Optimization in Machine Learning

Lecture 15: Algorithms for Optimization, Convergence Analysis of Gradient Descent under Lipschitz Continuity and Convexity, Enhancements via Smoothness and Strong Convexity

Ganesh Ramakrishnan

Department of Computer Science
Dept of CSE, IIT Bombay
https://www.cse.iitb.ac.in/~ganesh

March, 2025

[Recap] Convergence Rate For Smooth + Strongly Convex

- Setting $R^2 = ||x_0 - x^*||^2$, we get:

$$f(x_T) - f(x^*) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T R^2$$

- To get an error of ϵ, we require $\frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T R^2 \leq \epsilon$ which implies $T \geq \frac{L}{\mu}\log(\frac{R^2 L}{2\epsilon})$.
- To get an error of $\epsilon = 0.01$, we now need only $L/\mu \log(50R^2L)$ iterations as opposed to $50R^2L$ iterations in the smooth case!

Summary of Results so Far...(with convexity by default)

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then, we have the following:
  - $\epsilon = 0.1$, C: 10000, S $=$ 50, SS $=$ 8.49 iterations
  - $\epsilon = 0.01$, C: 1000000, S $=$ 500, SS $=$ 13.49 iterations
  - $\epsilon = 0.001$, C: 100000000, S $=$ 5000, SS $=$ 18.49 iterations
- As ϵ reduces by 10, the number of iterations of strongly + smooth case increases only by a additive constant! This is linear convergence!

Summary of Results so Far...(with convexity by default)

worst case for some intermediate ineq.

- Lipschitz continuous functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, achieve an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Smooth Functions (S): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{R^2 L}{\epsilon}$ iterations.
- Smooth + Strongly Convex (SS): With $\gamma = 1/L$, achieve an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2 L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then, we have the following:
  - $\epsilon = 0.1$, C: 10000, S = 50, SS = 8.49 iterations
  - $\epsilon = 0.01$, C: 1000000 ($*100$), S = 500 ($*10$), SS = 13.49 iterations
  - $\epsilon = 0.001$, C: 100000000 ($*100$), S = 5000 ($*10$), SS = 18.49 iterations

constant step increase for 10 fold higher precision

constant step increase for 10 fold higher precision

- As ϵ reduces by 10, the number of iterations of strongly + smooth case increases only by a additive constant! This is linear convergence!

Revisist optional slides and notice that this is Q-linear convergence

3/107

Can we do better for Lipschitz Continuous Functions?

- What if we have strong convexity? Can a function be **both** Lipschitz Continuous and Strongly Convex?

Can we do better for Lipschitz Continuous Functions?

- What if we have strong convexity? Can a function be **both** Lipschitz Continuous and Strongly Convex?

Very often (especially in ML) we cannot insist on differentiability
==> We cannot insist on L-smoothness
But we could add a regularizer and get strong convexity nevertheless!

Can we do better for Lipschitz Continuous Functions?

- What if we have strong convexity? Can a function be **both** Lipschitz Continuous and Strongly Convex?
- Unfortunately No!!!!

Unless we consider restriction of the fn

Can we do better for Lipschitz Continuous Functions?

- What if we have strong convexity? Can a function be **both** Lipschitz Continuous and Strongly Convex?
- Unfortunately No!!!!
- But on a bounded set, we can assume that they are (and the gradients are upper bounded)

Can we do better for Lipschitz Continuous Functions?

- What if we have strong convexity? Can a function be **both** Lipschitz Continuous and Strongly Convex?
- Unfortunately No!!!!
- But on a bounded set, we can assume that they are (and the gradients are upper bounded)

Within a bounded set eg: $\|w\|^2 \leq R^2$

Very often (especially in ML) we cannot insist on differentiability
==> We cannot insist on L-smoothness
But we could add a regularizer and get strong convexity nevertheless!

$$\text{Hinge}(w^T x, y) + \lambda \|w\|^2$$
$$\text{s.t. } \|w\|^2 \leq R$$

Can we do better for Lipschitz Continuous Functions?

