Mathematics of Image and Data Analysis
Math 5467

Nesterov's accelerated gradient descent

Instructor: Jeff Calder
Email: jcalder@umn.edu

http://www-users.math.umn.edu/~jwcalder/5467

Last time

- Heavy ball method

Today

- Continuum heavy ball method
- Nesterov's accelerated gradient descent

Heavy ball method

One of the oldest momentum based methods is the heavy ball method of Polyak. The heavy ball method iterates

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + \beta(x_k - x_{k-1}), \tag{1}$$

where α is the time step and $\beta \in [0,1]$ is the momentum parameter, where $x_1 = x_0$.

- The idea is that the descent direction has *memory*, or *momentum*. This averages out the bouncing effect in gradient descent, and accelerates convergence when the descent directions align over many iterations (near the minimizer).
- As we will see, the descent equations share similarities with the equations of motion for a ball rolling down the energy landscape, so it is also called the *heavy ball method*.

Continuum perspective: Heavy ball method

Recall the heavy ball method is a discretization of the ODE

$x' = -\nabla f$

GD.

$$x''(t) + ax'(t) = -\nabla f(x(t)), \tag{2}$$

where $a = \frac{1-\beta}{\sqrt{\alpha}}$.

Theorem 1. *Suppose $x(t)$ solves (2) with $x(0) = x_0 \in \mathbb{R}^n$, $x'(0) = 0$, and assume f is L-Lipschitz and μ-strongly convex. Let $x_* \in \mathbb{R}^n$ denote the unique minimizer of f. Then we have*

$$\|x(t) - x_*\|^2 \leq \frac{1}{3\mu}\left(3L + 2a^2\right)\|x_0 - x_*\|^2 \exp\left(-\frac{2\mu a t}{3L + 2a^2}\right). \tag{3}$$

Proof: Define the energy ($y(t) = x(t) - x_*$)

$$e(t) = 3\Big(\underbrace{\tfrac{1}{2}\|y'\|^2}_{\text{kinetic energy}} + \underbrace{f(x) - f(x_*)}_{\text{potential}}\Big) + \frac{a^2}{2}\|y\|^2 + a y^T y'$$

Total energy = kinetic + potential

<u>Goal</u>: $e'(t) \le -c\,e(t) \longrightarrow e(t) \le e(0)e^{-ct}$

① $\underline{e(t) \ge 0}$: By <u>strong convexity</u> of f

$$e(t) \ge \frac{1}{2}\|y'\|^2 + \frac{3\mu}{2}\|\underbrace{x - x_*}_{y}\|^2 + \frac{a^2}{2}\|y\|^2 + a y^T y'$$

$$= \frac{1}{2}\|y'\|^2 + \frac{3\mu}{2}\|y\|^2 + \frac{1}{2}\left(\|ay + y'\|^2 - \|y'\|^2\right)$$

$$= \|y'\|^2 + \frac{3\mu}{2}\|y\|^2 + \frac{1}{2}\|ay + y'\|^2 \ge 0$$

$$\ge \frac{3\mu}{2}\|y\|^2 = \frac{3\mu}{2}\|x(t) - x_*\|^2$$

then $$\boxed{\|x(t) - x_*\|^2 \le \frac{2}{3\mu} e(t)}$$

$$e(t) = 3\left(\frac{1}{2}\|y'\|^2 + f(x) - f(x_*)\right) + \frac{a^2}{2}\|y\|^2 + a y^T y'$$

