\boldsymbol{x},y)}[L(f(\boldsymbol{x};\boldsymbol{\theta}), y)] = \frac{1}{m}\sum_{i=1}^{m} L(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}), y^{(i)}),$$

What if the data are not independent?

number of samples

# Risk Minimization

- True risk in machine learning

$$\min \ \mathbb{E}_{\boldsymbol{x},\mathrm{y}\sim\hat{p}_{\mathrm{data}}(\boldsymbol{x},y)}[L(f(\boldsymbol{x};\boldsymbol{\theta}),y)] \qquad (1)$$

- Empirical risk minimization

$$\min \ \frac{1}{m}\sum_{i=1}^{m} L(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}),y^{(i)}) \qquad (2)$$

- Generalization error

$$\mathrm{P}[(1)-(2)\geq t]\leq \exp(-t), \forall t\geq 0$$

Zhang, Tong. "Data Dependent Concentration Bounds for Sequential Prediction Algorithms." In COLT, pp. 173-187. 2005.

# Risk Minimization

- True risk in machine learning

$$\min \ \mathbb{E}_{\boldsymbol{x},\mathrm{y}\sim\hat{p}_{\text{data}}(\boldsymbol{x},y)}[L(f(\boldsymbol{x};\boldsymbol{\theta}),y)] \qquad (1)$$

- Empirical risk minimization

$$\min \ \frac{1}{m}\sum_{i=1}^{m} L(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}),y^{(i)}) \qquad (2)$$

- Generalization error

What is the effect of m?

$$\left\{\begin{array}{l} O(\frac{1}{m}) \\ O(\frac{1}{m^2}) \end{array}\right.$$

$$\mathrm{P}[(1)-(2)\geq t]\leq \exp(-t), \forall t\geq 0$$

Zhang, Tong. "Data Dependent Concentration Bounds for Sequential Prediction Algorithms." In COLT, pp. 173-187. 2005.
Zhang, Lijun, Tianbao Yang, and Rong Jin. "Empirical Risk Minimization for Stochastic Convex Optimization: $ O (1/n) $- and $ O (1/n^ 2) $-type of Risk Bounds." In COLT, 2017.

# Problem Definition

- Empirical risk minimization (ERM)

$$\min \frac{1}{m}\sum_{i=1}^{m} L(f(\boldsymbol{x}^{(i)};\boldsymbol{\theta}), y^{(i)})$$

- Solution

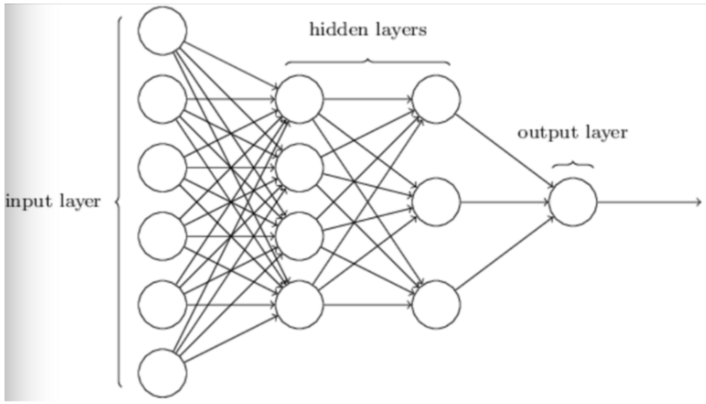
$$\boldsymbol{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\mathrm{model}}(\boldsymbol{x}^{(i)}, y^{(i)}; \boldsymbol{\theta})$$

- Square loss
  - Gaussian distribution in model errors
- Tools
  - SVM
  - Neural networks

# From ERM to Deep Learning

- Model of feedforward neural network

$$Y(\theta, X) = \underbrace{\theta_h \times \theta_{h-1} \times \theta_1}_{\theta} \times X$$