- What if we have strong convexity? Can a function be **both** Lipschitz Continuous and Strongly Convex?
- Unfortunately No!!!!
- But on a bounded set, we can assume that they are (and the gradients are upper bounded)
- Can we get an improved convergence rate in such a case?

improvement upon $O(\frac{1}{\epsilon^2})$

Can we do better for Lipschitz Continuous Functions?

- What if we have strong convexity? Can a function be **both** Lipschitz Continuous and Strongly Convex?
- Unfortunately No!!!!
- But on a bounded set, we can assume that they are (and the gradients are upper bounded)
- Can we get an improved convergence rate in such a case?
- We can obtain an improved $\boxed{O(1/\epsilon)}$ bound!!

Lipschitz Continuity + Strong Convexity I

- Recall: $g_t^T(x_t - x^*) = \frac{\gamma_t}{2}||g_t||^2 + \frac{1}{2\gamma_t}\left(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2\right)$
- Recall: Using stronger lower bound on the above expression(s) via strong convexity:
  $g_t^T(x_t - x^*) \geq f(x_t) - f(x^*) + \frac{\mu}{2}||x_t - x^*||^2$
- Combining, we obtained:
  $||x_{t+1} - x^*||^2 \leq 2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 + (1 - \mu\gamma)||x_t - x^*||^2$

- Recall: $g_t^T(x_t - x^*) = \frac{\gamma_t}{2}||g_t||^2 + \frac{1}{2\gamma_t}\left(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2\right)$ } Algebra
- Recall: Using stronger lower bound on the above expression(s) via strong convexity:

$g_t^T(x_t - x^*) \geq f(x_t) - f(x^*) + \frac{\mu}{2}||x_t - x^*||^2$ } we used this for L-smoothness only.

- Combining, we obtained:

$$\left[\,||x_{t+1} - x^*||^2 \leq \underline{2\gamma\,(f(x^*) - f(x_t))} + \overset{\mathbf{O}}{\underline{\gamma^2||g_t||^2}} + \underline{(1-\mu\gamma)||x_t - x^*||^2}\,\right] * \frac{1}{2\gamma}$$

$\leq \gamma^2 B^2$

Lipschitz Continuity + Strong Convexity I

- Recall: $g_t^T(x_t - x^*) = \frac{\gamma_t}{2}||g_t||^2 + \frac{1}{2\gamma_t}\left(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2\right)$
- Recall: Using stronger lower bound on the above expression(s) via strong convexity:
  $g_t^T(x_t - x^*) \geq f(x_t) - f(x^*) + \frac{\mu}{2}||x_t - x^*||^2$
- Combining, we obtained:
  $||x_{t+1} - x^*||^2 \leq 2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 + (1 - \mu\gamma)||x_t - x^*||^2$
- Now, with Lipschitz continuity, assume the gradients $||g_t|| \leq B \implies ||g_t||^2 \leq B^2$. Combining with strong convexity we get,

$$f(x_t) - f(x^*) \leq \frac{B^2\gamma_t}{2} + \left(\frac{1}{2\gamma_t} - \frac{\mu}{2}\right)||x_t - x^*|| - \frac{1}{2\gamma_t}||x_{t+1} - x^*||^2$$

Lipschitz Continuity + Strong Convexity I

- Recall: $g_t^T(x_t - x^*) = \frac{\gamma_t}{2}||g_t||^2 + \frac{1}{2\gamma_t}\left(||x_t - x^*||^2 - ||x_{t+1} - x^*||^2\right)$
- Recall: Using stronger lower bound on the above expression(s) via strong convexity:

$$g_t^T(x_t - x^*) \geq f(x_t) - f(x^*) + \frac{\mu}{2}||x_t - x^*||^2$$

- Combining, we obtained:

$$||x_{t+1} - x^*||^2 \leq 2\gamma\left(f(x^*) - f(x_t)\right) + \gamma^2||g_t||^2 + (1 - \mu\gamma)||x_t - x^*||^2$$

- Now, with Lipschitz continuity, assume the gradients $||g_t|| \leq B \implies ||g_t||^2 \leq B^2$. Combining with strong convexity we get,

To LHS

≤ 0 To make this disappear do I need specific choice of γ_t?

$$f(x_t) - f(x^*) \leq \frac{B^2\gamma_t}{2} + \left(\frac{1}{2\gamma_t} - \frac{\mu}{2}\right)||x_t - x^*|| - \frac{1}{2\gamma_t}||x_{t+1} - x^*||^2$$

We will do a scheme with γ_t to make things work! t-specific

Telescopic summing?

