Advanced optimization methods
Gradient descent and beyond. Part 1

Alexandr Katrutsa

MIPT department of applied mathematics and computer science

September 26, 2018

Plan for today

- Gradient descent
- Heavy-ball method
- Nesterov accelerated GD
- Why accelerated?

Problem statement

$$\min_{x \in \mathbb{R}^n} f(x)$$

- f is smooth and convex
- If $\mu \leq f''(x) \leq L,\ \mu \geq 0,\ L > 0$, then f' is Lipschitz with constant L
- If $\mu > 0$, then f is μ-strongly convex, and

$$f(y) \geq f(x) + \langle f'(x), y - x \rangle + \frac{\mu}{2}\|y - x\|_2^2$$

- Condition number $\kappa = \frac{L}{\mu}$

Gradient descent

$$x_{k+1} = x_k - \alpha_k f'(x_k)$$

- Explicit scheme for discretization of ODE

$$\frac{dx}{dt} = -f'(x), \quad x(0) = x_0$$

- Minimization of upper bound at x_k

$$\min_x f(x_k) + \langle f'(x_k), x - x_k \rangle + \frac{1}{2\alpha_k}\|x - x_k\|_2^2,$$

- The best local descent direction

$$f(x_k + h_k) \approx f(x_k) + \langle f'(x_k), h_k \rangle < f(x_k)$$

Step size selection

- Constant $\alpha_k \equiv \text{const} < \frac{2}{L}$
- Decreasing sequence such that $\sum\limits_{k=1}^{\infty} \alpha_k = \infty$, i.e. $\frac{1}{k}, \frac{1}{\sqrt{k}}$, etc
- Backtracking search: Armijo, Goldstein, Wolfe rules and others
- Steepest descent: find the best possible α_k

Main point

The best parameter you select gives you small gain in convergence!

Convergence: any L-smooth function

$$
\begin{aligned}
f(x_{k+1}) \le f(x_k) + \langle f'(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|_2^2 = \\
f(x_k) - \alpha_k \|f'(x_k)\|_2^2 + \frac{L\alpha_k^2}{2}\|f'(x_k)\|_2^2 = \\
f(x_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right)\|f'(x_k)\|_2^2
\end{aligned}
$$

- Descent condition: $\alpha_k - \frac{L\alpha_k^2}{2} > 0 \Rightarrow \alpha_k < \frac{2}{L}$
- The best $\alpha_k^* = \arg\max\limits_{\alpha_k} \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) = \frac{1}{L}$
- $f(x_k) - f(x_{k+1}) \ge \frac{1}{2L}\|f'(x_k)\|_2^2$
- $\frac{1}{2L}\sum\limits_{k=0}^{T}\|f'(x_k)\|_2^2 \le f(x_0) - f(x_{T+1}) \le f(x_0) - f^*$
- f is bounded below, $\|f'(x_k)\|_2 \to 0,\ k \to \infty$

Convergence: L-smooth convex case

Theorem

Let f be L-smooth convex function and $\alpha = \frac{1}{L}$, then GD converges as

$$f(x_{k+1}) - f^* \leq \frac{2L\|x - x_0\|_2^2}{k+4} = \mathcal{O}(1/k)$$

Convergence: μ-strongly convex case

- μ-strong convexity implication

$$f(z) \geq f(x_k) + \langle f'(x_k), z - x_k \rangle + \frac{\mu}{2}\|z - x_k\|_2^2$$

- Minimize both side on z

$$f(x^*) \geq f(x_k) - \frac{1}{2\mu}\|f'(x_k)\|_2^2, \quad \|f'(x_k)\|_2^2 \geq 2\mu(f(x_k) - f^*)$$

- Recall that for $\alpha_k \equiv \frac{1}{L}$

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{2L}\|f'(x_k)\|_2^2$$

- Finally get linear rate

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{1}{\kappa}\right)(f(x_k) - f^*)$$