- Batch learning

$$\min \frac{1}{m}\sum_{i=1}^{m} L(f(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}), y^{(i)})$$

- First order and second order in optimization

# First Order Method

- Gradient (a vector)

$$g = \nabla_{\theta} J^*(\theta) = \sum_{x} \sum_{y} p_{\text{data}}(x, y) \nabla_{\theta} L(f(x; \theta), y).$$

mapping is pre-defined

- Cons
  - Time consuming for each iteration
  - Not linear convergence rate
    - Convex case: $O\left(\frac{1}{T}\right)$ with $T$ being the total iteration
    - Acceleration case: $O\left(\frac{1}{T^2}\right)$
  - How to set the learning rate
- Pros
  - Exact gradient information

Nesterov, Yurii. Introductory lectures on convex optimization: A basic course. Vol. 87. Springer Science & Business Media, 2013.

# Second Order Method

- Hessian matrix (a square matrix)

$$H_{i,j} = \frac{\partial g_i}{\partial \theta_j}$$

- Cons
  - Ill-conditioning of matrix (zero eigenvalue)
  - Time consuming in each iteration, or even failure, in calculating the inverse of Hessian matrix
- Pros
  - linear convergence rate
    - Strongly convex case: $O(\rho^T)$ with $0 < \rho < 1$

$\Rightarrow$ $T = O(\ln(\frac{1}{\epsilon}))$ with $\epsilon$ being the accuracy

# Update Rules

- First order method

Learning rate

$$\theta_{t+1} = \theta_t - \eta g_t$$

- Second order method

$$\theta_{t+1} = \theta_t - H_t^{-1} g_t$$

1. Optimal learning rate
2. The inverse is not easily to solve
3. Estimation error of Hessian leads to large deviation in training of deep models

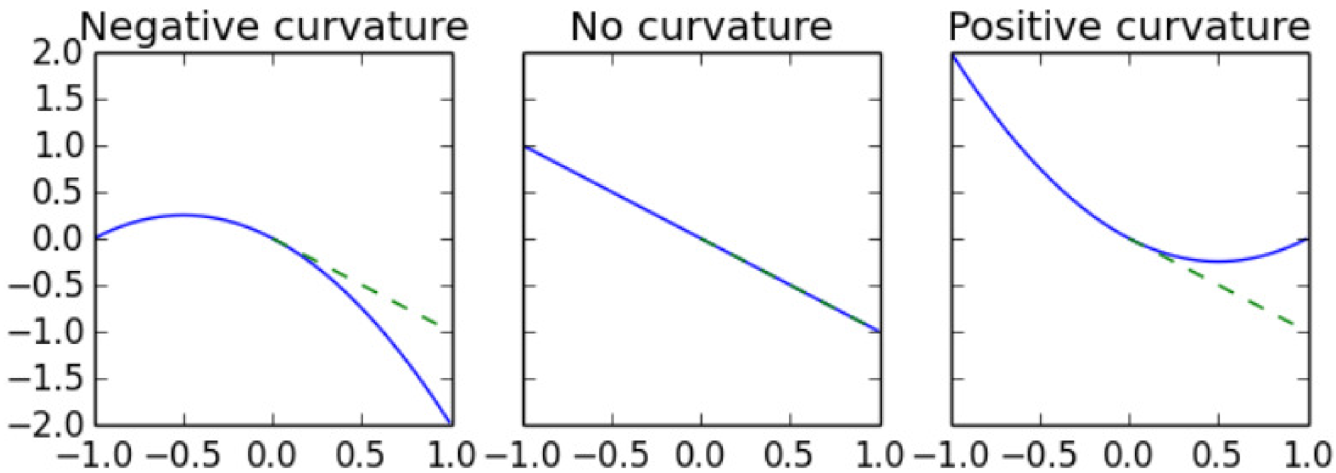
# Taylor Series Approximation

- Function approximation

$$f(x) = f(x_0) + (x - x_0) f'(x) + \frac{1}{2}(x - x_0)^2 f''(x) + \dots$$

- ERM problem

$$J(\boldsymbol{\theta}) = J(\boldsymbol{\theta}_0) + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \boldsymbol{g} + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^\top \boldsymbol{H} (\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \dots$$