Lipschitz Continuity + Strong Convexity II

- So Far:

$$\boxed{f(x_t) - f(x^*)} \leq \boxed{\frac{B^2\gamma_t}{2}} + \boxed{\left(\frac{1}{2\gamma_t} - \frac{\mu}{2}\right) ||x_t - x^*||} - \boxed{\frac{1}{2\gamma_t}||x_{t+1} - x^*||^2}$$

- So Far:

$$\sum_t f(x_t) - f(x^*) \leq \frac{B^2\gamma_t}{2} + \left(\frac{1}{2\gamma_t} - \frac{\mu}{2}\right)||x_t - x^*|| - \frac{1}{2\gamma_t}||x_{t+1} - x^*||^2$$

if $\gamma_t \propto \frac{1}{t}$

Lipschitz Continuity + Strong Convexity II

- So Far:

$$f(x_t) - f(x^*) \leq \frac{B^2\gamma_t}{2} + \left(\frac{1}{2\gamma_t} - \frac{\mu}{2}\right)||x_t - x^*|| - \frac{1}{2\gamma_t}||x_{t+1} - x^*||^2$$

- For the specific case of $\gamma_t^{-1} = \mu(1+t)/2$ and multiplying both sides by t:

$$t\left[f(x_t) - f(x^*)\right] \leq \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\left\{t(t-1)\,||x_t - x^*||^2 - (t+1)t\,||x_{t+1} - x^*||^2\right\}$$

$$\leq \frac{B^2}{\mu} + \frac{\mu}{4}\{t(t-1)\,||x_t - x^*||^2 - (t+1)t\,||x_{t+1} - x^*||^2\}$$

Lipschitz Continuity + Strong Convexity II

- So Far:

$$f(x_t) - f(x^*) \le \frac{B^2\gamma_t}{2} + \left(\frac{1}{2\gamma_t} - \frac{\mu}{2}\right)||x_t - x^*|| - \frac{1}{2\gamma_t}||x_{t+1} - x^*||^2$$

$\Gamma_t = \frac{2}{\mu(1+t)}$

- For the specific case of $\boxed{\gamma_t^{-1} = \mu(1+t)/2}$ and multiplying both sides by t:

$$\sum_{t=0}^{T-1}\left[\underbrace{t[f(x_t) - f(x^*)]}_{=0 \text{ for } t=0} \le \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\left\{t(t-1)\,||x_t - x^*||^2 - (t+1)t\,||x_{t+1} - x^*||^2\right\}\right.$$

$$\left.\le \frac{B^2}{\mu} + \frac{\mu}{4}\{\underbrace{t(t-1)}_{=0 \text{ for } t=0 \,\&\, t=1}\,||x_t - x^*||^2 - (t+1)\underbrace{t}_{=0 \text{ for } t=0}\,||x_{t+1} - x^*||^2\}\right]$$

Lowerbnd based on $f(x_T) - f(x^*)$ $\le \sum_{t=1}^{T-1} t[f(x_t) - f(x^*)] \le \frac{TB^2}{\mu} - T(T-1)||x_T - x^*||^2 \cdot \frac{\mu}{4} \le \frac{TB^2}{\mu}$

Lipschitz Continuity + Strong Convexity II

- So Far:

$$f(x_t) - f(x^*) \le \frac{B^2\gamma_t}{2} + \left(\frac{1}{2\gamma_t} - \frac{\mu}{2}\right)||x_t - x^*|| - \frac{1}{2\gamma_t}||x_{t+1} - x^*||^2$$