More precise estimate

Theorem

Let f be L-smooth and μ-strongly convex and $\alpha_k = \frac{2}{\mu + L}$, then GD converges as

$$f(x_k) - f^* \leq \frac{L}{2} \left(\frac{L - \mu}{L + \mu} \right)^{2k} \|x_0 - x^*\|_2^2$$

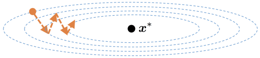
Gradient descent highlights

- Easy to implement
- It converges at least to stationary point
- Recent paper[1] shows that GD converges to a local minimizer **almost sure** with random initialization
- Linear convergence in strongly convex case
- It strongly depends on the condition number of $f''(x)$, random initial guess vector can help

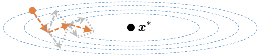
[1] https://arxiv.org/pdf/1602.04915.pdf

Heavy-ball method (Polyak, 1964)

$$x_{k+1} = x_k - \alpha_k f'(x_k) + \beta_k(x_k - x_{k-1})$$



gradient descent



heavy-ball method

Plot is from here[2]

- Two-step non-monotone method
- Discretization of the ODE with friction term

$$\ddot{x} + b\dot{x} + af'(x) = 0$$

- Connection between ODE and optimization methods
- CG is special case of this form

[2] http://www.princeton.edu/~yc5/ele538_optimization/lectures/accelerated_gradient.pdf

Convergence: μ-strongly convex

- Rewrite method as

$$\begin{bmatrix} x_{k+1} \\ x_k \end{bmatrix} = \begin{bmatrix} (1+\beta_k)I & -\beta_k I \\ I & 0 \end{bmatrix} \begin{bmatrix} x_k \\ x_{k-1} \end{bmatrix} + \begin{bmatrix} -\alpha_k f'(x_k) \\ 0 \end{bmatrix}$$

- Use theorem from calculus

$$\begin{bmatrix} x_{k+1} - x^* \\ x_k - x^* \end{bmatrix} = \underbrace{\begin{bmatrix} (1+\beta_k)I - \alpha_k \int_0^1 f''(x(\tau))d\tau & -\beta_k I \\ I & 0 \end{bmatrix}}_{=A_t} \begin{bmatrix} x_k - x^* \\ x_{k-1} - x^* \end{bmatrix},$$

  where $x(\tau) = x_k + \tau(x^* - x_k)$

- Convergence depends on the spectrum of the iteration matrix A_t
- Select α_k and β_k to make spectral radius the smallest

Parameter selection

Theorem

Let f be L-smooth and μ-strongly convex. Then $\alpha_k = \frac{4}{(\sqrt{L}+\sqrt{\mu})^2}$ and $\beta_k = \max(|1-\sqrt{\alpha_k L}|, |1-\sqrt{\alpha_k \mu}|)^2$ gives

$$\left\| \begin{bmatrix} x_{k+1} - x^* \\ x_k - x^* \end{bmatrix} \right\|_2 \leq \left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1} \right)^k \left\| \begin{bmatrix} x_1 - x^* \\ x_0 - x^* \end{bmatrix} \right\|_2$$

- Parameters depend on L and μ
- Faster than GD
- Similar to CG for μ-strongly convex quadratic
- Can such estimate be extend to L-smooth convex function?

Heavy-ball method highlights

- Simple two-step method
- Converges much faster than GD with appropriate α_k, β_k
- CG is particular case
- Proof only for μ-strongly convex functions

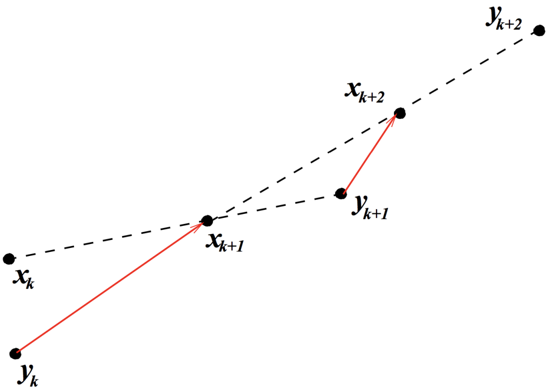
Nesterov accelerated GD (Nesterov, 1983)

