IE510 Applied Nonlinear Programming

Lecture 5: Fast Gradient Methods by Momentum

Ruoyu Sun

Feb 18, 2020

Review for Last Week (Analysis of GD)

Q1: If eigenvalues are $(3.8, 0.1, 0.05, 0.002)$, what plot do you expect with stepsize $1/3.8$?

Q2: What plots are possible for convex problems?

Today

- Today: GD with momentum; i.e., heavy ball method
- After today's course, you will be able to
  - explain why momentum is useful
  - pick the parameters of GD with momentum

Outline

Next Few Weeks: Faster Algorithm

Lesson: For huge problems, “slow” $\approx$ “not converging”

Convergence speed is as important as convergence itself, if not more.

Three classes of faster algorithms (besides preconditioning)

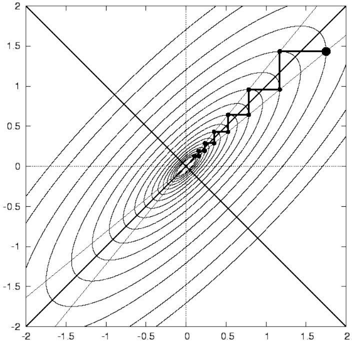
- Using momentum
- Using spectral info
- Decomposition (into small problems)

20+ different algorithms/variants. We’ll cover some of them.

Sin of Greedy

Recall GD: a ____________ method

"*Those who cannot remember the past are condemned to repeat it.*"
–Santayana



For GD, can we incorporate history to improve?

Add History Info to GD

Gradient descent is going too fast on the new direction.

$$\text{GD}: x^{k+1} = x^k - \alpha \nabla f(x^k).$$

Use the information of the old direction.

GD with momentum (**heavy ball** method): [Polyak'1974]

$$\text{Heavy ball}: x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta\_\_\_\_\_\_\_\_\_\_$$

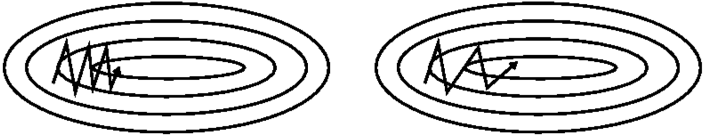
Fixed point verification:

Remark (don't forget history): Lots of similar methods before Polyak'1974 were proposed.

Constant α, β for linear algebra were studied by Frankel'1950 (call it 2nd order Richardson method).

System Designer's Motivation: History Helps

Explanation why momentum works: reduce "zigzag" behavior.



Nature's Motivation: Momentum in Driving

Recall: GD is doing hill clibming (in a greedy way).

What if you are driving uphill?

Difference: You change ________, insteading of changing position directly.

Polyak: heavy ball moving according to neg-gradient; but due to“inertia”, it has momentum.

Changing Speed by Gradient

$$v^k = \_\_\_ v^{k-1} - \_\_\_ \nabla f(x).$$
$$x^{k+1} = x^k - \_\_\_ v^k,$$

Graph: Draw $x^k, x^{k-1}, x^{k+1}, v^k, v^{k+1}$

This is equivalent to (set $v^0 = 0$)

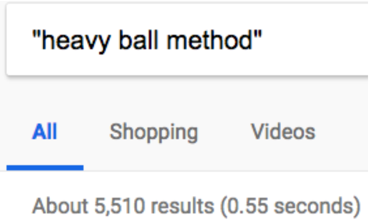
$$\text{Heavy ball method : } x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}). \qquad (1)$$

Task: Find the three coefficients in the red equations, so that the two red equations are equivalent to (1).

