Methods for Stochastic Optimisation

AdaGrad, RMSProp & Adam

Bigi Varghese Philip

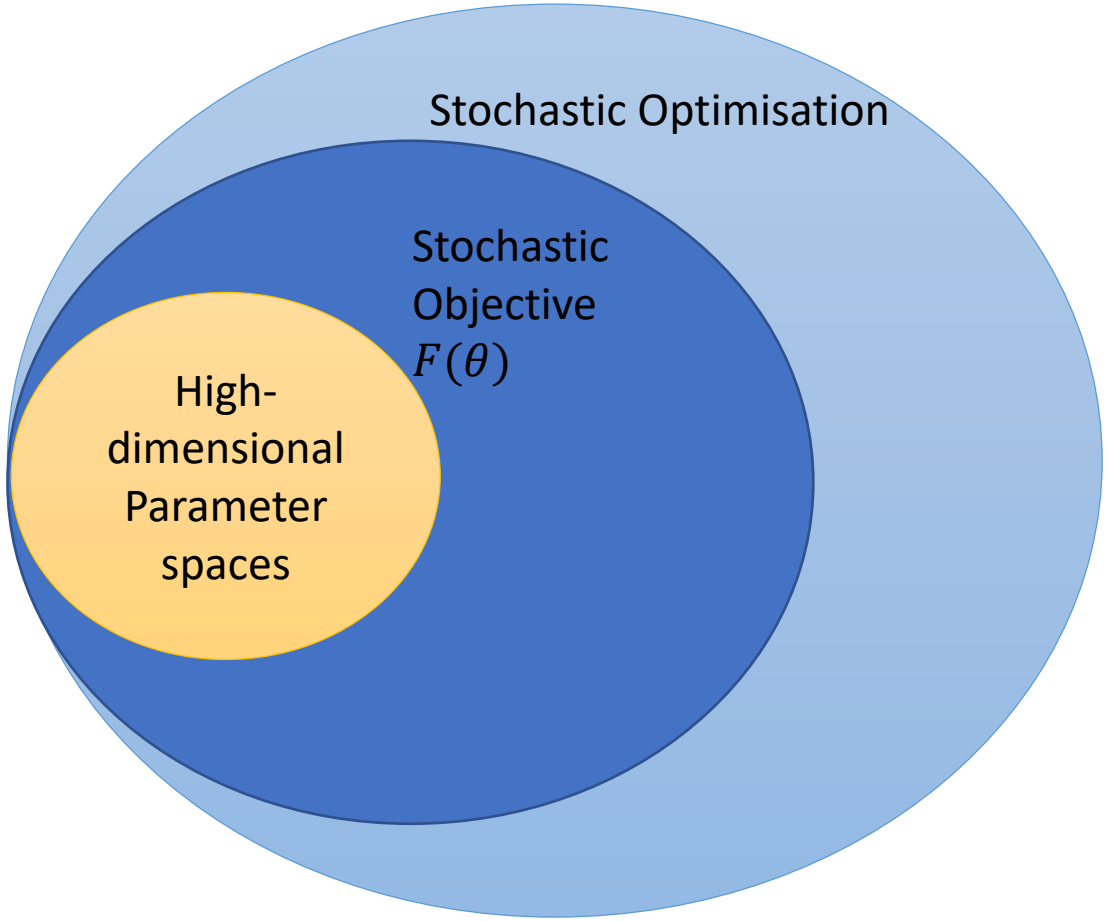
Contents

- Why Stochastic?
- Improvement on Gradient Descent
  - GD with Momentum
  - AdaGrad
  - RMS Prop
- ADAM: Adaptive Moment Estimation

Why Stochastic?

- Possible reasons for stochasticity :
  - Inherent function noise
  - Evaluation at random subsamples
    - Stochastic/mini-batch Vs Batch GD
  - Suitable for Online Learning

- $Min\ E[F(\theta)]$ w.r.to θ
- Perspective Algorithms should average gradients over Mini-batches



Ref: Srebro, Nathan, and Ambuj Tewari. "Stochastic optimization for machine learning." *ICML Tutorial* (2010).