- For the specific case of $\gamma_t^{-1} = \mu(1+t)/2$ and multiplying both sides by t:

$$t\left[f(x_t) - f(x^*)\right] \le \frac{B^2 t}{\mu(t+1)} + \frac{\mu}{4}\left\{t(t-1)\,||x_t - x^*||^2 - (t+1)t\,||x_{t+1} - x^*||^2\right\}$$

$$\le \frac{B^2}{\mu} + \frac{\mu}{4}\{t(t-1)\,||x_t - x^*||^2 - (t+1)t\,||x_{t+1} - x^*||^2\}$$

- Now we can use the telescoping sum and obtain...

$$\sum_{t=1}^{T} t(f(x_t) - f(x^*)) \le \frac{TB^2}{\mu} + \mu/4(\,0 - T(T+1)\,||x_{T+1} - x^*||^2) \le \frac{TB^2}{\mu}$$

Lipschitz + Strongly Convex III

- So Far:

$$\boxed{\sum_{t=1}^{T} t(f(x_t) - f(x^*))} \leq \boxed{\frac{TB^2}{\mu}}$$

$f(x^*) \leq f(x_T) \leq f(x_{T-1}) \leq \dots \leq f(x_1)$

$0 \leq f(x_T) - f(x^*) \leq \dots \leq f(x_1) - f(x^*)$

- So Far:

$$\boxed{\sum_{t=1} t\,(f(x_T) - f(x^*))} < \sum_{t=1}^{T} t(f(x_t) - f(x^*)) \leq \frac{TB^2}{\mu}$$

$$\geq \left(\sum_{t=1}^{T-1} t\right)(f(x_T) - f(x^*))$$

$$= \frac{T(T-1)}{2}(f(x_T) - f(x^*))$$

Lipschitz + Strongly Convex III

- So Far:

$$\boxed{\sum_{t=1}^{T} t(f(x_t) - f(x^*))} \leq \boxed{\frac{TB^2}{\mu}}$$

- Multiply by $2/(T(T+1))$ on both sides to make it a convex combination:

$$\sum_{t=1}^{T} \frac{2t}{T(T+1)}(f(x_t) - f(x^*)) \leq \frac{2B^2}{\mu(T+1)}$$

Lipschitz + Strongly Convex III

- So Far:

$$\boxed{\sum_{t=1}^{T} t(f(x_t) - f(x^*))} \leq \boxed{\frac{TB^2}{\mu}}$$

- Multiply by $2/(T(T+1))$ on both sides to make it a convex combination:

$$\sum_{t=1}^{T} \frac{2t}{T(T+1)}(f(x_t) - f(x^*)) \leq \frac{2B^2}{\mu(T+1)}$$

- Thus, for the error $\leq \epsilon$, we need $T \geq \frac{2B^2}{\mu\epsilon} - 1$

Summary of Results so Far (with convexity by default)...

- Lipschitz continuous Functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, attain an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Lipschitz continuous + Strongly Convex Functions (CS) with $\gamma_t = \mu(1+t)/2$ attain an an ϵ-approximate solution in $2B^2/\epsilon - 1$ iterations
- Smooth Functions (S): With $\gamma = 1/L$, attain an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
- Smooth + Strongly Convex Functions (SS): With $\gamma = 1/L$, attain an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then, we have the following:
  - $\epsilon = 0.1$, C: 10000, CS: 2000, S $= 50$, SS $= 8.49$ iterations
  - $\epsilon = 0.01$, C: 1000000, CS: 20000, S $= 500$, SS $= 13.49$ iterations
  - $\epsilon = 0.001$, C: 100000000, CS: 200000, S $= 5000$, SS $= 18.49$ iterations

Summary of Results so Far (with convexity by default)...

- Lipschitz continuous Functions (C). With $\gamma = \frac{R}{B\sqrt{T}}$, attain an ϵ-approximate solution in R^2B^2/ϵ^2 iterations
- Lipschitz continuous + Strongly Convex Functions (CS) with $\gamma_t = \mu(1+t)/2$ attain an an ϵ-approximate solution in $2B^2/\epsilon - 1$ iterations
- Smooth Functions (S): With $\gamma = 1/L$, attain an ϵ-approximate solution in $\frac{R^2L}{\epsilon}$ iterations.
  + convexity
- Smooth + Strongly Convex Functions (SS): With $\gamma = 1/L$, attain an ϵ-approximate solution in $\frac{L}{\mu}\log(\frac{R^2L}{2\epsilon})$ iterations.
- Concrete examples. Let $L = B = 10, R = 1, \mu = 1$. Then, we have the following:
  - $\epsilon = 0.1$, C: 10000, CS: 2000, S = 50, SS = 8.49 iterations
  - $\epsilon = 0.01$, C: 1000000, CS: 20000, S = 500, SS = 13.49 iterations
  - $\epsilon = 0.001$, C: 100000000, CS: 200000, S = 5000, SS = 18.49 iterations

