Adaptive Gradient Descent without Descent

Yura Malitsky

Variational Analysis and Optimisation Webinar series, 19 May 2021

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.

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?

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

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:

Cauchy (1847), Curry (1944), Goldshtein (1962), Polyak (1963), Armijo (1966)

Gradient descent

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

History:

Cauchy (1847), Curry (1944), Goldshtein (1962), Polyak (1963), Armijo (1966)

Def. f is L-smooth $\iff$ ∇f is L-Lipschitz $\iff$ $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$

Gradient descent

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

History:

Cauchy (1847), Curry (1944), Goldshtein (1962), Polyak (1963), Armijo (1966)

Def. f is L-smooth $\iff$ ∇f is L-Lipschitz $\iff$ $||\nabla f(x) - \nabla f(y)|| \leq L||x - y||$

Theorem

Suppose f is convex, L-smooth, and $\lambda \in \left(0, \frac{2}{L}\right)$.

Gradient descent

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

History:

Cauchy (1847), Curry (1944), Goldshtein (1962), Polyak (1963), Armijo (1966)

Def. f is L-smooth $\iff$ ∇f is L-Lipschitz $\iff$ $||\nabla f(x) - \nabla f(y)|| \le L||x - y||$

Theorem

Suppose f is convex, L-smooth, and $\lambda \in \left(0, \frac{2}{L}\right)$. Then $x^k \to x^* \in \operatorname{argmin} f$.

Gradient descent

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

History:

Cauchy (1847), Curry (1944), Goldshtein (1962), Polyak (1963), Armijo (1966)

Def. f is L-smooth $\iff$ ∇f is L-Lipschitz $\iff$ $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$

Theorem

Suppose f is convex, L-smooth, 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

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

From discrete to continuous

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

From discrete to continuous

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

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

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 $t = \lambda k$,

$$x(t + \lambda) = x(t) - \lambda \nabla f(x(t))$$

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 $t = \lambda k$,

$$x(t+\lambda) = x(t) - \lambda \nabla f(x(t))$$

$$\iff$$

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

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 $t = \lambda k$,

$$x(t+\lambda) = x(t) - \lambda \nabla f(x(t))$$

$$\Longleftrightarrow$$

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

If $\lambda \to 0$,

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

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 $t = \lambda k$,

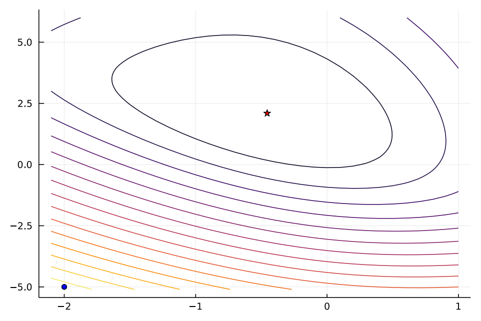
$$x(t+\lambda) = x(t) - \lambda \nabla f(x(t))$$

$$\Longleftrightarrow$$

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

If $\lambda \to 0$,

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



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 $t = \lambda k$,

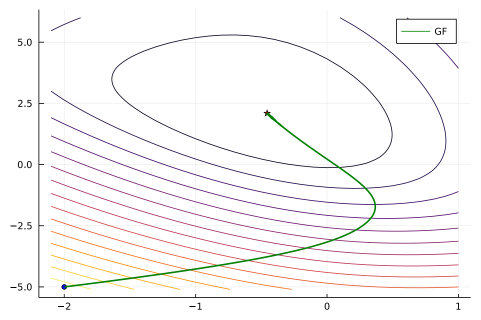
$$x(t + \lambda) = x(t) - \lambda \nabla f(x(t))$$