# SGD in Deep Learning

- Stochastic gradient descent (SGD)

$$\theta_{t+1} = \theta_t - \eta \hat{g}_t$$

  $\hat{g}_t$ is calculated based on: 1) one sample (online learning); 2) a small subset of samples (mini batch)

- Pros
  - Improve the efficiency in each iteration
- Cons
  - Noise in the estimation of gradient

How to deal with this issue?

# An Ideal Assumption

- Gradient in (conditional) expectation

$$g_t = \mathrm{E}[\hat{g}_t] \Rightarrow g_t = \mathrm{E}[\hat{g}_t | \boldsymbol{F}_{t-1}]$$

Why this assumption works?

# An Ideal Assumption

- Gradient in (conditional) expectation

$$g_t = \mathrm{E}[\hat{g}_t] \Rightarrow g_t = \mathrm{E}[\hat{g}_t | \boldsymbol{F}_{t-1}]$$

- Stochastic (convex) optimization

$$\min F(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}\left[f_t(\theta)\right]$$

  – Convex convergence rate: $O(\frac{1}{\sqrt{T}})$

  – Acceleration rate: $O\left(\frac{1}{T}\right)$ for strongly convex and smooth function

Shalev-Shwartz, Shai, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. "Stochastic Convex Optimization." In COLT. 2009.

Hazan, Elad, and Satyen Kale. "Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization." Journal of Machine Learning Research 15, no. 1 (2014): 2489-2512.

# A Realistic Assumption

- Gradient in (conditional) expectation

$$g_t = \mathrm{E}[\hat{g}_t] \Rightarrow g_t = \mathrm{E}[\hat{g}_t | \boldsymbol{F}_{t-1}]$$

- Stochastic (convex) optimization

$$\min F(\theta) = \frac{1}{m}\sum_{i=1}^{m} \mathrm{E}\left[f_t(\theta)\right]$$

$$\left|\left|\mathrm{E}[\hat{g}_t] - g_t\right|\right| \leq B$$

  – Convex convergence rate: $O(\frac{1}{\sqrt{T}})$

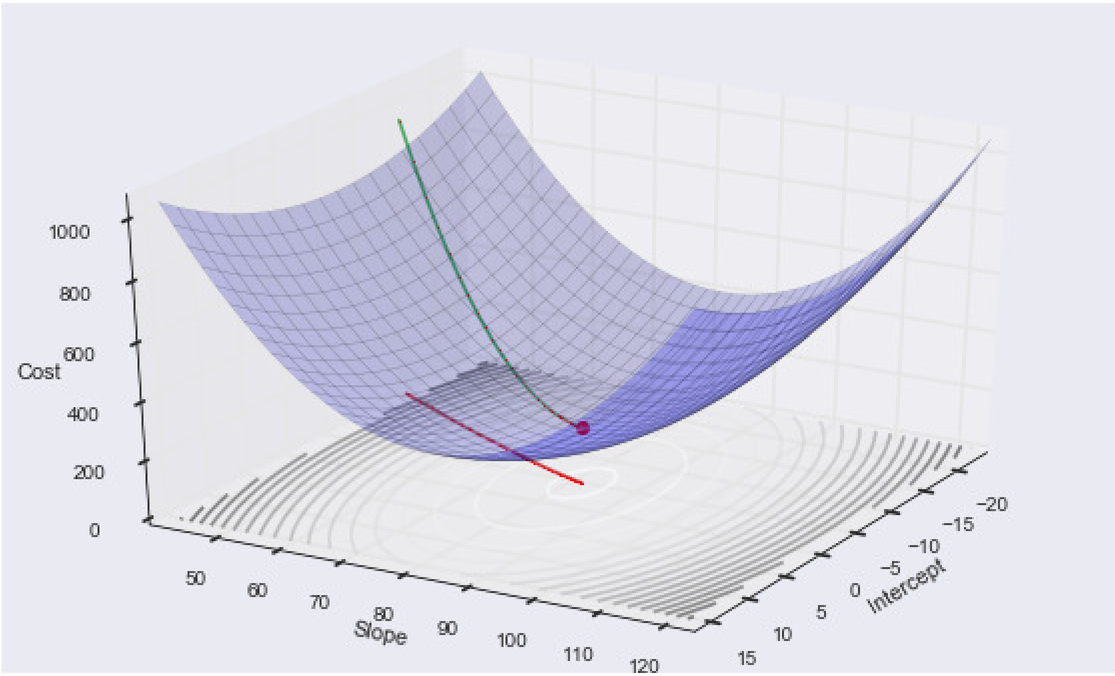
  – Acceleration rate: $O\left(\frac{1}{T}\right)$ for strongly convex and smooth function