Side note: Smoothness has yeilded overall less wall clock time (however the dependence is still $1/\epsilon$ } Like CS

Lower Bounds with convexity (No proof will be discussed in the class)

- Case I: Lipschitz Continuous Functions: Any black-box procedure will have an error of at least $\frac{RB}{2(1+\sqrt{T})}$ (GD: $\frac{RB}{\sqrt{T}}$)
- Case II: Lipschitz Continuous + Strongly Convex Functions: Any black-box procedure have an error of at least $\frac{B^2}{2\mu T}$ (GD: $\frac{2B^2}{\mu(T-1)}$)
- Case III: Smooth Functions: Any black box procedure will have an error of at least $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ (GD: $\frac{LR^2}{2T}$) - notice the gap between lower bound and actual time!
  - Can some improvement to the GD algorithm help bridge this? Ans: YES - Accelerated GD
- Case IV: Smooth + Strongly Convex Functions: Define $\kappa = \frac{L}{\mu}$. Then, any black box procedure will have an error of at least $\frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(T-1)}$

  (contrast with GD: $\frac{L}{2}\left(1-\frac{\mu}{L}\right)^T = \frac{L}{2}(\frac{\kappa-1}{\kappa})^T$)

For proofs, see Section 3.5 (page 53 onwards of) `https://arxiv.org/pdf/1405.4980.pdf`

Lower Bounds with convexity (No proof will be discussed in the class)

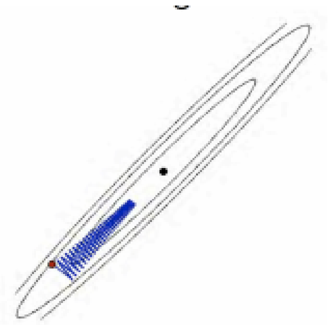
- Case I: Lipschitz Continuous Functions: Any black-box procedure will have an error of at least $\frac{RB}{2(1+\sqrt{T})}$ (GD: $\frac{RB}{\sqrt{T}}$) $T \geq \alpha\, 1/\epsilon^2$.
- Case II: Lipschitz Continuous + Strongly Convex Functions: Any black-box procedure have an error of at least $\frac{B^2}{2\mu T}$ (GD: $\frac{2B^2}{\mu(T-1)}$) $T \geq \alpha\, 1/\epsilon$
- Case III: Smooth Functions: Any black box procedure will have an error of at least $\frac{3L}{32}\frac{R^2}{(T+1)^2}$ (GD: $\frac{LR^2}{2T}$) - notice the gap between lower bound and actual time!
  - Can some improvement to the GD algorithm help bridge this? Ans: YES - Accelerated GD
- Case IV: Smooth + Strongly Convex Functions: Define $\kappa = \frac{L}{\mu}$. Then, any black box procedure will have an error of at least $\frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2(T-1)}$ $\kappa \geq 1$

  (contrast with GD: $\frac{L}{2}\left(1-\frac{\mu}{L}\right)^T = \frac{L}{2}(\frac{\kappa-1}{\kappa})^T$)

For proofs, see Section 3.5 (page 53 onwards of) `https://arxiv.org/pdf/1405.4980.pdf`

Accelerated GD? Why could GD be slow?

- GD has suboptimal rates for smooth (and to a smaller extent for smooth + strongly convex).
- GD relies just on local gradient information
- Can we add some momentum from the progress made so far to push it faster towards the optimal?