$$\iff$$

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

If $\lambda \to 0$,

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



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 $t = \lambda k$,

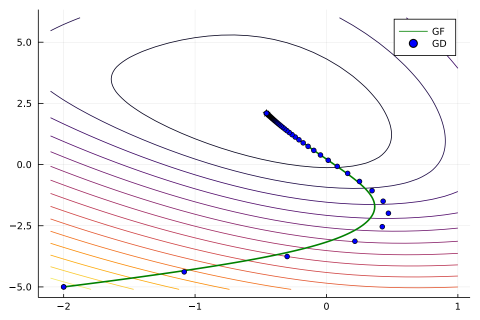
$$x(t + \lambda) = x(t) - \lambda \nabla f(x(t))$$

$$\iff$$

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

If $\lambda \to 0$,

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



Gradient flow

Continuous counterpart of GD:

$$\begin{aligned} x(0) &= x_0 \\ x'(t) &= -\nabla f(x(t)) \end{aligned}$$

Gradient flow

Continuous counterpart of GD:

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

Let $\Psi(t) = \|x(t) - x^*\|^2$ be a Lyapunov function.

Gradient flow

Continuous counterpart of GD:

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

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

$$\frac{\mathrm{d}}{\mathrm{d}t}\|x(t) - x^*\|^2 = 2\langle x(t) - x^*, x'(t)\rangle$$

Gradient flow

Continuous counterpart of GD:

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

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

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

Gradient flow

Continuous counterpart of GD:

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

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

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

Gradient flow

Continuous counterpart of GD:

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

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

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

Gradient flow

Continuous counterpart of GD:

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

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

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

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

From continuous to discrete: possible issues

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

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).
   Example: x^p, with $p > 2$; $\exp x$; $\log x$; $\tan x$

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).
   Example: x^p, with $p > 2$; $\exp x$; $\log x$; $\tan x$
2. GD is not a free lunch: one needs to guess λ.

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).
   Example: x^p, with $p > 2$; $\exp x$; $\log x$; $\tan x$
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.

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).
   Example: x^p, with $p > 2$; $\exp x$; $\log x$; $\tan x$
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?

Workaround-1

What to do?

- GD is not a free lunch: one needs to guess λ.

Workaround-1

What to do?

- GD is not a free lunch: one needs to guess λ.
  Solution: line search?

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

Workaround-1

What to do?

- GD is not a free lunch: one needs to guess λ.
  Solution: line search?

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

  Cons: more expensive than GD

Workaround-2

- GD is slow

Workaround-2

- GD is slow

  Solution: Polyak's stepsize?

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

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

Workaround-2

- 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_*

Workaround-2

- 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?

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

Workaround-2

- 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:

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

Required tools

Law of cosines:

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

Convexity:

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

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:

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

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:

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

$$\overset{\text{convexity}}{\iff}$$

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

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:

Convexity:

Smoothness:

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:

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

Smoothness:

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

$\iff$

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

Standard analysis of GD

If $\lambda \leq \frac{1}{L}$,

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

Standard analysis of GD

If $\lambda \leq \frac{1}{L}$,

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

Almost the same as in the continuous case:

$$\frac{\mathrm{d}}{\mathrm{d}t}\|x(t) - x^*\|^2 + 2(f(x(t) - f(x^*)) \leq 0$$

Standard analysis of GD

If $\lambda \leq \frac{1}{L}$,

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

Almost the same as in the continuous case:

$$\frac{\mathrm{d}}{\mathrm{d}t}\|x(t) - x^*\|^2 + 2(f(x(t) - f(x^*)) \leq 0$$

If $\Psi_k = \|x^k - x^*\|^2$ and $\Psi(t) = \|x(t) - x^*\|^2$,

$$\Psi_{k+1} + 2\lambda(f(x^{k+1}) - f(x^*)) \leq \Psi_k \qquad \text{vs.} \qquad \frac{\mathrm{d}}{\mathrm{d}t}\Psi(t) + 2(f(x(t)) - f(x^*)) \leq 0$$

Proposed algorithm

$$f \text{ is } L\text{-smooth} \iff \nabla f \text{ is } L\text{-Lipschitz} \iff \|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$$

Proposed algorithm

f is L-smooth $\iff$ ∇f is L-Lipschitz $\iff$ $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$