Shalev-Shwartz, Shai, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. "Stochastic Convex Optimization." In COLT. 2009.

Hazan, Elad, and Satyen Kale. "Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization." Journal of Machine Learning Research 15, no. 1 (2014): 2489-2512.

# An Illustration

- A Comparison of GD and SGD

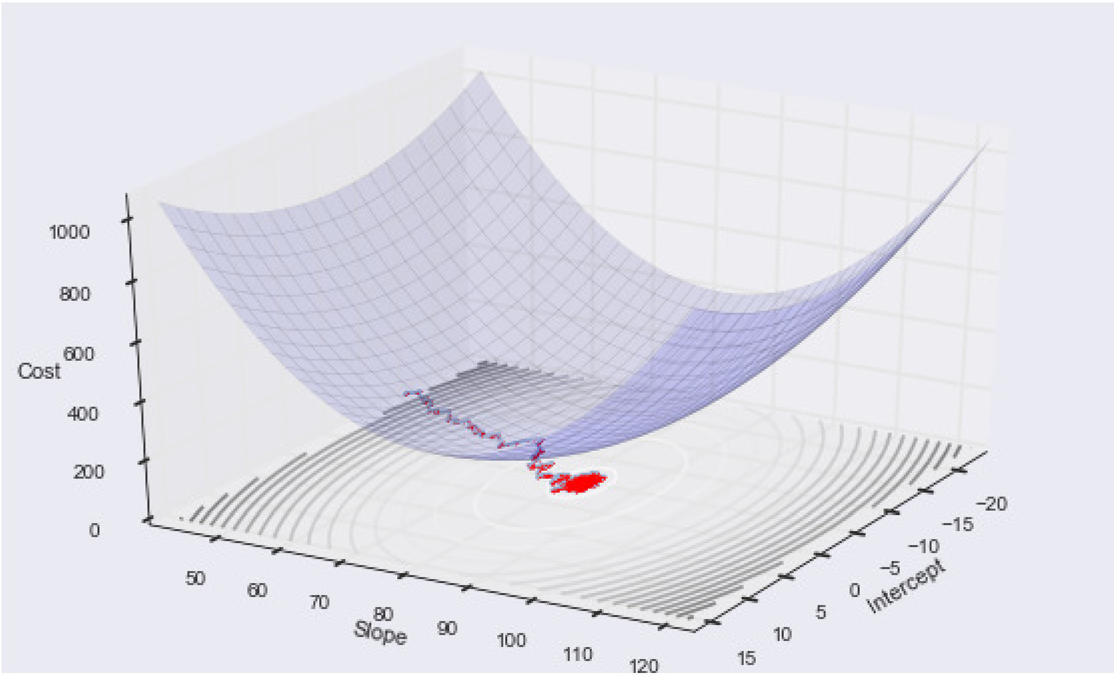



GD: O(1/T)