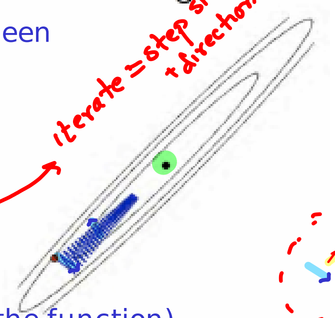
Accelerated GD? Why could GD be slow?

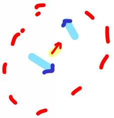
- GD has suboptimal rates for smooth (and to a smaller extent for smooth + strongly convex).
- GD relies just on local gradient information
- Can we add some momentum from the progress made so far to push it faster towards the optimal?

Motivation1: Bridging the gap
Concern: The gap could have been worse for some other function that Nesterov would have come up with?

Motivation2: Emperically could the next iterate have been influenced by the series of previous iterates (which also account for some curvature of the function)

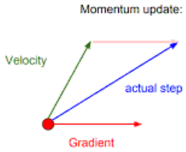
iterate = step size + direction





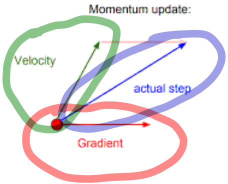
Avg movement

Efficient alternatives to GD: Momentum



- **Momentum Accelerated GD:** $x_{act} = x_{old} + \hat{\mathbf{v}}$ where $\hat{\mathbf{v}} = \mu \mathbf{v} - \eta_k \frac{\partial f}{\partial x}$

Efficient alternatives to GD: Momentum



- **Momentum Accelerated GD:** $x_{act} = x_{old} + \hat{\mathbf{v}}$ where $\hat{\mathbf{v}} = \mu \mathbf{v} - \eta_k \frac{\partial f}{\partial x}$

Accounting for momentum

$\propto$

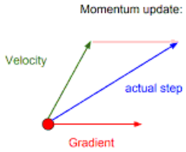
Multipicative factor associated with the history (captured through the velocity)

$\cancel{\mu > 1}$?

$\cancel{\mu < 0}$?

$\mu \in (0,1)$ ✓

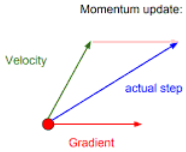
Efficient alternatives to GD: Momentum



- **Momentum Accelerated GD:** $x_{act} = x_{old} + \hat{\mathbf{v}}$ where $\hat{\mathbf{v}} = \mu\mathbf{v} - \eta_k \frac{\partial f}{\partial x}$
- $\mathbf{v}$ plays the role of velocity - it is the direction and speed at which the parameter x_i moves through parameter space.
- Set to an exponentially decaying average of the negative gradient: Decay factor will be

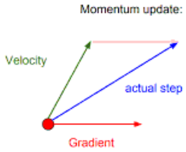
$\mu \in [0,1)$

Efficient alternatives to GD: Momentum



- **Momentum Accelerated GD:** $x_{act} = x_{old} + \hat{\mathbf{v}}$ where $\hat{\mathbf{v}} = \mu \mathbf{v} - \eta_k \frac{\partial f}{\partial x}$
  - $\mathbf{v}$ plays the role of velocity - it is the direction and speed at which the parameter x_i moves through parameter space.
  - Set to an exponentially decaying average of the negative gradient: Decay factor will be $\mu \in [0, 1)$. Common values of μ used in practice are .5, .9, and .99.
- Physical analogy $\Rightarrow$ negative gradient is a force moving a particle through parameter space, according to Newton's laws of motion
- Step size largest when

Efficient alternatives to GD: Momentum

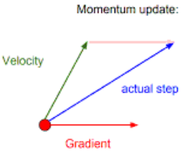


- **Momentum Accelerated GD:** $x_{act} = x_{old} + \hat{\mathbf{v}}$ where $\hat{\mathbf{v}} = \mu \mathbf{v} - \eta_k \frac{\partial f}{\partial x}$

- $\mathbf{v}$ plays the role of velocity - it is the direction and speed at which the parameter x_i moves through parameter space.
- Set to an exponentially decaying average of the negative gradient: Decay factor will be $\mu \in [0, 1)$. Common values of μ used in practice are .5, .9, and .99