Momentum is Popular

Heavy ball method is not commonly covered in traditional optimization courses.

but popularized by ML community recently, under the name “GD with momentum”



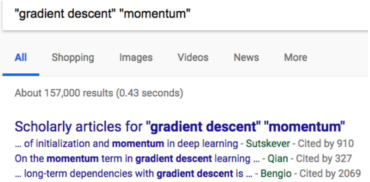
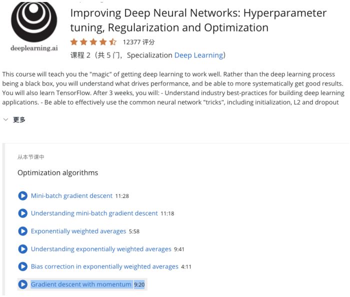


Figure: Left: Heavy ball; Right: Momentum

Momentum is Popular



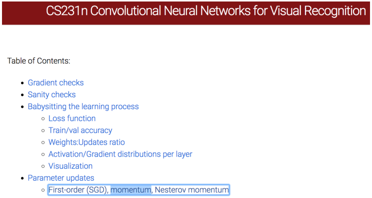


Figure: Snapshots of Online Resources on Momentum

Momentum is Popular (cont'd)

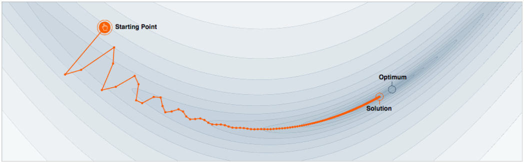
Conclusion

In this post we looked at the optimization algorithms for neural nets beyond SGD. We looked at two classes of algorithms: momentum based and adaptive learning rate methods.

We also implement all of those methods in Python and Numpy with the use case of our neural nets stated in the last post.

Most of those methods above are currently implemented in the popular Deep Learning

Why Momentum Really Works



Step-size α = 0.02 — 0, 0.003, 0.006

Momentum β = 0.99 — 0.00, 0.500, 0.990

We often think of Momentum as a means of dampening oscillations and speeding up the iterations, leading to faster convergence. But it has other interesting behavior. It allows a larger range of step-sizes to be used, and creates its own oscillations. What is going on?

GABRIEL GOH
UC Davis | April. 4 2017 | Citation: Goh, 2017

When They Talk about Momentum Method, What do They Talk About?

They talk about: why use the momentum; why it works better.

As "Why Momentum Really Works" implied, there is something else that is missing in most ML/engineering courses/posts.

- Partially due to the time/space limit in ML/engineering courses/posts

I'll mainly talk about how optimizers understand momentum method.

Outline

Result for Quadratic Problem

The analysis for 2-dim can be extended to general n-dim strongly convex quadratic problem.

Proposition 3.1: Suppose Q is symmetric PD. Consider solving

$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}) \triangleq \mathbf{x}^T Q\mathbf{x} + 2\mathbf{b}^T x + \mathbf{c}.$$

Use heavy ball method with

$$\alpha = \frac{1}{L}, \quad \beta = (1 - \sqrt{1/\kappa})^2,$$

where $L = \lambda_{\max}(Q)$, $\kappa =$ is the condition number.

Then the asymptotic convergence rate

$$\exp(\lim_{r\to\infty} \frac{1}{r} \log(\|x^r - x^*\|/\|x^0 - x^*\|)) = 1 - \sqrt{1/\kappa}.$$

Remark: To achieve relative error $\frac{\|\mathbf{x}^r - x^*\|}{\|\mathbf{x}^0 - x^*\|} < \epsilon$, only need $\#$ of iterations

$$\sqrt{\kappa}\log(\frac{1}{\epsilon}).$$

Theoretical Implication

$$x^+ = x - \frac{1}{L}\nabla f(x) + (1 - \sqrt{1/\kappa})^2 (x - x^-).$$

Original GD: $\tilde{O}(\kappa)$ iterations. (ignore $\log 1/\epsilon$ term).

Adding a momentum term reduces to $\tilde{O}(\sqrt{\kappa})$ iterations.

Accelerate by $\sqrt{\kappa}$ times.

- Example: $\kappa = 10^4$, accelerate by 100 times.

This is one of major reasons why people believe momentum can help!

Exercise

True or False? For strongly convex problems with strongly convexity parameter κ, gradient descent method with momentum (when picking stepsize properly) has an asymptotic convergence rate of $1 - 1/\sqrt{\kappa}$, thus is faster than gradient descent method with only convergence rate of $1 - 1/\kappa$.

Some Drawbacks

Drawback 1: For non-quadratic function, such acceleration is lost – at least no one knows how to pick α, β to achieve it.