https://am207.github.io/2017/wiki/gradientdescent.html

# An Illustration

- A Comparison of GD and SGD




SGD: O(1/T^0.5)

https://am207.github.io/2017/wiki/gradientdescent.html

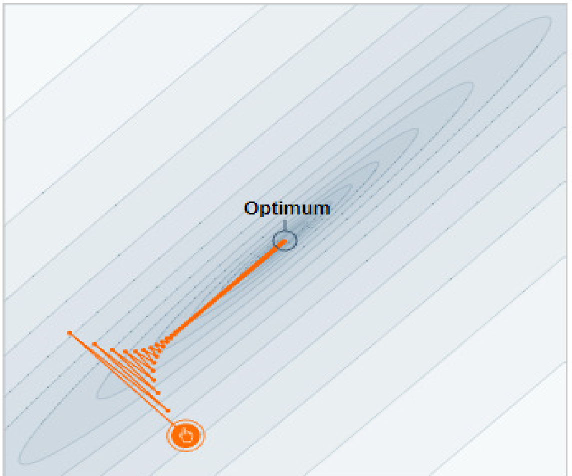
# Ill-Conditioning of Objectives

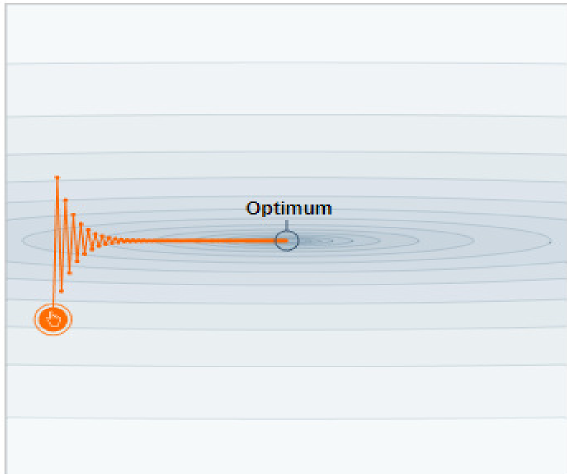
- Ill-conditioning is general in deep models
  - Metric: condition number