Use $x' = y'$, $x'' = y''$

$$e'(t) = 3 y'^T y'' + 3\nabla f(x)^T x' + a^2 y^T y' + a\|y'\|^2 + a y^T y''$$

$$= 3 y'^T \underbrace{(y'' + \nabla f(x))}_{-ax' = -ay'} + a y^T \underbrace{(y'' + a y')}_{-\nabla f(x)} + a\|y'\|^2$$

$$x''(t) + ax'(t) = -\nabla f(x(t)),$$

$$= -3a\|y'\|^2 - a\nabla f(x)^T y + a\|y'\|^2$$

$$= -2a\|y'\|^2 - a\underbrace{(\nabla f(x) - \underbrace{\nabla f(x_*)}_{=0})^T(x - x_*)}_{\geq \mu\|x - x_*\|^2 = \mu\|y\|^2}$$

strong convexity

$$\leq -a(\mu\|y\|^2 + 2\|y'\|^2)$$

$$e'(t) \leq \underbrace{-a(\mu\|y\|^2 + 2\|y'\|^2)}_{\text{want } \leq -Cae(t)}$$

Need upper bound for $e(t)$

$$a y^T y' \leq (a\|y\|)\|y'\| \leq \frac{a^2}{2}\|y\|^2 + \frac{1}{2}\|y'\|^2$$

Cauchy-Schwarz

$$ab \leq \frac{1}{2}a^2 + \frac{1}{2}b^2$$

Cauchy's Ineq.

$$0 \leq (a-b)^2 = a^2 - 2ab + b^2$$

$$f(x) - f(x_*) \leq \underbrace{\nabla f(x_*)}_{=0} + \frac{L}{2}\|x - x_*\|^2$$

∇f L-Lipschitz

$$f(x) - f(x_*) \leq \frac{L}{2}\|y\|^2$$

$$e(t) = 3\left(\frac{1}{2}\|y'\|^2 + f(x) - f(x_*)\right) + \frac{a^2}{2}\|y\|^2 + a y^T y'$$

$$\le \frac{3}{2}\|y'\|^2 + \frac{3L}{2}\|y\|^2 + \frac{a^2}{2}\|y\|^2 + \frac{a^2}{2}\|y\|^2 + \frac{1}{2}\|y'\|^2$$

$$= \left(\frac{3L}{2} + a^2\right)\|y\|^2 + 2\|y'\|^2$$

$$= \underbrace{\left(\frac{3L+2a^2}{2\mu}\right)}_{\ge 1 \ \text{since} \ \frac{3L}{2\mu} \ge \frac{L}{\mu} \ge 1,\ \mu \le L} \mu\|y\|^2 + 2\|y'\|^2$$

$$e(t) \le \left(\frac{3L+2a^2}{2\mu}\right)\left(\mu\|y\|^2 + 2\|y'\|^2\right)$$

$$\mu\|y\|^2 + 2\|y'\|^2 \geq \frac{2\mu e(t)}{3L+2a^2}$$

$$e'(t) \leq -a\left(\mu\|y\|^2 + 2\|y'\|^2\right)$$

$$\leq \boxed{-\frac{2\mu a}{3L+2a^2}}\, e(t)$$

$$\Longrightarrow \quad e(t) \leq e(0)\exp\left(\frac{-2\mu a t}{3L+2a^2}\right)$$

$$\|x(t) - x_*\|^2 \le \frac{2}{3\mu} e(t) \le \frac{2}{3\mu} \underline{e(0)} \exp\left(\frac{-2\mu a t}{3L + 2a^2} \right)$$

To complete the proof,

$$e(0) \le \left(\frac{3L + 2a^2}{2\mu} \right) \left(\underbrace{\mu \|y\|^2}_{\mu \|x_0 - x_*\|^2} + \overbrace{2 \|y'\|^2}^{=0} \right)$$

$$= \left(\frac{3L + 2a^2}{2} \right) \|x_0 - x_*\|^2$$

Nesterov's Accelerated Gradient Descent

Set $\lambda_0 = 0$ and define λ_k by

$$\lambda_k = \frac{1 + \sqrt{1 + 4\lambda_{k-1}^2}}{2}. \tag{4}$$

Nesterov's accelerated gradient descent method then corresponds to the iteration scheme