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

$$L_k = \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|}$$

$$\lambda_k = \frac{1}{L_k}$$

Proposed algorithm

f is L-smooth $\iff$ ∇f is L-Lipschitz $\iff$ $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$

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

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

Proposed algorithm

f is L-smooth $\iff$ ∇f is L-Lipschitz $\iff$ $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\|$

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

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

Algorithm description

Iteration k

Algorithm

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

$$L_k = \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|}$$

$$\lambda_k = \min\left\{\sqrt{1 + \theta_{k-1}}\lambda_{k-1}, \frac{1}{2L_k}\right\}$$

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

Algorithm description

Iteration k

Given x^k, $\nabla f(x^{k-1})$, θ_{k-1}

Algorithm

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

$$L_k = \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|}$$

$$\lambda_k = \min\left\{\sqrt{1 + \theta_{k-1}}\lambda_{k-1}, \frac{1}{2L_k}\right\}$$

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

Algorithm description

Iteration k

Given x^k, $\nabla f(x^{k-1})$, θ_{k-1}

1. Compute $\nabla f(x^k)$ and L_k

Algorithm

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

$$L_k = \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|}$$

$$\lambda_k = \min\left\{\sqrt{1+\theta_{k-1}}\lambda_{k-1}, \frac{1}{2L_k}\right\}$$

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

Algorithm description

Iteration k

Given x^k, $\nabla f(x^{k-1})$, θ_{k-1}

1. Compute $\nabla f(x^k)$ and L_k

Algorithm

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

$$L_k = \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|}$$

$$\lambda_k = \min\left\{\sqrt{1 + \theta_{k-1}}\lambda_{k-1}, \frac{1}{2L_k}\right\}$$

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

Algorithm description

Iteration k

Given x^k, $\nabla f(x^{k-1})$, θ_{k-1}

1. Compute $\nabla f(x^k)$ and L_k
2. Compute λ_k

Algorithm

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

$$L_k = \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|}$$

$$\lambda_k = \min\left\{\sqrt{1+\theta_{k-1}}\lambda_{k-1}, \frac{1}{2L_k}\right\}$$

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

Algorithm description

Iteration k

Given x^k, $\nabla f(x^{k-1})$, θ_{k-1}

1. Compute $\nabla f(x^k)$ and L_k
2. Compute λ_k
3. Compute x^{k+1} and θ_k

Algorithm

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

$$L_k = \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|}$$

$$\lambda_k = \min\left\{\sqrt{1 + \theta_{k-1}}\lambda_{k-1}, \frac{1}{2L_k}\right\}$$

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

Algorithm description

Iteration k

Given x^k, $\nabla f(x^{k-1})$, θ_{k-1}

1. Compute $\nabla f(x^k)$ and L_k
2. Compute λ_k
3. Compute x^{k+1} and θ_k

Algorithm

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

$$L_k = \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|}$$

$$\lambda_k = \min\left\{\sqrt{1+\theta_{k-1}}\lambda_{k-1}, \frac{1}{2L_k}\right\}$$

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

Algorithm description

Iteration k

Given x^k, $\nabla f(x^{k-1})$, θ_{k-1}

1. Compute $\nabla f(x^k)$ and L_k
2. Compute λ_k
3. Compute x^{k+1} and θ_k
4. Set $k = k + 1$

Algorithm

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

$$L_k = \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{\|x^k - x^{k-1}\|}$$

$$\lambda_k = \min\left\{\sqrt{1 + \theta_{k-1}}\lambda_{k-1}, \frac{1}{2L_k}\right\}$$

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

Adaptive Gradient Descent without Descent

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

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

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

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:

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

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:

$$\Psi_{k+1} \le \Psi_k$$

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:

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

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:

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

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:

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

Convergence

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*

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

Convergence

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*

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

- Local Lipschitzness $\iff$ Lipschitzness in the small neighborhood: x^p, with $p \geq 2$, $\exp(x)$, $\tan(x)$ all satisfy.