Largest eigenvalue of Hessian matrix

$$r = \frac{\lambda_{\max}(H)}{\lambda_{\min}(H)}$$

  - Large $r$ means ill-condition




https://distill.pub/2017/momentum/

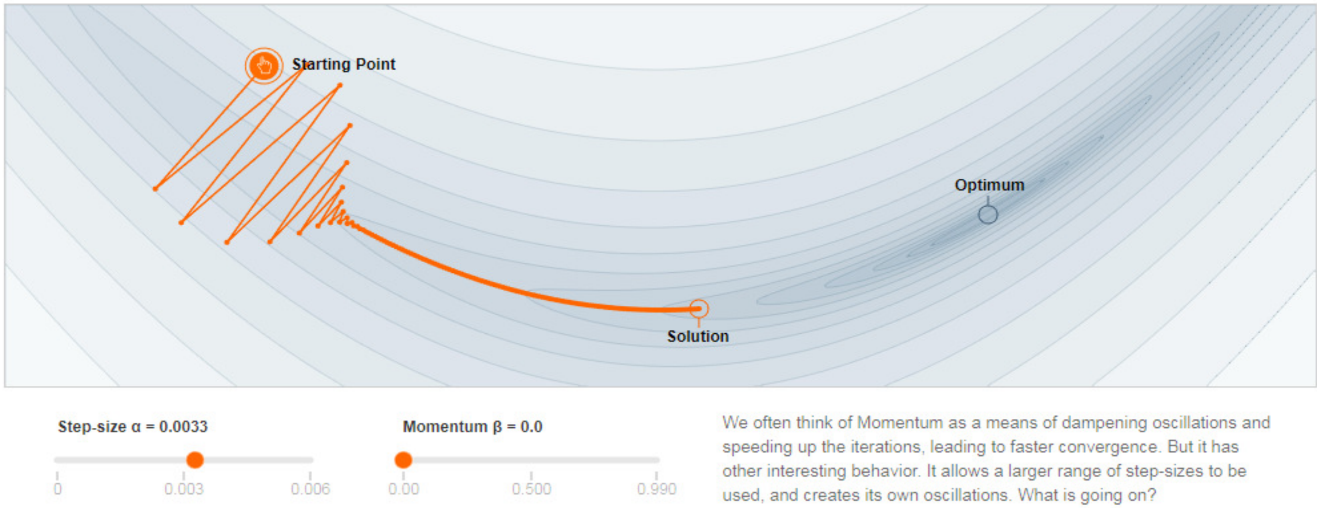
# Momentum

- Classical method

A new vector

$$v_{t+1} = \mu v_t - \eta g_t(\theta_t), \mu \in [0,1]$$

$$\theta_{t+1} = \theta_t + v_{t+1}$$




https://distill.pub/2017/momentum/

Polyak, Boris T. "Some methods of speeding up the convergence of iteration methods." USSR Computational Mathematics and Mathematical Physics 4, no. 5 (1964): 1-17.