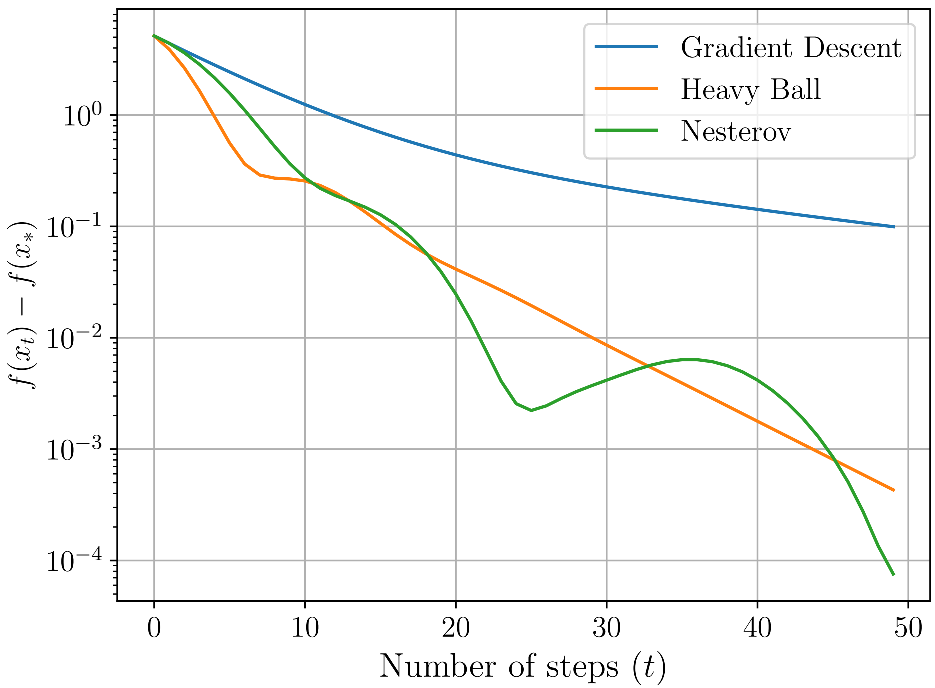
$$\begin{cases} y_{k+1} = x_k - \alpha \nabla f(x_k) \\ x_{k+1} = y_{k+1} + \dfrac{\lambda_k - 1}{\lambda_{k+1}}(y_{k+1} - y_k), \end{cases} \tag{5}$$

Theorem 2. *Assume f is convex and ∇f is L-Lipschitz. If $\alpha \le \frac{1}{L}$ then Nesterov's accelerated gradient descent satisfies*

$$f(y_t) - f(x_*) \le \frac{2\|x_1 - x_*\|^2}{\alpha(t-1)^2}. \tag{6}$$

G.D. $O(\frac{1}{t})$

Comparison



Gradient Descent
Heavy Ball
Nesterov
10^0
10^{-1}
10^{-2}
10^{-3}
10^{-4}
$f(x_t) - f(x_*)$
0
10
20
30
40
50
Number of steps (t)

Proposition about λ_k

Proposition 3. *For all $k \geq 1$ we have*

$$\frac{k}{2} \leq \lambda_k \leq \frac{k}{2} + \frac{1}{4}(3 + \log(k)). \tag{7}$$

$\lambda_0 = 0, \lambda_1 = 1$

$$\lambda_k = \frac{1 + \sqrt{1 + 4\lambda_{k-1}^2}}{2}.$$

λ_k solves $\lambda_k^2 - \lambda_k - \lambda_{k-1}^2 = 0$

$$\boxed{\lambda_k(\lambda_k - 1) = \lambda_{k-1}^2}$$

$$\rightarrow \lambda_k \geq \frac{1}{2} + \frac{1}{2}\sqrt{4\lambda_{k-1}^2}$$

$$= \frac{1}{2} + \lambda_{k-1} \xrightarrow{\text{induction}} \lambda_k \geq \frac{k}{2}$$

$$\lambda_k \le \frac{1 + 1 + 2\lambda_{k-1}}{2} \qquad \sqrt{a^2+b^2} \le a+b$$

$$= 1 + \lambda_{k-1} \longrightarrow \lambda_k \le k$$

For k large, $\lambda_k \sim \frac{k}{2}$

$$\begin{cases} y_{k+1} = x_k - \alpha \nabla f(x_k) \\ x_{k+1} = y_{k+1} + \dfrac{\lambda_k - 1}{\lambda_{k+1}}(y_{k+1} - y_k), \end{cases}$$

$$\frac{\lambda_k - 1}{\lambda_{k+1}} \sim \frac{\frac{k}{2} - 1}{\frac{k+1}{2}} = \frac{k-2}{k+1}$$

