

I'm a bandit

Random topics on optimization, probability, and statistics. By Sébastien Bubeck



- [Home](#)
- [ORF523: The complexities of optimization](#)
- [Guest posts](#)
- [Archives](#)
- [About me](#)

# ORF523: Nesterov's Accelerated Gradient Descent

Posted on [April 1, 2013](#) by [Sebastien Bubeck](#)

In this lecture we consider the same setting than in the previous post (that is we want to minimize a smooth convex function over  $\mathbb{R}^n$ ). Previously we saw that the plain Gradient Descent algorithm has a rate of convergence of order  $1/t$  after  $t$  steps, while the lower bound that we proved is of order  $1/t^2$ .

We present now a beautiful algorithm due to Nesterov, called Nesterov's Accelerated Gradient Descent, which attains a rate of order  $1/t^2$ . First we define the following sequences:

$$\lambda_0 = 0, \quad \lambda_s = \frac{1 + \sqrt{1 + 4\lambda_{s-1}^2}}{2}, \quad \text{and} \quad \gamma_s = \frac{1 - \lambda_s}{\lambda_{s+1}}.$$

(Note that  $\gamma_s \leq 0$ .) Now the algorithm is simply defined by the following equations, with an arbitrary initial point  $x_1 = y_1$ ,

$$\begin{aligned} y_{s+1} &= x_s - \frac{1}{\beta} \nabla f(x_s), \\ x_{s+1} &= (1 - \gamma_s) y_{s+1} + \gamma_s y_s. \end{aligned}$$

In other words, Nesterov's Accelerated Gradient Descent performs a simple step of gradient descent to go from  $x_s$  to  $y_{s+1}$ , and then it 'slides' a little bit further than  $y_{s+1}$  in the direction given by the previous point  $y_s$ .

The intuition behind the algorithm is quite difficult to grasp, and unfortunately the analysis will not be very enlightening either. Nonetheless Nesterov's Accelerated Gradient is an optimal method (in terms of oracle

complexity) for smooth convex optimization, as shown by the following theorem. **[Added in September 2015: we now have a simple geometric explanation of the phenomenon of acceleration, [see this post](#).]**

**Theorem (Nesterov 1983)** Let  $f$  be a convex and  $\beta$ -smooth function, then Nesterov's Accelerated Gradient Descent satisfies

$$f(y_t) - f(x^*) \leq \frac{2\beta\|x_1 - x^*\|^2}{t^2}.$$

We follow here the proof by Beck and Teboulle from the paper '[A fast iterative shrinkage-thresholding algorithm for linear inverse problems](#)'.

*Proof:* We start with the following observation, that makes use of Lemma 1 and Lemma 2 from the previous lecture: let  $x, y \in \mathbb{R}^n$ , then

$$\begin{aligned} & f\left(x - \frac{1}{\beta}\nabla f(x)\right) - f(y) \\ & \leq f\left(x - \frac{1}{\beta}\nabla f(x)\right) - f(x) + \nabla f(x)^\top(x - y) \\ & \leq \nabla f(x)^\top\left(x - \frac{1}{\beta}\nabla f(x) - x\right) + \frac{\beta}{2}\left\|x - \frac{1}{\beta}\nabla f(x) - x\right\|^2 + \nabla f(x)^\top(x - y) \\ & = -\frac{1}{2\beta}\|\nabla f(x)\|^2 + \nabla f(x)^\top(x - y). \end{aligned}$$

Now let us apply this inequality to  $x = x_s$  and  $y = y_s$ , which gives

$$\begin{aligned} f(y_{s+1}) - f(y_s) &= f\left(x_s - \frac{1}{\beta}\nabla f(x_s)\right) - f(y_s) \\ &\leq -\frac{1}{2\beta}\|\nabla f(x_s)\|^2 + \nabla f(x_s)^\top(x_s - y_s) \\ &= -\frac{\beta}{2}\|y_{s+1} - x_s\|^2 - \beta(y_{s+1} - x_s)^\top(x_s - y_s). \end{aligned} \tag{1}$$

Similarly we apply it to  $x = x_s$  and  $y = x^*$  which gives

$$f(y_{s+1}) - f(x^*) \leq -\frac{\beta}{2}\|y_{s+1} - x_s\|^2 - \beta(y_{s+1} - x_s)^\top(x_s - x^*). \tag{2}$$

Now multiplying (1) by  $(\lambda_s - 1)$  and adding the result to (2), one obtains with  $\delta_s = f(y_s) - f(x^*)$ ,

$$\lambda_s \delta_{s+1} - (\lambda_s - 1) \delta_s \leq -\frac{\beta}{2} \lambda_s \|y_{s+1} - x_s\|^2 - \beta(y_{s+1} - x_s)^\top(\lambda_s x_s - (\lambda_s - 1)y_s - x^*).$$

Multiplying this inequality by  $\lambda_s$  and using that by definition  $\lambda_{s-1}^2 = \lambda_s^2 - \lambda_s$  one obtains

$$\begin{aligned} & \lambda_s^2 \delta_{s+1} - \lambda_{s-1}^2 \delta_s \\ & \leq -\frac{\beta}{2} \left( \lambda_s \|y_{s+1} - x_s\|^2 + 2\lambda_s(y_{s+1} - x_s)^\top(\lambda_s x_s - (\lambda_s - 1)y_s - x^*) \right). \end{aligned} \tag{3}$$