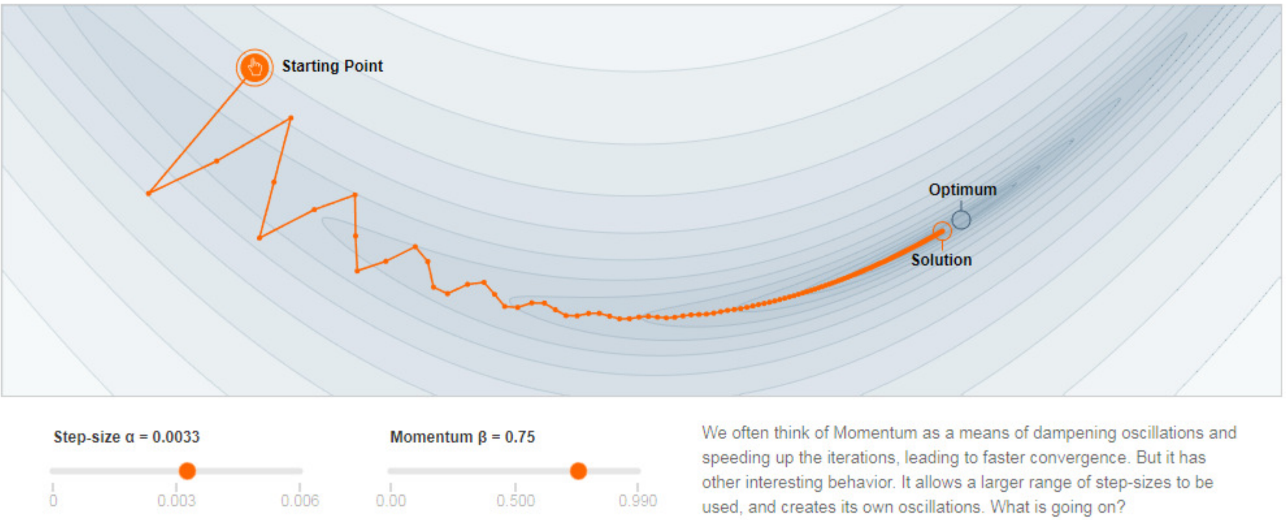
# Momentum

- Classical method

A new vector

$$v_{t+1} = \mu v_t - \eta g_t(\theta_t), \mu \in [0,1]$$

$$\theta_{t+1} = \theta_t + v_{t+1}$$




https://distill.pub/2017/momentum/

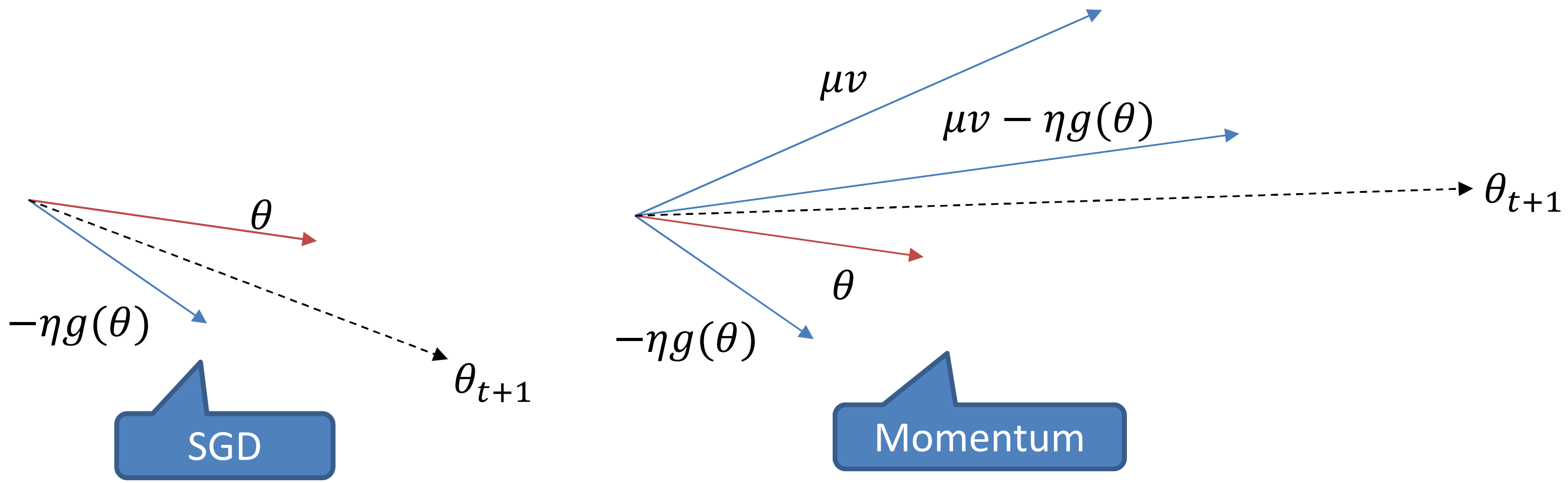
Polyak, Boris T. "Some methods of speeding up the convergence of iteration methods." USSR Computational Mathematics and Mathematical Physics 4, no. 5 (1964): 1-17.

# Momentum Intuition

- Classical method

$$v_{t+1} = \mu v_t - \eta g_t(\theta_t), \mu \in [0,1]$$

$$\theta_{t+1} = \theta_t + v_{t+1}$$




Sutskever, Ilya, James Martens, George Dahl, and Geoffrey Hinton. "On the importance of initialization and momentum in deep learning." In International conference on machine learning, pp. 1139-1147. 2013.

# Momentum Intuition

- Classical method

$$v_{t+1} = \mu v_t - \eta g_t(\theta_t), \mu \in [0,1]$$

$$\theta_{t+1} = \theta_t + v_{t+1}$$

- Pros
  - Partially solve ill-conditioning
  - Help to adjust the learning rate
  - Faster convergence, and less oscillation
  - Set $\mu = 0$, we have GD/SGD
- Cons
  - A new parameter

# Nesterov Accelerated Gradient

- Update rules

$$v_{t+1} = \mu v_t - \eta \boxed{g_t(\theta_t + \mu v_t)}, \mu \in [0,1]$$

$$\theta_{t+1} = \theta_t + v_{t+1}$$

- Convergence rate in convex case
  - $O(1/T^2)$

Nesterov, Yurii. "A method of solving a convex programming problem with convergence rate O (1/k2)." In Soviet Mathematics Doklady, vol. 27, no. 2, pp. 372-376. 1983.