Gradient Descent Improvements

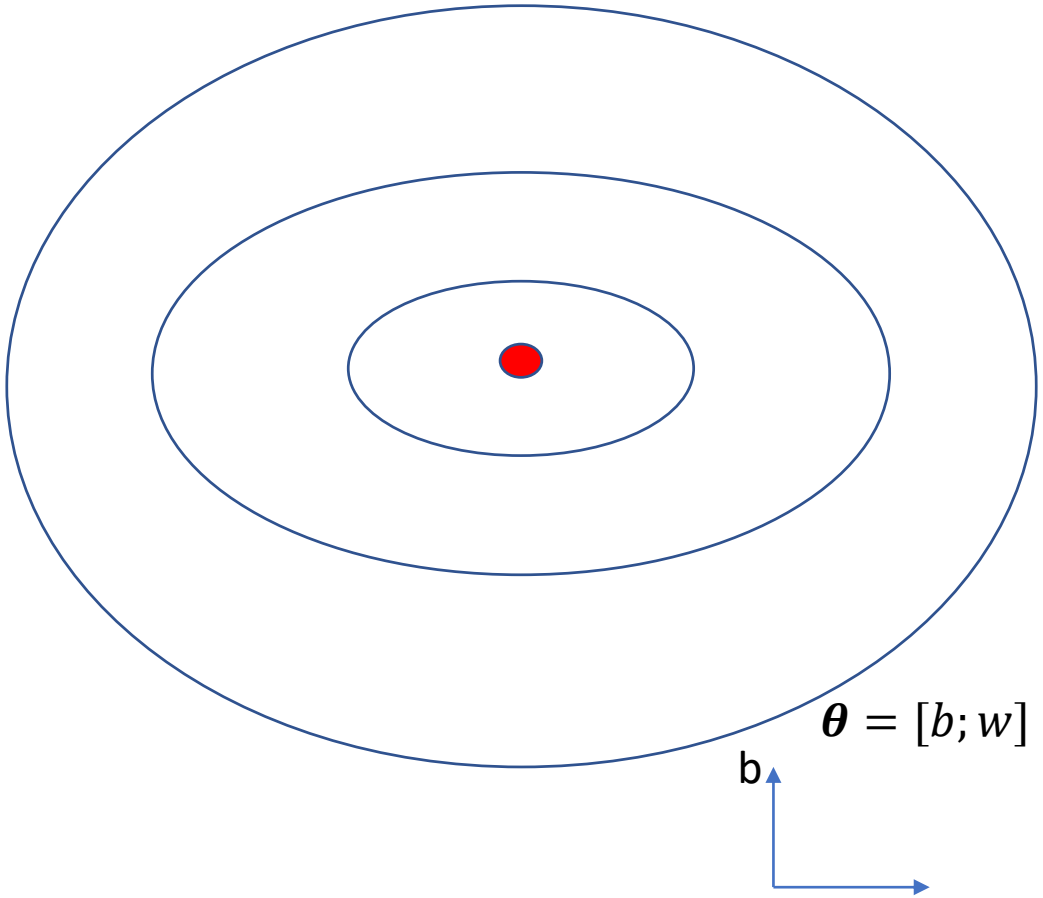
- The Problem with GD:

$$\boldsymbol{\theta_t} = \boldsymbol{\theta}_{t-1} - \alpha_t \nabla F_{t-1}$$

  - Oscillations
  - Stuck at flat portions
- Solutions :
  - Adaptive learning rates
  - Perspective Algorithms should average gradients over Mini-batches
  - General Pattern:

$$\boldsymbol{\theta_t} = \boldsymbol{\theta}_{t-1} - \alpha G(\nabla F_{t-1} \dots \nabla F_{0,} \, higher \; moments)$$

- E.g.
  - Gradient Descent With Momentum
  - AdaGrad
  - RMS Prop
  - Adam and Many more
- Why no simple/ Universal soln.?