Convergence

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*

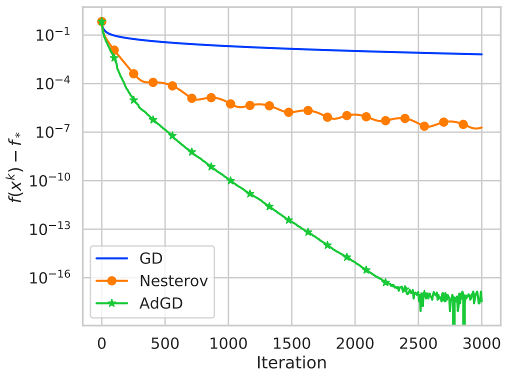
$$f(\hat{x}^k) - f(x^*) \leq \frac{C}{\sum_{i=1}^{k} \lambda_i} = \mathcal{O}\Big(\frac{1}{k}\Big).$$

- Local Lipschitzness $\iff$ Lipschitzness in the small neighborhood: x^p, with $p \geq 2$, $\exp(x)$, $\tan(x)$ all satisfy.
- If ∇f is L-Lipschitz, then $\lambda_i \geq \frac{1}{2L_i} \geq \frac{1}{2L} \implies \mathcal{O}\left(\frac{1}{k}\right)$ rate.

How good is it?

l_2-regularized logistic regression:

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



mushroom dataset

Strongly convex case

Let f be μ-strongly convex, i.e.,

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

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 $\mathcal{O}(\frac{L}{\mu}\log\frac{1}{\varepsilon})$

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 $\mathcal{O}(\frac{L'}{\mu'}\log\frac{1}{\varepsilon})$,

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 $\overline{\mathrm{conv}}\{x_0, x_1, \dots\}$

Heuristics

Acceleration (heuristic)

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

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

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

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]

GD complexity: $\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$ vs. **Accelerated GD complexity:** $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$

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]

GD complexity: $\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$ vs. **Accelerated GD complexity:** $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$

- We know how to estimate L locally.

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]

GD complexity: $\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$ vs. **Accelerated GD complexity:** $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$

- We know how to estimate L locally.
- What about μ?

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]

GD complexity: $\mathcal{O}\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$ vs. **Accelerated GD complexity:** $\mathcal{O}\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$

- We know how to estimate L locally.
- What about μ? f is μ-strongly convex $\Longrightarrow$ f^* is $\frac{1}{\mu}$-smooth.

Adaptive “accelerated” gradient descent

$$\lambda_k = \min\left\{\sqrt{1 + \frac{\theta_{k-1}}{2}}\lambda_{k-1}, \frac{\|x^k - x^{k-1}\|}{2\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}$$
$$\Lambda_k = \min\left\{\sqrt{1 + \frac{\Theta_{k-1}}{2}}\Lambda_{k-1}, \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{2\|x^k - x^{k-1}\|}\right\}$$
$$\beta_k = \frac{\sqrt{1/\lambda_k} - \sqrt{\Lambda_k}}{\sqrt{1/\lambda_k} + \sqrt{\Lambda_k}}$$
$$y^{k+1} = x^k - \lambda_k \nabla f(x^k)$$
$$x^{k+1} = y^{k+1} + \beta_k (y^{k+1} - y^k)$$
$$\theta_k = \frac{\lambda_k}{\lambda_{k-1}}, \ \Theta_k = \frac{\Lambda_k}{\Lambda_{k-1}}$$

Adaptive "accelerated" gradient descent

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

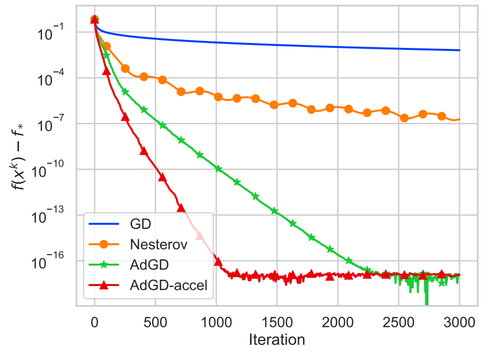
$$\Lambda_k = \min\left\{\sqrt{1+\frac{\Theta_{k-1}}{2}}\Lambda_{k-1}, \frac{\|\nabla f(x^k) - \nabla f(x^{k-1})\|}{2\|x^k - x^{k-1}\|}\right\}$$