Drawback 2: Nonconvex problems? Unknown.

Drawback 3: Two parameters α, β to tune in practice. Harder to tune.

Optimal Stepsize

- For strongly convex functions, $1/L$ is not the optimal stepsize.
- Use GD to solve strongly convex problem, the optimal stepsize is

$$\alpha = \frac{1}{L+\mu}.$$

Convergence rate is $\frac{\kappa-1}{\kappa+1}$.
$\#$ of iterations is at most $\frac{\kappa+1}{2}\log\frac{1}{\epsilon}$.

- Use GD with momentum (heavy ball method) to solve strongly convex quadratic problem, the optimal stepsize is

$$\alpha = \left(\frac{2}{\sqrt{\lambda_{\max}}+\sqrt{\lambda_{\min}}}\right)^2, \quad \beta = \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^2.$$

Convergence rate is $\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}$.
$\#$ of iterations is at most $\frac{\sqrt{\kappa}+1}{2}\log\frac{1}{\epsilon}$.

Practical Stepsize Choice

These are theoretical stepsizes; any practical guidance to engineers?

- GD: Theoretical stepsize $2/(L+\mu)$; diverge for $> 2/L$, so...
- **GD practice**: "knife's edge". Slightly smaller than the converge/diverge threshold
- Heavy ball (HB): theoretical α slightly smaller than $4/L$, $\beta \approx (1 - \frac{2}{\sqrt{\kappa}})^2$
- **HB Practice**: pick β slightly smaller than 1, and tune α up before divergence ("knife's edge")

Remark: These guidelines work for convex case; for nonconvex, may or may not work, but can be a starting point to try.

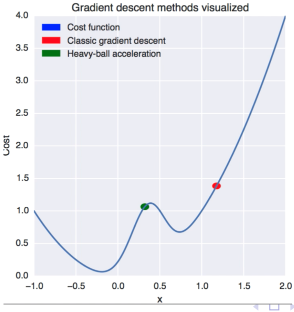
Does Momentum Work Better for Nonconvex Problems?

A popular claim: momentum helps avoid local-min.

Polyak'74 argued: for GD $x^{k+1} = x^k - \alpha \nabla f(x^k)$, when $\nabla f(x^k)$ is small, the ball will stay in the "valley".

But for HB, the extra $x^k - x^{k-1}$ term may pull the ball out of the basin.

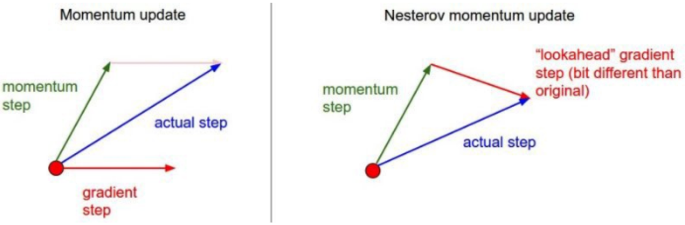
Can be a research topic or course project.



Outline

Nesterov's accelerated gradient method (Nesterov momentum)

View: slip (by momentum) then update. Difference lies in: HB uses old gradient, Nesterov's uses new gradient



Source (Stanford CS231n class)

HB $\begin{cases} x_{\text{ahead}} = x + \beta(x - x_{\text{old}}), \\ x_{\text{new}} = x_{\text{ahead}} - \alpha \nabla f(x). \end{cases}$ Nesterov $\begin{cases} x_{\text{ahead}} = x + \beta(x - x_{\text{old}}), \\ x_{\text{new}} = x_{\text{ahead}} - \alpha \nabla f(x_{\text{ahead}}). \end{cases}$

Nesterov's Momentum

Now we define a simple version of Nesterov's accelerated gradient method (1983).

$$\begin{cases} y^r = x^r + \beta_r(x^r - x^{r-1}), & \text{slip due to momentum} \\ x^{r+1} = y^r - \alpha \nabla f(y^r). & \text{move along new gradient} \end{cases} \tag{2}$$

- Simplest stepsize (for strongly convex case):

$$\alpha = 1/L, \quad \beta_r = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}. \quad \text{constant parameters} \tag{3}$$