Now one can verify that

$$\begin{aligned} & \|\lambda_s(y_{s+1} - x_s)\|^2 + 2\lambda_s(y_{s+1} - x_s)^\top (\lambda_s x_s - (\lambda_s - 1)y_s - x^*) \\ &= \|\lambda_s y_{s+1} - (\lambda_s - 1)y_s - x^*\|^2 - \|\lambda_s x_s - (\lambda_s - 1)y_s - x^*\|^2. \end{aligned} \quad (4)$$

Next remark that, by definition, one has

$$\begin{aligned} x_{s+1} &= y_{s+1} + \gamma_s(y_s - y_{s+1}) \\ \Leftrightarrow \lambda_{s+1}x_{s+1} &= \lambda_{s+1}y_{s+1} + (1 - \lambda_s)(y_s - y_{s+1}) \\ \Leftrightarrow \lambda_{s+1}x_{s+1} - (\lambda_{s+1} - 1)y_{s+1} &= \lambda_s y_{s+1} - (\lambda_s - 1)y_s. \end{aligned} \quad (5)$$

Putting together (3), (4) and (5) one gets with  $u_s = \lambda_s x_s - (\lambda_s - 1)y_s - x^*$ ,

$$\lambda_s^2 \delta_{s+1} - \lambda_{s-1}^2 \delta_s^2 \leq \frac{\beta}{2} \left( \|u_s\|^2 - \|u_{s+1}\|^2 \right).$$

Summing these inequalities from  $s = 1$  to  $s = t - 1$  one obtains:

$$\delta_t \leq \frac{\beta}{2\lambda_{t-1}^2} \|u_1\|^2.$$

By induction it is easy to see that  $\lambda_{t-1} \geq \frac{t}{2}$  which concludes the proof.

□

This entry was posted in [Optimization](#). Bookmark the [permalink](#).

[← ORF523: Oracle complexity of smooth convex functions](#)

[Guest post by Amir Ali Ahmadi: Sum of Squares \(SOS\) Techniques: An Introduction. Part I/II →](#)

## 9 Responses to "ORF523: Nesterov's Accelerated Gradient Descent"

- By [Backpropagation for dummies | Sachin Joglekar's blog](#) December 6, 2015 - 11:52 am

[...] the search will eventually come back to the required point as the momentum till go on reducing. Nesterov Momentum is another way of optimizing the use of momentum in gradient [...]

[Reply](#)

- By [Lagrange duality via the Fenchel conjugate | Look at the corners!](#) October 28, 2015 - 2:58 pm

[...] by some -strongly convex "regularizer", which will make the dual smooth, such that Nesterov's Accelerated Gradient Descent can be applied. Of course, we also need to control the approximation error [...]

[Reply](#)



- 

By mistake? October 14, 2015 - 7:24 pm

After 5, if you sum from  $s=1$  to  $s=t-1$  you should get on the right side  $(\|u_1\|^2 - \|u_t\|^2)$  right?

[Reply](#)



By Anonymous October 13, 2015 - 11:28 am

can we use this method with active set method ??

[Reply](#)



By Coordinate Descent September 7, 2015 - 10:28 pm

Is it possible to apply the nesterov acceleration to the second order newton method? and to the block coordinate descent method?

[Reply](#)



By Sebastien Bubeck September 10, 2015 - 11:13 pm

These are very good questions, both answered by Nesterov. For accelerating Newton's method see this paper: <http://link.springer.com/article/10.1007%2Fs10107-006-0089-x> ; and for accelerating coordinate descent see this: [http://www.optimization-online.org/DB\\_FILE/2010/01/2527.pdf](http://www.optimization-online.org/DB_FILE/2010/01/2527.pdf) .

- By [Nesterov's Accelerated Gradient Descent](#) | December 18, 2013 - 3:15 am

[...] 转自:<http://blogs.princeton.edu/imabandit/2013/04/01/acceleratedgradientdescent/> [...]

[Reply](#)

- By [NIPS 2013 | spider's space](#) | December 15, 2013 - 1:29 pm

[...] Both SDCA and SAG have a linear dependency on the condition number . For the deterministic case Nesterov's accelerated gradient descent attains a linear dependency on . This paper partially bridges the gap between these results and [...]

[Reply](#)

- By [The Zen of Gradient Descent | Moody Rd](#) | September 7, 2013 - 2:29 pm

[...] Bubeck's course notes are [...]

[Reply](#)

# Leave a reply

Name

Email(will not be published)

Website

☐ Notify me of follow-up comments by email.

☐ Notify me of new posts by email.

• Search for:

## . Archives

Archives  

## . Categories

Categories  

## . Recent Posts


- [Guest post by Miklos Racz: The fundamental limits of dimension estimation in random geometric graphs](#)
- [Guest post by Miklos Racz: Estimating the dimension of a random geometric graph on a high-dimensional sphere](#)
- [Guest post by Miklos Racz: A primer on exact recovery in the general stochastic block model](#)
- [Kernel-based methods for convex bandits, part 3](#)
- [Kernel-based methods for convex bandits, part 2](#)

## . Subscribe to Blog via Email

Enter your email address to subscribe to this blog and receive notifications of new posts by email.

Join 249 other subscribers

## . Meta

- [Log in](#)
-  [RSS - Posts](#)
-  [RSS - Comments](#)

## . Blogroll

- [Combinatorics and more](#)
- [Computational Complexity](#)
- [Godel's Lost Letter](#)
- [Gowers's Weblog](#)
- [hunch.net](#)
- [in theory](#)
- [Normal Deviate](#)
- [Nuit Blanche](#)
- [Shtetl-Optimized](#)
- [Stochastic Analysis Seminar](#)
- [The Geomblog](#)
- [What's new](#)

[Theme by Simple Themes](#)

[Princeton University](#)

[I'm a bandit](#)

© 2016 The Trustees of [Princeton University](#)