Adaptive Gradient Descent without Descent

Yura Malitsky

MAI Optimization Seminar, 19 November 2020

LINKÖPING UNIVERSITY

Paper

Reference: ICML–2020, arxiv:1910.09529

Konstantin Mishchenko
(PhD student, KAUST)

Problem

We want to solve

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

where $f : \mathbb{R}^d \to \mathbb{R}$ is convex and differentiable.

How?

- Gradient descent
- Accelerated gradient methods
- Newton's methods
- Tensor methods
- Stochastic methods
- Coordinate methods

Gradient descent

$$x^{k+1} = x^k - \lambda \nabla f(x^k)$$

History:

O. Cauchy (1847), H. Curry (1944), A. Goldshtein (1962), B. Polyak (1963), L. Armijo (1966)

Theorem

Suppose f is convex, ∇f is L-Lipschitz, and $\lambda \in \left(0, \frac{2}{L}\right)$. Then $x^k \to x^* \in \operatorname{argmin} f$. For $\lambda = \frac{1}{L}$, the rate is

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

From discrete to continuous

$$x^{k+1} = x^k - \lambda \nabla f(x^k)$$

Let $x(t)$ be a continuous curve with $x(\lambda k) = x^k$.
For $t = \lambda k$,

$$x(t+\lambda) = x(t) - \lambda \nabla f(x(t))$$
$$\iff$$
$$\frac{x(t+\lambda) - x(t)}{\lambda} = -\nabla f(x(t))$$

If $\lambda \to 0$,

$$x'(t) = -\nabla f(x(t))$$

Continuous counterpart of GD:

$$x(0) = x_0$$
$$x'(t) = -\nabla f(x(t))$$

Let $\Psi(t) = \frac{1}{2}\|x(t) - x^*\|^2$ be a Lyapunov function. Then

$$\begin{aligned}
\frac{1}{2}\frac{\mathrm{d}}{\mathrm{d}t}\|x(t) - x^*\|^2 &= \langle x(t) - x^*, x'(t)\rangle \\
&= \langle x(t) - x^*, -\nabla f(x(t))\rangle \\
&\leq f(x^*) - f(x(t)) && //\text{convexity} \\
&\leq 0
\end{aligned}$$

$$\implies \qquad x(t) \to x^* \in \operatorname{argmin} f \qquad \text{and} \qquad f(x(t)) - f(x^*) \leq \frac{1}{2t}\|x_0 - x^*\|^2$$

From continuous to discrete: possible issues

$$x^{k+1} = x^k - \lambda \nabla f(x^k)$$

1. GD is not general: many functions are not L-smooth (i.e., gradients are not L-Lipschitz).
2. GD is not a free lunch: one needs to guess λ.
3. GD is not robust: with $\lambda \geq \frac{2}{L}$ may lead to divergence.
4. GD is slow: even if L is finite, it might be larger than local smoothness.

Workaround-1

What to do?

1. GD is not general: many functions are not L-smooth.
   Solutions: mirror descent with relative smoothness / dual preconditioning?
   [Birnbaum et.al., 2011, Bauschke et.al., 2016, Lu et.al., 2016, Maddison et.al.,2019]

$$\nabla h(x^{k+1}) = \nabla h(x^k) - \lambda \nabla f(x^k) \qquad \text{or} \qquad x^{k+1} = x^k - \lambda \nabla h(\nabla f(x^k))$$

   Cons: work only for specific f, still need to guess λ.
2. GD is not a free lunch: one needs to guess λ.
   Solution: line search?

$$\begin{aligned} \texttt{try} \;\; & \lambda = \gamma^i \\ & x^{k+1} = x^k - \lambda \nabla f(x^k) \\ \texttt{until} \;\; & f(x^{k+1}) \le f(x^k) - c\|\nabla f(x^k)\|^2 \end{aligned}$$

   Cons: more expensive than GD

Workaround-2

3. GD is slow

   Solution: Polyak's stepsize?

$$\lambda_k = \frac{f(x^k) - f_*}{\|\nabla f(x^k)\|^2}$$
$$x^{k+1} = x^k - \lambda_k \nabla f(x^k)$$

   Cons: needs f_*

   Solution-2: Barzilai-Borwein stepsize?

$$\lambda_k = \frac{\langle \nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k-1} \rangle}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2}$$
$$x^{k+1} = x^k - \lambda_k \nabla f(x^k)$$

   Cons: guarantees only for quadratic f, doesn't work in general.
   Counterexample in [Burdakov et.al., 2019]