Ref:
Deep Learning, Andrew Ng (video Lectures)
Lecture 6.5 — Rmsprop: normalize the gradient [Neural Networks for Machine Learning]

Gradient Descent Improvements

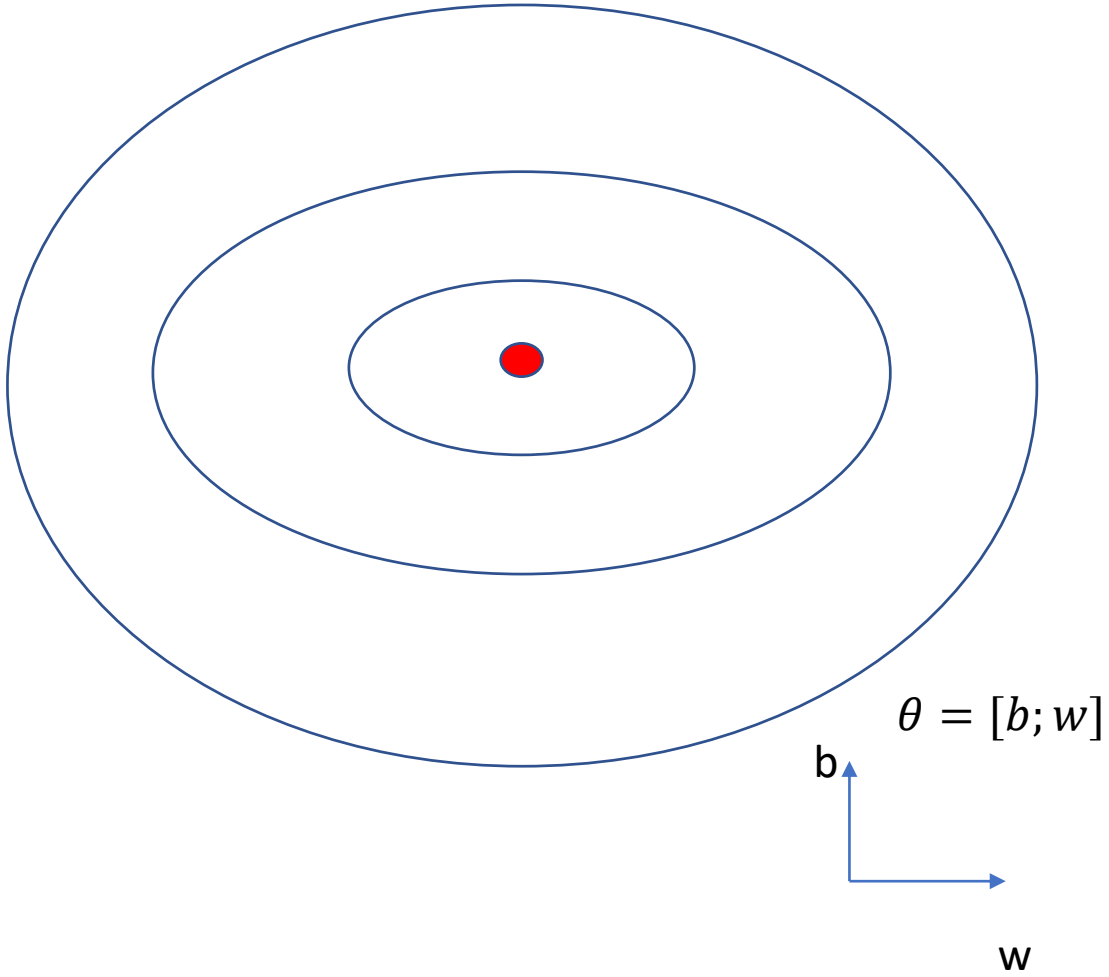
- The Problem with GD:

$$\boldsymbol{\theta_t = \theta_{t-1} - \alpha_t \nabla F_{t-1}}$$

- Gradient Descent With Momentum:
  - Motivation: to average out the gradient to reduce oscillations and deal with flat regions

$$\boldsymbol{m_t = \beta m_{t-1} + (1-\beta)\nabla F_{t-1}}$$
$$\boldsymbol{\theta_t = \theta_{t-1} - \alpha m_t}$$