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

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

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

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



mushroom dataset

Stochastic extensions (heuristic)

$$\min_x \frac{1}{n}\sum_{i=1}^{n} f_i(x), \qquad n \text{ is big}$$

Stochastic extensions (heuristic)

$$\min_x \frac{1}{n}\sum_{i=1}^{n} f_i(x), \qquad n \text{ is big}$$

SGD:

1. Sample $\xi^k \in \{1, \dots, n\}$
2. $x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$

Stochastic extensions (heuristic)

$$\min_x \frac{1}{n} \sum_{i=1}^{n} f_i(x), \qquad n \text{ is big}$$

SGD:

1. Sample $\xi^k \in \{1, \dots, n\}$
2. $x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$ ▷ $\lambda_k \to 0$ in theory, small λ_k in practice

Stochastic extensions (heuristic)

$$\min_x \frac{1}{n}\sum_{i=1}^n f_i(x), \qquad n \text{ is big}$$

SGD:

1. Sample $\xi^k \in \{1, \dots, n\}$
2. $x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$ ▷ $\lambda_k \to 0$ in theory, small λ_k in practice

Adaptive SGD:

Stochastic extensions (heuristic)

$$\min_x \frac{1}{n}\sum_{i=1}^{n} f_i(x), \qquad n \text{ is big}$$

SGD:

1. Sample $\xi^k \in \{1, \dots, n\}$
2. $x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$ ▷ $\lambda_k \to 0$ in theory, small λ_k in practice

Adaptive SGD:

1. Sample $\xi^k \in \{1, \dots, n\}$

Stochastic extensions (heuristic)

$$\min_x \frac{1}{n} \sum_{i=1}^{n} f_i(x), \qquad n \text{ is big}$$

SGD:

1. Sample $\xi^k \in \{1, \dots, n\}$
2. $x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$

▷ $\lambda_k \to 0$ in theory, small λ_k in practice

Adaptive SGD:

1. Sample $\xi^k \in \{1, \dots, n\}$
2. $L_k = \frac{\|\nabla f_{\xi^k}(x^k) - \nabla f_{\xi^k}(x^{k-1})\|}{\|x^k - x^{k-1}\|}$

▷ Cannot use $\nabla f_{\xi^{k-1}}(x^{k-1})$

Stochastic extensions (heuristic)

$$\min_x \frac{1}{n}\sum_{i=1}^n f_i(x), \qquad n \text{ is big}$$

SGD:

1. Sample $\xi^k \in \{1, \dots, n\}$
2. $x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$

▷ $\lambda_k \to 0$ in theory, small λ_k in practice

Adaptive SGD:

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

▷ Cannot use $\nabla f_{\xi^{k-1}}(x^{k-1})$

▷ α, β should be tuned

Stochastic extensions (heuristic)

$$\min_x \frac{1}{n}\sum_{i=1}^{n} f_i(x), \qquad n \text{ is big}$$

SGD:

1. Sample $\xi^k \in \{1, \dots, n\}$
2. $x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$

▷ $\lambda_k \to 0$ in theory, small λ_k in practice

Adaptive SGD:

1. Sample $\xi^k \in \{1, \dots, n\}$
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. $\theta_k = \frac{\lambda_k}{\lambda_{k-1}}$