- Physical analogy $\Rightarrow$ negative gradient is a force moving a particle through parameter space, according to Newton's laws of motion
- Step size largest when

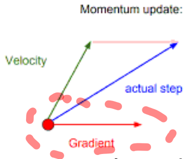
Suggestion: When the gradient and velocity are in the same direction
More generally: When successive gradients are in the same direction



- **Momentum Accelerated GD:** $x_{act} = x_{old} + \hat{\mathbf{v}}$ where $\hat{\mathbf{v}} = \mu \mathbf{v} - \eta_k \frac{\partial f}{\partial x}$
- $\mathbf{v}$ plays the role of velocity - it is the direction and speed at which the parameter x_i moves through parameter space.
- Set to an exponentially decaying average of the negative gradient: Decay factor will be $\mu \in [0, 1)$. Common values of μ used in practice are .5, .9, and .99.

- Physical analogy $\Rightarrow$ negative gradient is a force moving a particle through parameter space, according to Newton's laws of motion
- Step size largest when many successive gradients point in exactly the same direction

Efficient alternatives to GD: Momentum



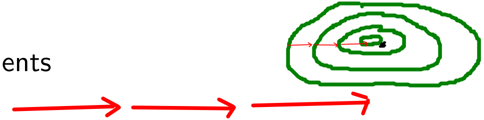
option 2: $\frac{\partial f}{\partial x_{new}}$

option 1: $\frac{\partial f}{\partial x_{old}}$

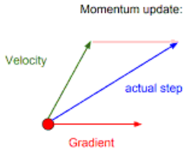
- **Momentum Accelerated GD:** $x_{act} = x_{old} + \hat{\mathbf{v}}$ where $\hat{\mathbf{v}} = \mu \mathbf{v} - \eta_k \frac{\partial f}{\partial x}$

  Q: x is x_{old} OR x_{new}

  - $\mathbf{v}$ plays the role of velocity - it is the direction and speed at which the parameter x_i moves through parameter space.
  - Set to an exponentially decaying average of the negative gradient: Decay factor will be $\mu \in [0, 1)$. Common values of μ used in practice are .5, .9, and .99. fading history
- Physical analogy $\Rightarrow$ negative gradient is a force moving a particle through parameter space, according to Newton's laws of motion
- Step size largest when many successive gradients point in exactly the same direction



Efficient alternatives to GD: Momentum

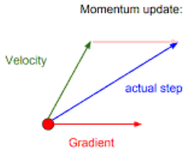


- **Momentum Accelerated GD:** $x_{act} = x_{old} + \hat{\mathbf{v}}$ where $\hat{\mathbf{v}} = \mu \mathbf{v} - \eta_k \frac{\partial f}{\partial x}$

- $\mathbf{v}$ plays the role of velocity - it is the direction and speed at which the parameter x_i moves through parameter space.
- Set to an exponentially decaying average of the negative gradient: Decay factor will be $\mu \in [0, 1)$. Common values of μ used in practice are .5, .9, and .99.

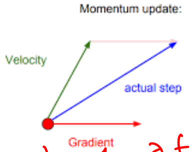
- Physical analogy $\Rightarrow$ negative gradient is a force moving a particle through parameter space, according to Newton's laws of motion
- Step size largest when many successive gradients point in exactly the same direction

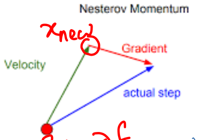
Efficient alternatives to GD: Momentum



- **Momentum Accelerated GD:** $x_{act} = x_{old} + \hat{\mathbf{v}}$ where $\hat{\mathbf{v}} = \mu \mathbf{v} - \eta_k \frac{\partial f}{\partial x}$
- $\mathbf{v}$ plays the role of velocity - it is the direction and speed at which the parameter x_i moves through parameter space.
- Set to an exponentially decaying average of the negative gradient: Decay factor will be $\mu \in [0, 1)$. Common values of μ used in practice are .5, .9, and .99.
- Physical analogy $\Rightarrow$ negative gradient is a force moving a particle through parameter space, according to Newton's laws of motion
- Step size largest when many successive gradients point in exactly the same direction