  - Initialization: $\boldsymbol{m_0 = 0}, < \beta = 0.9 >$



Ref: Deep Learning, Andrew Ng (video Lectures)

Gradient Descent Improvements

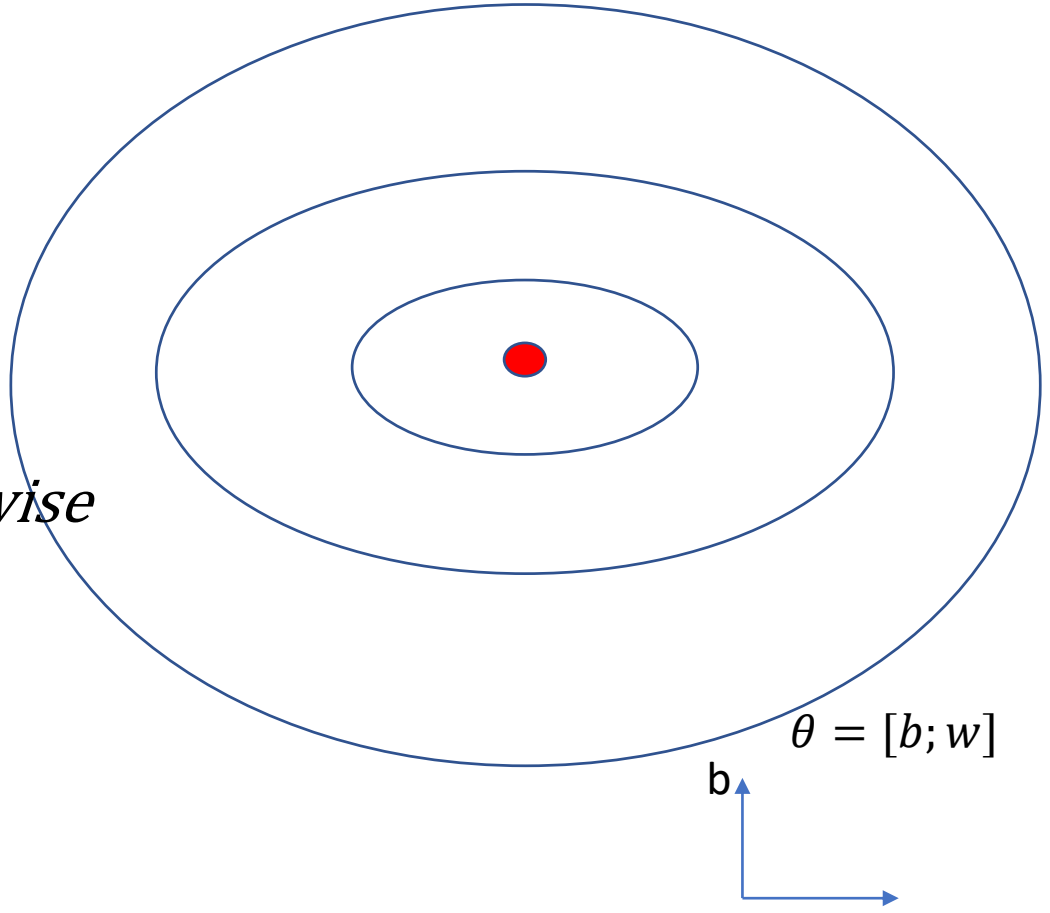
- The Problem with GD:

$$\theta_t = \theta_{t-1} - \alpha_t \nabla F_{t-1}$$

- Gradient Descent With Momentum
- AdaGrad
  - Motivation: Adaptively scaling the gradient

$$v_t = \sum \nabla F_i^2 \mid i = 0 \; to \; t-1 \rightarrow Element\ wise$$

$$\theta_t = \theta_{t-1} - \frac{\alpha\{\nabla F_{t-1}\}}{\sqrt{v_t} + \epsilon}$$