- Simplest stepsize (for convex case):

$$\alpha = 1/L, \quad \beta_r = \frac{r-1}{r+3}. \quad \text{time-dependent parameters} \tag{4}$$

Nesterov's Momentum: Results

Theorem 5.1 For strongly convex problems, Nesterov's method (5) with the stepsize choice in (3) satisfies

$$f(x^r) - f^* \leq L(1 - \frac{1}{\sqrt{\kappa}})^{2r} \|x^0 - x^*\|^2.$$

For convex problems, Nesterov's method (5) with the stepsize choice in (4) satisfies

$$f(x^r) - f^* \leq \frac{2L}{(r+1)^2} \|x^0 - x^*\|^2.$$

- For strongly convex case, iteration complexity $O(\sqrt{\kappa} \log 1/\epsilon)$, faster than $O(\kappa \log 1/\epsilon)$ of GD.
- For convex case, iteration complexity $O(\qquad \epsilon \;)$, faster than ____________ of GD.

See Nesterov "Introductory lectures on convex optimization" for details. A shorter introduction in Donoghue,Candes "Adaptive Restart for Accelerated Gradient Schemes".

Lots of Interpretations

People found Nesterov's original proof HARD to understand.

Lots of interpretations:

- Chebychev polynomial (related); approximation theory
  - Hardt blog: "Zen of Gradient Descent", but does not recover Nesterov's method
  - HB method equivalent to Chebychev iteration method
- ODE interpretation (2nd order ODE, Hamiltonian system; not simple)
- geometric idea (related to ellipsoid method; different method)
- game in primal-dual method
- upper/lower bound estimate (still magical)
- ......

General Stepsize Rule (optional)

The original stepsize rule by Nesterov is rather general:

$$\begin{cases} y^r = x^r + \beta_r(x^r - x^{r-1}), \\ x^{r+1} = y^r - \alpha_r \nabla f(y^r). \end{cases} \tag{5}$$

$\alpha_r \leq 1/L$, and one general choice of β_r is:

General Rule 1: $q \in [0,1]$, $\theta_{r+1}^2 = (1-\theta_{r+1})\theta_r^2 + q$.

$$\beta_{r+1} = \frac{\theta_r(1-\theta_r)}{\theta_r^2 + \theta_{r+1}}.$$

Consider $\theta_0 = 1$ in this page.
Let $q = 1$: then $\theta_r = 1$, $\beta_r = 0$. Recover GD.

When convex, let $q = 0$: then $\theta_{r+1} = \frac{\theta_r(\sqrt{\theta_r^2+4}-\theta_r)}{2}$, then

$$f(x^r) - f^* \leq \frac{4L}{(r+2)^2}\|x^0 - x^*\|^2.$$

When strongly convex, let $q = 1/\kappa = \mu/L$, linear convergence:

$$f(x^r) - f^* \leq L\left(1 - \frac{1}{\sqrt{\kappa}}\right)^r \|x^0 - x^*\|^2.$$

Derive Two Simplest Stepsize Rules (optional)

Let the initial point of the auxiliary sequence and the parameter be

$$\theta_0 = 1/\sqrt{\kappa}, q = 1/\kappa.$$

Then we obtain

$$\beta_r = \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1},$$

which recovers (3).

Another general stepsize rule different from Rule 2 (for convex case):

General Rule 2: $a_0 \in [0,1],\ a_{r+1} = (1+\sqrt{4a_{r-1}^2+1})/2.$

$$\beta_r = \frac{a_r - 1}{a_{r+1}}.$$

Let $a_0 = 0$, then $a_r = \frac{r+1}{2}$, and

$$\beta_r = 1 - \frac{4}{r+3},$$

which recovers (4).

Further Improvement of Momentum? (optional)

- : Question: can we do better than momentum methods?
  - "Better" can mean many things...
  - What Nesterov's method does: extend $\sqrt{\kappa}$ result from quadratic to convex/strongly convex
- **Question**: Can we improve the bound $\tilde{O}(\sqrt{\kappa})$, even just for quadratic case?
- More history info: can we use three or four terms in history, to get $\tilde{O}(\kappa^{1/3})$ or even better bound?
- Answer: No (within a certain theoertical framework). Nesterov's method is "optimal" for solving convex problems.
- For engineers: can we get a "faster" gradient-type method than Nesterov's method?