One of possible notation variant

$$
\begin{aligned}
y_0 &= x_0 \\
x_{k+1} &= y_k - \alpha_k f'(y_k) \\
y_{k+1} &= x_{k+1} + \frac{k}{k+3}(x_{k+1} - x_k)
\end{aligned}
$$

- Heavy-ball comparison
- ODE interpretation again
- Non-monotone, too
- More details and options see in Part 2

Nesterov method visualization



Convergence: L-smooth convex

Theorem

Let f be convex and L-smooth. Assume $\alpha_k = \frac{1}{L}$. Then Nesterov method converges as

$$f(x_k) - f^* \leq \frac{2L\|x_0 - x^*\|_2^2}{(k+1)^2} = \mathcal{O}(1/k^2)$$

- Compare with GD convergence
- Iteration cost is almost the same

Convergence: μ-strongly convex

Theorem

Nesterov method for μ-strongly convex function f (with some additional assumptions) converges as

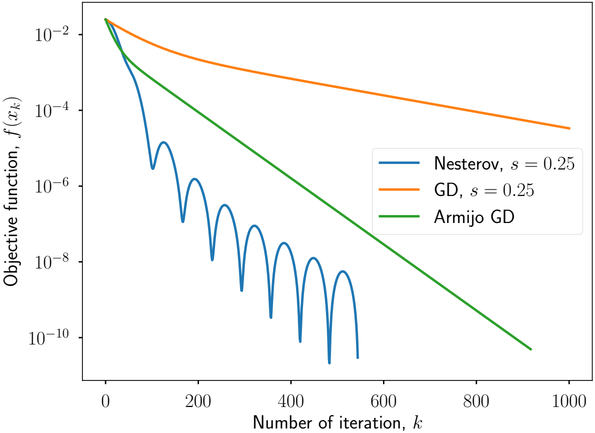
$$f(x_k) - f^* \leq L\|x_k - x_0\|_2^2 \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k$$

- Faster than GD
- Similar to heavy-ball method

Practical issues

- Rippling behaviour and restarts

$$f(x_1, x_2) = 2 \cdot 10^{-2} x_1^2 + 5 \cdot 10^{-3} x_2^2 \to \min, \; x_0 = (1, 1)$$



- Estimate of μ and L is separate problem

Why acceleration?

- It is faster than GD
- Is there even faster FOM for considered problems?

Lower bound concept

If we have access only to (sub)gradient in any point:

Convex functions	Lower bound
Nonsmooth	$f(x_k) - f^* \geq \frac{G\|x^* - x_0\|_2^2}{2(1+\sqrt{k+1})}$
L-smooth convex	$f(x_k) - f^* \geq \frac{3L\|x_0 - x^*\|_2^2}{32(k+1)^2}$
μ-strongly convex	$f(x_k) - f^* \geq \frac{\mu}{2}\left(\frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}\right)^{2k}\|x_0 - x^*\|_2^2$

To be discussed in part 2

- Proximal methods
- Mirror descent

Next class announce

Some proofs from this class.

Stochastic modifications of the considered methods today

- SAG
- SAGA
- SVRG
- SEGA
- ...

References I

Beck, A. (2017).
First-Order Methods in Optimization, volume 25.
SIAM.

Nemirovsky, A. S. and Yudin, D. B. (1983).
Problem complexity and method efficiency in optimization.

Nesterov, Y. E. (1983).
A method for solving the convex programming problem with convergence rate $O(1/k^2)$.
In *Dokl. Akad. Nauk SSSR*, volume 269, pages 543–547.

O'Donoghue, B. and Candes, E. (2015).
Adaptive restart for accelerated gradient schemes.
Foundations of computational mathematics, 15(3):715–732.

References II

Polyak, B. T. (1964).
Some methods of speeding up the convergence of iteration methods.
USSR Computational Mathematics and Mathematical Physics, 4(5):1–17.

Su, W., Boyd, S., and Candes, E. (2014).
A differential equation for modeling nesterov's accelerated gradient method: Theory and insights.
In *Advances in Neural Information Processing Systems*, pages 2510–2518.