Ref: Deep Learning, Andrew Ng (video Lectures)

Duchi, John, Elad Hazan, and Yoram Singer. "Adaptive subgradient methods for online learning and stochastic optimization." Journal of machine learning research 12.7 (2011).

Gradient Descent Improvements

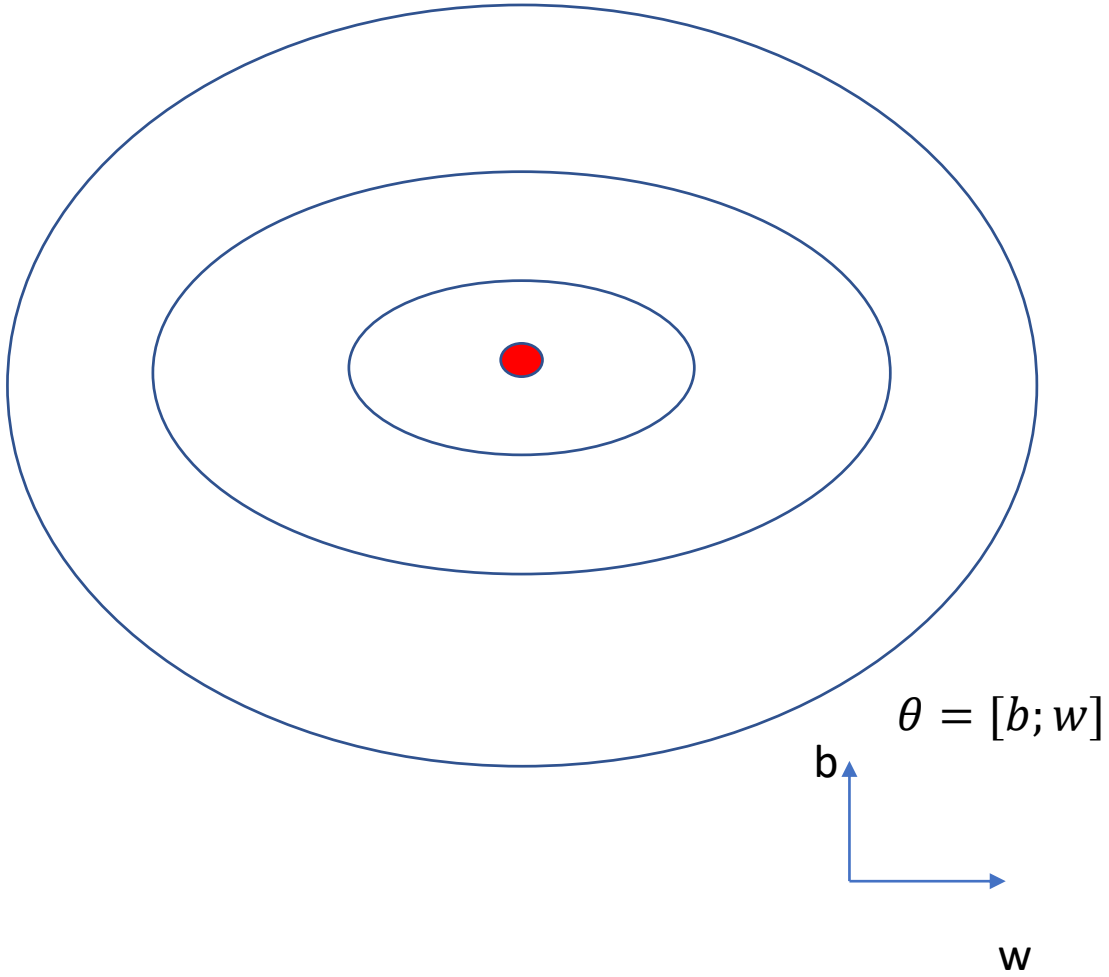
- The Problem with GD:

$$\theta_t = \theta_{t-1} - \alpha \nabla F_{t-1}$$

- Gradient Descent With Momentum
- AdaGrad
- RMSProp

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) \nabla F_{t-1}^2$$

$$\theta_t = \theta_{t-1} - \frac{\alpha \{\nabla F_{t-1}\}}{\sqrt{v_t}}$$



Ref: Deep Learning, Andrew Ng (video Lectures)

Lecture 6.5 — Rmsprop: normalize the gradient [Neural Networks for Machine Learning]

Gradient Descent Improvements

- Gradient Descent With Momentum: Simple, Suitable for online algorithm
- AdaGrad: Works well with sparse gradients, popular for linear, less complex models
- RMSProp : Works fine for online and Non-stationary settings

Ref: Deep Learning, Andrew Ng (video Lectures)

Adam: Adaptive Moment Estimation

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
$m_0 \leftarrow 0$ (Initialize 1st moment vector)
$v_0 \leftarrow 0$ (Initialize 2nd moment vector)
$t \leftarrow 0$ (Initialize timestep)
while θ_t not converged **do**
$t \leftarrow t + 1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
$\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
$\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)
end while
return θ_t (Resulting parameters)

- Initialisations:
  $m_0 = \mathbf{0}, v_0 = \mathbf{0}$
- Bias correction to avoid bias towards **0**
- ϵ to avoid divide by zero

Ref: Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

Adam: Adaptive Moment Estimation

- Recommended Parameters

$\beta_1 = 0.9$

$\beta_2 = 0.999$

$\epsilon = 10^{-8}$

α Needs tuning.

- Update Rule:

(Step-size upper bounds)

$$|\Delta t| = \alpha . \frac{m_t}{\sqrt{v_t}}$$

$|\Delta t| \leq \alpha . \frac{1-\beta_1}{\sqrt{1-\beta_2}}$, iff $\frac{1-\beta_1}{\sqrt{1-\beta_2}} > 1$

$\leq \alpha, Otherwise$

Require: α: Stepsize
Require: $\beta_1, \beta_2 \in [0, 1)$: Exponential decay rates for the moment estimates
Require: $f(\theta)$: Stochastic objective function with parameters θ
Require: θ_0: Initial parameter vector
$m_0 \leftarrow 0$ (Initialize 1st moment vector)
$v_0 \leftarrow 0$ (Initialize 2nd moment vector)
$t \leftarrow 0$ (Initialize timestep)
while θ_t not converged **do**
$t \leftarrow t + 1$
$g_t \leftarrow \nabla_\theta f_t(\theta_{t-1})$ (Get gradients w.r.t. stochastic objective at timestep t)
$m_t \leftarrow \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$ (Update biased first moment estimate)
$v_t \leftarrow \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$ (Update biased second raw moment estimate)
$\widehat{m}_t \leftarrow m_t / (1 - \beta_1^t)$ (Compute bias-corrected first moment estimate)
$\widehat{v}_t \leftarrow v_t / (1 - \beta_2^t)$ (Compute bias-corrected second raw moment estimate)
$\theta_t \leftarrow \theta_{t-1} - \alpha \cdot \widehat{m}_t / (\sqrt{\widehat{v}_t} + \epsilon)$ (Update parameters)
end while
return θ_t (Resulting parameters)

Ref: Kingma, Diederik P., and Jimmy Ba. "Adam: A method for stochastic optimization." arXiv preprint arXiv:1412.6980 (2014).

Thank You !!

Let us discuss …