Appendix

(Starting next page. Not required for the course)

Review of GD Analysis

Starting from the simplest case:

$$\min_{x_1,x_2\in\mathbb{R}} \frac{1}{2}(\lambda_1 x_1^2 + \lambda_2 x_2^2).$$

$$\text{GD}: \quad x^+ =$$

$$\begin{bmatrix} x_1^+ \\ x_2^+ \end{bmatrix} = \begin{bmatrix} (1-\alpha\lambda_1)x_1 \\ (1-\alpha\lambda_2)x_1 \end{bmatrix}$$

First eigen-mode rate: $|1-\alpha\lambda_1|$.

Second eigen-mode rate: $|1-\alpha\lambda_2|$.

Pick $\alpha = \frac{1}{\lambda_1}$, then the speed $\max\{0, |1-\frac{\lambda_2}{\lambda_1}|\} = 1-\frac{1}{\kappa}$.

Analysis: Starting From 2-dim

Starting from the simplest case:

$$\min_{x_1, x_2 \in \mathbb{R}} \frac{1}{2}(\lambda_1 x_1^2 + \lambda_2 x_2^2).$$

Heavy ball method:

$$x^+ =$$

$$\begin{bmatrix} x_1^+ \\ x_2^+ \end{bmatrix} =$$

Analysis of 2-dim (cont'd)

Observation: The two sequences x_1^k and x_2^k are independent.

thus the convergence speed of x^k depends on _______________

Each sequence is like

$$a_{k+2} = \gamma a_k - \beta a_k, \tag{6}$$

where $\gamma = 1 - \alpha\lambda_i + \beta$.

e.g. Fibonacci sequence $a_{k+2} = a_{k+1} + a_k$,
$1, 1, 2, 3, 5, 8, 13, 21, 34, \ldots$.

General formula:

$$a_k = c_1\sigma_1^k + c_2\sigma_2^k,$$

where σ_1, σ_2 are the two roots of

$$z^2 + \qquad\qquad\qquad = 0.$$

Convergence of 2-term Recursion

$$a_k = c_1 \sigma_1^k + c_2 \sigma_2^k,$$

where $\sigma_{1,2} = \frac{\gamma \pm \sqrt{\gamma^2 - 4\beta}}{2}$ and $\gamma = 1 - \alpha\lambda_i + \beta$.

Case 1: $\gamma^2 - 4\beta > 0$. Two Real roots. Can ignore it.

Case 2: $\gamma^2 - 4\beta \leq 0$. Two conjugate complex roots.

Practice: in Case 2, for what α, β the sequence converges?

Conclusion: When

$$\leq \sqrt{\beta} < \qquad ,$$

the sequence converges at rate ____.

The optimal speed is achieved when $\beta =$.

2-dim Case

Now let's come back to the 2-dim case.

Two sequences x_1^k, x_2^k:

$$\{x_1^k\} \text{ converges if } |1-\sqrt{\alpha\lambda_1}| \le \sqrt{\beta} < 1, \tag{7}$$

$$\{x_2^k\} \text{ converges if } |1-\sqrt{\alpha\lambda_2}| \le \sqrt{\beta} < 1, \tag{8}$$

both with rate $\sqrt{\beta}$.

The overall convergence rate is optimal when

$$\sqrt{\beta} = \max\{|1-\sqrt{\alpha\lambda_1}|,\ \ |1-\sqrt{\alpha\lambda_2}|\}.$$

Pick $\alpha = \frac{1}{\lambda_1} = 1/L$, then the optimal $\beta =$, and the rate is

$$1 -$$

Appendix: Convergence of 2-term Recursion (skip in class)

$$a_k = c_1\sigma_1^k + c_2\sigma_2^k,$$

where $\sigma_{1,2} = \frac{\gamma \pm \sqrt{\gamma^2 - 4\beta}}{2}$ and $\gamma = 1 - \alpha\lambda_i + \beta$.