Alternative Nesterov

$$x_{k+1} = y_{k+1} + \left(\frac{k-2}{k+1}\right)(y_{k+1} - y_k)$$

Proof of convergence rate:

$$\begin{cases} y_{k+1} = x_k - \alpha \nabla f(x_k) \\ x_{k+1} = y_{k+1} + \dfrac{\lambda_k - 1}{\lambda_{k+1}}(y_{k+1} - y_k), \end{cases}$$

$$f(y_{k+1}) = f(x_k - \alpha \nabla f(x_k))$$ ← ∇f L-Lipschitz

$$\leq f(x_k) + \nabla f(x_k)^T(-\alpha \nabla f(x_k)) + \frac{L}{2}\|-\alpha \nabla f\|^2$$

$$= f(x_k) - \alpha \|\nabla f(x_k)\|^2 + \frac{L\alpha^2}{2}\|\nabla f(x_k)\|^2$$

Assume $\alpha \leq \frac{1}{L}$, $\frac{L\alpha^2}{2} \leq \frac{\alpha}{2}$

$$\leq f(x_k) - \frac{\alpha}{2}\left\|\frac{1}{\alpha}(y_{k+1} - x_k)\right\|^2$$

$$\boxed{f(y_{k+1}) = f(x_k) - \frac{1}{2\alpha}\|y_{k+1} - x_k\|^2}$$

Since f is convex:

$\nabla f(x_k) = \frac{1}{\alpha}(x_k - y_{k+1})$

$$f(y) \geq f(x_k) + \nabla f(x_k)^T (y - x_k)$$

$$f(x_k) \leq f(y) - \nabla f(x_k)^T (y - x_k)$$

$$= f(y) - \frac{1}{\alpha}(x_k - y_{k+1})^T (y - x_k)$$

$$= f(y) - \frac{1}{\alpha}(y_{k+1} - x_k)^T (x_k - y)$$

for any $y \in \mathbb{R}^n$:

$$f(y_{k+1}) = f(x_k) - \frac{1}{2\alpha}\|y_{k+1} - x_k\|^2$$

$$\leq f(y) - \frac{1}{2\alpha}\|y_{k+1} - x_k\|^2 - \frac{1}{\alpha}(y_{k+1} - x_k)^T(x_k - y)$$

$$f(y_{k+1}) - f(y) \leq -\frac{1}{2\alpha}\left(\|y_{k+1} - x_k\|^2 + 2(y_{k+1} - x_k)^T(x_k - y)\right)$$

Set $y = y_k$ and mult. by $\lambda_k - 1$

1

$$(\lambda_k - 1)\left(f(y_{k+1}) - f(y_k)\right) \leq -\frac{\lambda_k - 1}{2\alpha}\left(\|y_{k+1} - x_k\|^2 + \underline{2(y_{k+1} - x_k)^T}(x_k - y_k)\right)$$

Set $y = x_*$

2

$$f(y_{k+1}) - f(x_*) \le -\frac{1}{2\alpha}\left(\|y_{k+1} - x_k\|^2 + 2(y_{k+1} - x_k)^T(x_k - x_*)\right)$$