Required tools

Law of cosines:

$$\|a+b\|^2 = \|a\|^2 + 2\langle a, b\rangle + \|b\|^2$$

Convexity:

$$\langle \nabla f(x), y-x\rangle \leq f(y) - f(x)$$

Smoothness:

$$\|\nabla f(y) - \nabla f(x)\| \leq L\|y-x\|$$

$$\overset{\text{convexity}}{\Longleftrightarrow}$$

$$f(y) \leq f(x) + \langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2$$

descent inequality

Standard analysis of GD

$$x^{k+1} = x^k - \lambda \nabla f(x^k)$$

Law of cosines:

$$\begin{aligned} \|x^{k+1} - x^*\|^2 &= \|x^{k+1} - x^k + x^k - x^*\|^2 \\ &= \|x^k - x^*\|^2 + 2\langle x^{k+1} - x^k, x^k - x^* \rangle + \|x^{k+1} - x^k\|^2 \\ &= \|x^k - x^*\|^2 + 2\lambda \langle \nabla f(x^k), x^* - x^k \rangle + \|x^{k+1} - x^k\|^2 \end{aligned}$$

Convexity:

$$2\lambda \langle \nabla f(x^k), x^* - x^k \rangle \leq 2\lambda \big(f(x^*) - f(x^k) \big)$$

Smoothness:

$$f(x^{k+1}) \leq f(x^k) + \langle \nabla f(x^k), x^{k+1} - x^k \rangle + \frac{L}{2} \|x^{k+1} - x^k\|^2$$

$\iff$

$$f(x^{k+1}) \leq f(x^k) - \frac{2 - \lambda L}{2\lambda} \|x^{k+1} - x^k\|^2$$

Standard analysis of GD

Summing up,

$$\|x^{k+1} - x^*\|^2 + 2\lambda(f(x^{k+1}) - f(x^*)) \leq \|x^k - x^*\|^2 - (1 - \lambda L)\|x^{k+1} - x^k\|^2$$

Let $\Psi_k = \|x^k - x^*\|^2$, $\lambda \leq \frac{1}{L}$

$$\Psi_{k+1} + 2\lambda(f(x^{k+1}) - f(x^*)) \leq \Psi_k$$

Proposed algorithm

$$
\begin{aligned}
x^{k+1} &= x^k - \lambda_k \nabla f(x^k) \\
L_k &= \frac{\|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|} \\
\lambda_k &= \frac{1}{L_k}
\end{aligned}
$$

$$
\begin{aligned}
x^{k+1} &= x^k - \lambda_k \nabla f(x^k) \\
L_k &= \frac{\|x^k - x^{k-1}\|}{\|\nabla f(x^k) - \nabla f(x^{k-1})\|} \\
\lambda_k &= \min\left\{\sqrt{1+\theta_{k-1}}\lambda_{k-1} \quad , \frac{1}{2L_k}\right\} \\
\theta_k &= \frac{\lambda_k}{\lambda_{k-1}}
\end{aligned}
$$

Adaptive Gradient Descent without Descent

$$x^{k+1} = x^k - \lambda_k \nabla f(x^k)$$

$$\lambda_k = \min\left\{ \sqrt{1+\theta_{k-1}}\lambda_{k-1}, \frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|} \right\}$$

$$\theta_k = \frac{\lambda_k}{\lambda_{k-1}}$$

New energy:

$$\Psi_{k+1} = \|x^{k+1} - x^*\|^2 + 2\lambda_k(1+\theta_k)\big(f(x^k) - f(x^*)\big) + \frac{1}{2}\|x^{k+1} - x^k\|^2$$

Decrease of energy:

$$\Psi_{k+1} \le \Psi_k + \left(\lambda_k^2\|\nabla f(x^k) - \nabla f(x^{k-1})\|^2 - \frac{1}{4}\|x^k - x^{k-1}\|^2\right)$$
$$+ 2\left(\lambda_{k-1}(1+\theta_{k-1}) - \lambda_k\theta_k\right)\left(f(x^{k-1}) - f(x^*)\right)$$

Convergence

Adaptive gradient descent without descent:

$$\lambda_k = \min\left\{\sqrt{1+\theta_{k-1}}\lambda_{k-1}, \frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}$$
$$x^{k+1} = x^k - \lambda_k \nabla f(x^k)$$
$$\theta_k = \frac{\lambda_k}{\lambda_{k-1}}$$

Theorem

Suppose that $f : \mathbb{R}^d \to \mathbb{R}$ is convex with locally Lipschitz gradient ∇f.
Then $x^k \to x^* \in \operatorname{argmin} f$ and

$$f(\hat{x}^k) - f(x^*) \le \frac{C}{\sum_{i=1}^k \lambda_i} = \mathcal{O}\left(\frac{1}{k}\right).$$

How good is it?

l_2-regularized logistic regression:

$$\frac{1}{n}\sum_{i=1}^{n}\log(1+e^{-b_i a_i^\top x})+\frac{\gamma}{2}\|x\|^2$$

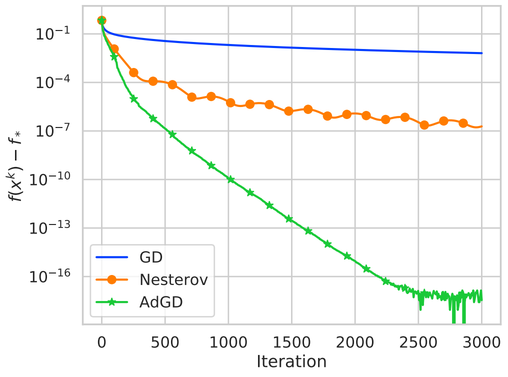

Figure 1: `mushroom` dataset

Strongly convex case

Let f be μ-strongly convex, i.e.,

$$\alpha f(x) + (1-\alpha) f(y) \geq f(\alpha x + (1-\alpha) y) + \frac{\alpha(1-\alpha)}{2} \mu \|x - y\|^2$$