First question: when does the sequence converge?
Answer:

$$\max\{|\sigma_1|, |\sigma_2|\} < 1.$$

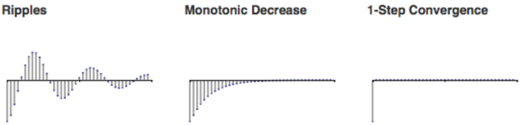
A sufficient condition for convergence is

$$0 \le \beta < 1, \quad 0 < \alpha\lambda_i < 2 + 2\beta.$$

Remark: When β is close to 1, stepsize $\alpha < 4/\lambda_i$; 2 times larger than GD upper bound $2/\lambda_i$.

Appendix: Fastest Convergence Rate (skip in class)

Second question: when does the sequence converge the fastest?



- If α and β are both free, the best choice is

$$\beta = 0, \quad \alpha = 1/\lambda_i.$$

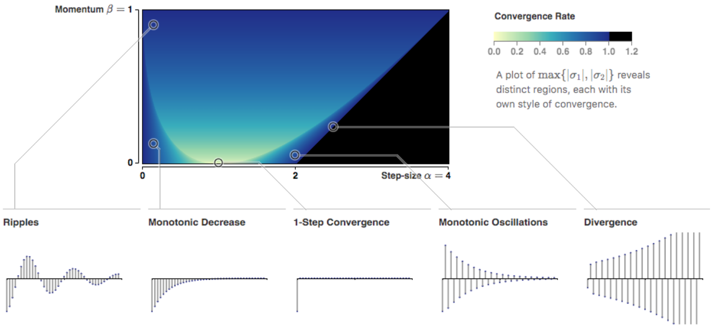
  1 step converge.

- But α is not completely free (why?)

- Best $\beta = (1 - \sqrt{\alpha\lambda_i})^2$, and converge at rate $\sqrt{\beta}$.
  - This happens when $\sigma_1 = \sigma_2$, i.e., $\gamma^2 - 4\beta = (1 - \alpha\lambda_i + \beta)^2 - 4\beta = 0$.

Appendix: Convergence Rate v.s. Parameter Choice (skip in class)

Figures of convergence rate v.s. α, β. Here α replaces $\alpha\lambda_i$.



Two boundaries: $\alpha = 2\beta + 2$; $\quad \beta = (1 - \sqrt{\alpha})^2$.

Conclusion of Today

Can you summarize yourself?

- Heavy ball method adds a momentum term to GD
- It works better because faster by a factor of $\sqrt{\kappa}$, for (strongly convex) quadratic problem
- Practical choice of stepsize: β slightly smaller than 1, and α as large as possible

Conclusion of Today

Can you summarize yourself?

- Heavy ball method adds a momentum term to GD
- It works better because faster by a factor of $\sqrt{\kappa}$, for (strongly convex) quadratic problem
- Practical choice of stepsize: β slightly smaller than 1, and α as large as possible

Questions for Last Week

- **Q1**: Physical interpretation of momentum method?

- **Q2**: Your colleague implemented GD and found it takes one day to run the algorithm. He/she already tuned the stepsize for a while, but the improvement is little.

  You suggest using momentum. He/she asked why. What will you say?

Last Week: Convergence Rate of Momentum

- **Proposition 3.1**: Suppose f is a strongly convex quadratic function, and consider
$$\min_{\mathbf{x}\in\mathbb{R}^n} f(\mathbf{x}).$$
Solving the problem by GD with momentum (heavy ball method) with stepsize α and momentum coefficient β.
  - The iterates converges at a rate $1 - 1/\sqrt{\kappa}$ if
$$\alpha = \frac{1}{\lambda_{\max}}, \beta = (1 - \sqrt{\alpha\lambda_{\min}})^2,$$
  - The iterates converges at a rate $1 - 2/(\sqrt{\kappa} + 1)$ if
$$\alpha = \left(\frac{2}{\sqrt{\lambda_{\max}} + \sqrt{\lambda_{\min}}}\right)^2, \beta = (1 - \sqrt{\alpha\lambda_{\min}})^2,$$
- **Practical Choice**: Pick β slightly smaller than 1, and α as large as possible (still converging).