▷ Cannot use $\nabla f_{\xi^{k-1}}(x^{k-1})$

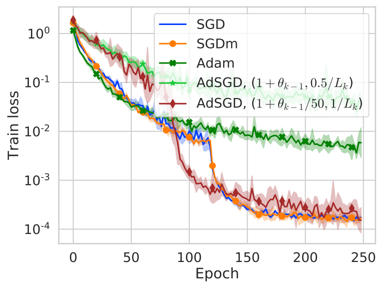
▷ α, β should be tuned

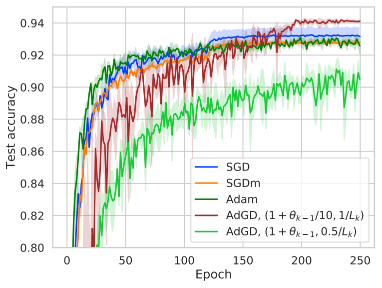
ResNet-18

$$\lambda_k = \min\left\{\sqrt{1 + \frac{\theta_{k-1}}{\beta}}\lambda_{k-1}, \frac{\alpha}{L_k}\right\} \qquad x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$$

ResNet-18

$$\lambda_k = \min\left\{\sqrt{1 + \frac{\theta_{k-1}}{\beta}}\lambda_{k-1}, \frac{\alpha}{L_k}\right\} \qquad x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$$



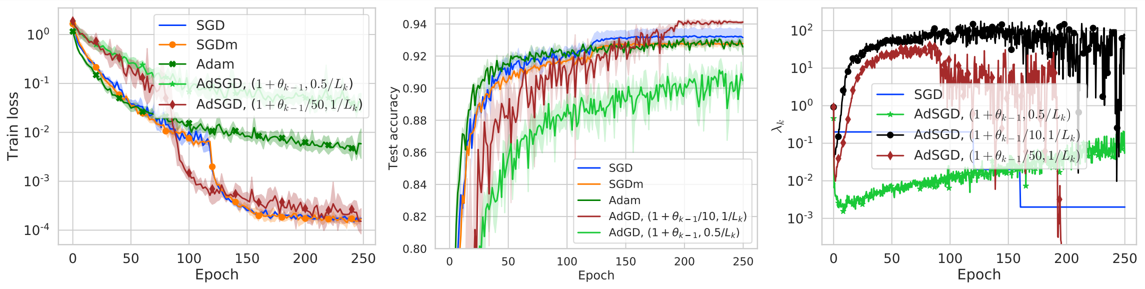


Train loss

Test accuracy

ResNet-18

$$\lambda_k = \min\left\{\sqrt{1 + \frac{\theta_{k-1}}{\beta}}\lambda_{k-1}, \frac{\alpha}{L_k}\right\} \qquad x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$$



Train loss　　Test accuracy　　Learning rate

Open questions

- Acceleration

Open questions

- Acceleration
- Mirror descent variant

Open questions

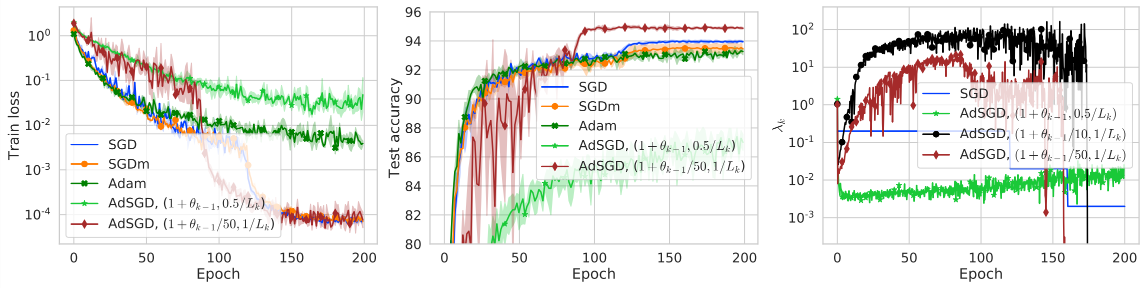
- Acceleration
- Mirror descent variant
- Nonconvexity

Open questions

- Acceleration
- Mirror descent variant
- Nonconvexity
- Robust version of adaptive SGD

DenseNet-121

$$\lambda_k = \min\left\{\sqrt{1+\frac{\theta_{k-1}}{\beta}}\lambda_{k-1}, \frac{\alpha}{L_k}\right\} \qquad x^{k+1} = x^k - \lambda_k \nabla f_{\xi^k}(x^k)$$



Train loss | Test accuracy | Learning rate