GD complexity for $\|x^k - x^*\|^2 \leq \varepsilon$ is $\mathcal{O}(\frac{L}{\mu} \log \frac{1}{\varepsilon})$

Our complexity for $\|x^k - x^*\|^2 \leq \varepsilon$ is $\mathcal{O}(\frac{L'}{\mu'} \log \frac{1}{\varepsilon})$,

where L', μ' are *local* smoothness and strong convexity on $\overline{\text{conv}}\{x_0, x_1, \dots\}$

Acceleration (heuristic)

When f is μ-strongly convex and L-smooth, the "best" GD-type method is

$$y^{k+1} = x^k - \frac{1}{L}\nabla f(x^k),$$
$$x^{k+1} = y^{k+1} + \beta(y^{k+1} - y^k),$$

where $\beta = \frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}$ [Nesterov, 2004]

We know how to estimate L locally:

$$\lambda_k = \min\left\{\sqrt{1+\frac{\theta_{k-1}}{2}}\lambda_{k-1}, \frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}$$

What about μ? f is μ-strongly convex $\implies f^*$ is $\frac{1}{\mu}$-smooth. Hence,

$$\Lambda_k = \min\left\{\sqrt{1+\frac{\Theta_{k-1}}{2}}\Lambda_{k-1}, \frac{\|p^k - p^{k-1}\|}{2\|\nabla f^*(p^k) - \nabla f^*(p^{k-1})\|}\right\}$$

What is p_k? Let's set $p_k = \nabla f(x^k)$ and use $\nabla f^*(\nabla f(x)) = x$

Adaptive "accelerated" gradient descent

$$\lambda_k = \min\left\{\sqrt{1+\frac{\theta_{k-1}}{2}}\lambda_{k-1}, \frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}$$

$$\Lambda_k = \min\left\{\sqrt{1+\frac{\Theta_{k-1}}{2}}\Lambda_{k-1}, \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{2\|x^k - x^{k-1}\|}\right\}$$

$$\beta_k = \frac{\sqrt{1/\lambda_k} - \sqrt{\Lambda_k}}{\sqrt{1/\lambda_k} + \sqrt{\Lambda_k}}$$

$$y^{k+1} = x^k - \lambda_k \nabla f(x^k)$$

$$x^{k+1} = y^{k+1} + \beta_k (y^{k+1} - y^k)$$

$$\theta_k = \frac{\lambda_k}{\lambda_{k-1}}, \; \Theta_k = \frac{\Lambda_k}{\Lambda_{k-1}}$$

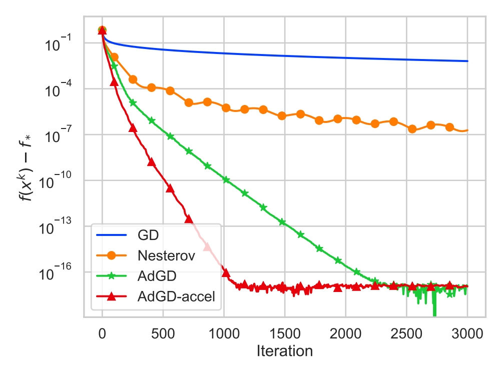

Figure 2: `mushroom` dataset

Stochastic extensions (heuristic)

$$\min_x \frac{1}{n}\sum_{i=1}^{n} f_i(x)$$

SGD:

$x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$

Adaptive SGD:

1. Sample ξ^k
2. $L_k = \frac{\|\nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(x^{k-1})\|}{\|x^k - x^{k-1}\|}$
3. $\lambda_k = \min\left\{\sqrt{1 + \frac{\theta_{k-1}}{\beta}}\lambda_{k-1}, \frac{\alpha}{L_k}\right\}$
4. $x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$
5. $\theta_k = \frac{\lambda_k}{\lambda_{k-1}}$

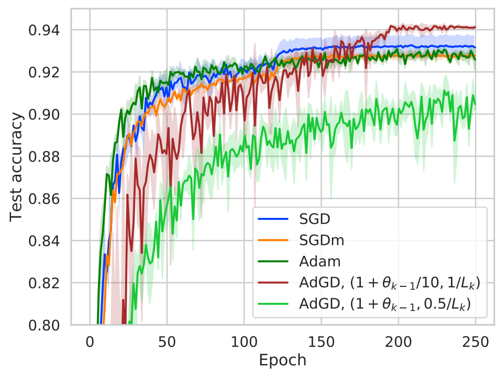

Figure 3: Test accuracy

Open questions

- Acceleration
- Mirror descent variant
- Nonconvexity