We add 1 + 2

$$\underline{LHS} = \lambda_k f(y_{k+1}) - (\lambda_k - 1) f(y_k) - f(x_*)$$

$$= \lambda_k\left(f(y_{k+1}) - f(x_*)\right) - (\lambda_k - 1) f(y_k) - f(x_*) + \lambda_k f(x_*)$$

$$= \lambda_k \delta_{k+1} - (\lambda_k - 1)\delta_k$$

$$\delta_k = f(y_{k+1}) - f(x_*)$$

RHS:

$$-\frac{\lambda_k}{2\alpha}\|y_{k+1}-x_k\|^2 - \frac{1}{\alpha}(y_{k+1}-x_k)^T\left((\lambda_k-1)(x_k-y_k)+x_k-x_*\right)$$

$$= -\frac{\lambda_k}{2\alpha}\|y_{k+1}-x_k\|^2 - \frac{1}{\alpha}(y_{k+1}-x_k)^T\left(\lambda_k x_k + (\lambda_k-1)y_k - x_*\right)$$

Multiply LHS and RHS by λ_k

$$\text{LHS} = \lambda_k^2\delta_{k+1} - \underbrace{\lambda_k(\lambda_k-1)}_{\lambda_{k-1}^2}\delta_k$$

$$= \underbrace{\lambda_k^2 \delta_{k+1} - \lambda_{k-1}^2 \delta_k}_{\text{Telescoping}} \quad , \quad \lambda_k \sim \frac{k}{2}$$

RHS =

$$-\frac{1}{2\alpha}\Big(\|\lambda_k(y_{k+1}-x_k)\|^2 + 2\lambda_k (y_{k+1}-x_k)^T(\lambda_k x_k + (\lambda_k -1) y_k - x_*)\Big)$$

$$= -\frac{1}{2\alpha}\Big(\|\lambda_k(y_{k+1} - \cancel{x_k}) + \cancel{\lambda_k x_k} + (\lambda_k-1)y_k - x_*\|^2 - \|\lambda_k x_k + (\lambda_k-1)y_k - x_*\|^2 \Big)$$

$$= -\frac{1}{2\alpha}\Big(\|\lambda_k y_{k+1} + (\lambda_k - 1) y_k - x_*\|^2 - \|\lambda_k x_k + (\lambda_k - 1) y_k - x_*\|^2 \Big)$$

$$= -\frac{1}{2\alpha}\Big(\|\underbrace{(\lambda_k - 1)(y_{k+1} - y_k)} + y_{k+1} - x_*\|^2 - \|\lambda_k x_k + (\lambda_k - 1) y_k - x_*\|^2 \Big)$$

$$\begin{cases} y_{k+1} = x_k - \alpha \nabla f(x_k) \\ x_{k+1} = y_{k+1} + \underbrace{\frac{\lambda_k - 1}{\lambda_{k+1}}(y_{k+1} - y_k)}, \end{cases}$$

$$(\lambda_k - 1)(y_{k+1} - y_k) = \lambda_{k+1}(x_{k+1} - y_{k+1})$$

$$= -\frac{1}{2\alpha}\Big(\|\lambda_{k+1} x_{k+1} - (\lambda_{k+1} - 1) y_{k+1} - x_*\|^2 - \|\lambda_k x_k + (\lambda_k - 1) y_k - x_*\|^2 \Big)$$

Telescoping

Sum LHS $\leq$ RHS over

$k=1$ to $t-1$

$$\sum_{k=1}^{t-1} \text{LHS} = \lambda_{t-1}^2 \delta_t = \lambda_{t-1}^2 \left(f(y_t) - f(x_*)\right)$$

$$\sum_{k=1}^{t-1} \text{RHS} \leq \frac{1}{2\alpha} \| \lambda_1 x_1 - (\lambda_1 - 1) y_1 - x_* \|^2$$

$\lambda_1 = 1$

$$= \frac{1}{2\alpha} \|x_1 - x_*\|^2$$

$$\lambda_{t-1}^2 \left(f(y_t) - f(x_*)\right) \leq \frac{\|x_1 - x_*\|^2}{2\alpha}$$

Use $\lambda_{t-1} \geq \frac{t-1}{2}$

$\Rightarrow$

$$f(y_t) - f(x_*) \leq \frac{4\,\|x_1 - x_*\|^2}{2\alpha\,(t-1)^2}$$