# Momentum in Deep Learning

- Some deep models
  - SGD cannot obtain good performance
  - Try momentum technique
- Random initialization
  - Good performance in FNN and RNN
  - Constant initialization leads to failure of training

Sutskever, Ilya, James Martens, George Dahl, and Geoffrey Hinton. "On the importance of initialization and momentum in deep learning." In International conference on machine learning, pp. 1139-1147. 2013.

# Function Properties in Optimization

- Revisit SGD

$$\theta_{t+1} = \theta_t - \eta \hat{g}_t$$

What if g has some good properties?

# Function Properties in Optimization

- Revisit SGD

$$\theta_{t+1} = \theta_t - \eta \hat{g}_t$$

- Bernstein condition

Optimal parameter

$$\mathrm{E}\left[L\left(x^i, y^i, \theta_t\right)^2\right] \leq B\left(\mathrm{E}\left[L\left(x^i, y^i, \theta_t\right) - L\left(x^i, y^i, \theta^*\right)\right]\right)^\gamma$$

- Convergence rate
  - $O(T^{\frac{1-\gamma}{2-\gamma}-1})$ with $\gamma \in [0,1]$

Van Erven, Tim, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. "Fast rates in statistical and online learning." Journal of Machine Learning Research 16 (2015): 1793-1861.

# Holderian Error Bound

- Local Holderian error bound

$$||\theta_t - \theta^*|| \leq C\left(\mathrm{E}\left[L\left(x^i, y^i, \theta_t\right)\right] - \mathrm{E}\left[L\left(x^i, y^i, \theta^*\right)\right]\right)^{\gamma}$$

- Convergence rate
  - $O(T^{\frac{1-\gamma}{2-\gamma}-1})$ with $\gamma \in [0,1]$

Can one design deep models to have this property?

Xu, Yi, Qihang Lin, and Tianbao Yang. "Stochastic Convex Optimization: Faster Local Growth Implies Faster Global Convergence." In International Conference on Machine Learning, pp. 3821-3830. 2017.

# Discussions

- ERM problem in deep models
- Optimization to solve ERM problem
- SGD and momentum
- Function properties help to solve optimization

# References


1. Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016.
2. Zhang, Tong. "Data Dependent Concentration Bounds for Sequential Prediction Algorithms." In COLT, pp. 173-187. 2005.
3. Zhang, Lijun, Tianbao Yang, and Rong Jin. "Empirical Risk Minimization for Stochastic Convex Optimization: $ O (1/n) $-and $ O (1/n^ 2) $-type of Risk Bounds." In COLT, 2017.
4. Nesterov, Yurii. Introductory lectures on convex optimization: A basic course. Vol. 87. Springer Science & Business Media, 2013.
5. Shalev-Shwartz, Shai, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. "Stochastic Convex Optimization." In COLT. 2009.
6. Hazan, Elad, and Satyen Kale. "Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization." Journal of Machine Learning Research 15, no. 1 (2014): 2489-2512.
7. Polyak, Boris T. "Some methods of speeding up the convergence of iteration methods." USSR Computational Mathematics and Mathematical Physics 4, no. 5 (1964): 1-17.
8. Nesterov, Yurii. "A method of solving a convex programming problem with convergence rate O (1/k2)." In Soviet Mathematics Doklady, vol. 27, no. 2, pp. 372-376. 1983.
9. Sutskever, Ilya, James Martens, George Dahl, and Geoffrey Hinton. "On the importance of initialization and momentum in deep learning." In International conference on machine learning, pp. 1139-1147. 2013.
10. Van Erven, Tim, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. "Fast rates in statistical and online learning." Journal of Machine Learning Research 16 (2015): 1793-1861.
11. Xu, Yi, Qihang Lin, and Tianbao Yang. "Stochastic Convex Optimization: Faster Local Growth Implies Faster Global Convergence." In International Conference on Machine Learning, pp. 3821-3830. 2017.