Heavy-ball/Polyak



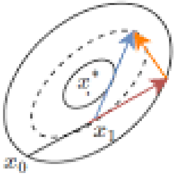


option 1: $\frac{\partial f}{\partial x_{old}}$ option 2: $\frac{\partial f}{\partial x_{new}}$

Momentum and its variants

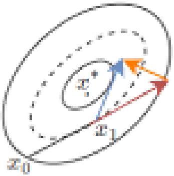
- Modified Nesterov momentum $\Rightarrow$ gradient is evaluated after the current velocity is applied. Speeds up rate of convergence from $O(1/T)$ to $O(1/T^2)$

Polyak's Momentum



$$x_{t+1} = x_t - \alpha\nabla f(x_t) + \mu(x_t - x_{t-1})$$

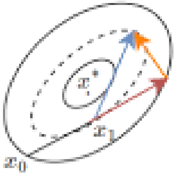
Nesterov Momentum



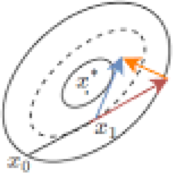
$$x_{t+1} = x_t + \mu(x_t - x_{t-1}) \\ - \gamma\nabla f(x_t + \mu(x_t - x_{t-1}))$$

- Modified Nesterov momentum $\Rightarrow$ gradient is evaluated after the current velocity is applied. Speeds up rate of convergence from $O(1/T)$ to $O(1/T^2)$ for L-continuity and strong convexity case

Polyak's Momentum



Nesterov Momentum



$$x_{t+1} = x_t \underbrace{- \alpha\nabla f(x_t)}_{} + \mu(x_t - x_{t-1})$$

$$x_{t+1} = x_t + \mu(x_t - x_{t-1}) \underbrace{- \gamma\nabla f(x_t + \mu(x_t - x_{t-1}))}_{}$$

$-\alpha \nabla f(x_t) \leftarrow$ Difference $\rightarrow -r\nabla f(x_t + u(x_t - x_{t-1})) = -r\nabla f(x_{new})$

Attempt 1: Polyak's Heavy Ball Momentum

- Recall standard gradient descent: $x_{t+1} = x_t - \alpha_t \nabla f(x_t)$
- Polyak's Heavy Ball Momentum adds inertia: $x_{t+1} = x_t - \alpha_t \nabla f(x_t) + \mu_t(x_t - x_{t-1})$
- Heavy Ball result: For smooth + strongly convex functions, the heavy ball algorithm converges in $\frac{R^2}{2}\left(1 - \sqrt{\frac{1}{\kappa}}\right)^T = \frac{L}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}}\right)^T$ in comparison with GD's $\frac{R^2}{2}\left(1 - \frac{1}{\kappa}\right)^T = \frac{L}{2}\left(\frac{\kappa-1}{\kappa}\right)^T$ iterations.
- Heavy Ball momentum is not optimal for the Smooth case (though it is optimal for the strongly convex + smooth class)

Attempt 1: Polyak's Heavy Ball Momentum

Conservatively exploits curvature

In this term which is standard GD
the curvature is not being exploited

Velocity exploits
curvature empirically

Exponential Decay μ_t

- Recall standard gradient descent: $x_{t+1} = x_t - \alpha_t \nabla f(x_t)$
- Polyak's Heavy Ball Momentum adds inertia: $x_{t+1} = x_t - \alpha_t \nabla f(x_t) + \mu_t(x_t - x_{t-1})$
- Heavy Ball result: For smooth + strongly convex functions, the heavy ball algorithm converges in $\frac{R^2}{2}\left(1-\sqrt{\frac{1}{\kappa}}\right)^T = \frac{L}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}}\right)^T$ in comparison with GD's $\frac{R^2}{2}\left(1-\frac{1}{\kappa}\right)^T = \frac{L}{2}\left(\frac{\kappa-1}{\kappa}\right)^T$ iterations.

Comparable and Commensurate with Lower Bo

- Heavy Ball momentum is not optimal for the Smooth case (though it is optimal for the strongly convex + smooth class)

+ vanilla convexity

We have both an upper bound
and a lower bound on the curvature
We might have to be a bit careful in going aggressively in exploiting curvature