AGDwoD for Quadratic function

Derivation

Let f be a quadratic function $f(x) = \frac{1}{2}\langle x, Ax\rangle, \quad A \succeq 0.$

▼ **Analysis of GD for quadratic functions**

The general form of a quadratic function is $g(x) = \frac{1}{2}\langle x, Ax\rangle + \langle b, x\rangle + c, \quad A \succeq 0$. However, we can simplify it for analysis by variable substitution:
Let x^* be any minimizer of function. It exists given that $b \in R(A)$. We know that

$$\nabla g(x^*) = 0 \Rightarrow Ax^* = -b \tag{1}$$

Set $x = y + x^*$. Then, our function becomes

$$g(x) = \frac{1}{2}\left(\langle y, Ay\rangle + \langle x^*, Ay\rangle + \langle y, Ax^*\rangle + \langle x^*, Ax^*\rangle\right) + \langle b, x^*\rangle + \langle b, y\rangle + c$$

Using (1), as well as symmetry of A, the above simplifies to:

$$g(x) = \frac{\langle y, Ay\rangle + \langle x^*, Ax^*\rangle}{2} + \langle b, x^*\rangle + c = \frac{\langle y, Ay\rangle}{2} + c_0$$

Finally, we can redefine the function as $f(y) = g(x) - c_0$ to get a much more convenient representation for analysis.
Note that

$$y_* = 0 \quad \text{is a minimizer of } f \tag{2}$$
$$f(y_*) = f_* = 0 \tag{3}$$

▼ **Algorithm**

Algorithm 1: Adaptive gradient descent

1: Input: $x^0 \in \mathbb{R}^d, \lambda_0 > 0, \theta_0 = +\infty$
2: $x^1 = x^0 - \lambda_0 \nabla f(x^0)$
3: for $k = 1, 2, \ldots$ **do**
4: $\quad \lambda_k = \min\left\{\sqrt{1+\theta_{k-1}}\lambda_{k-1}, \frac{3\|x^k - x^{k-1}\|}{4\|\nabla f(x^k) - \nabla f(x^{k-1})\|}\right\}$
5: $\quad x^{k+1} = x^k - \lambda_k \nabla f(x^k)$
6: $\quad \theta_k = \frac{\lambda_k}{\lambda_{k-1}}$
7: end for

We assumed that for quadratic problem a bigger stepsize can be taken. That is why set λ_k such satisfies two inequalities

$$\begin{cases} \lambda_k^2 \le (1+\theta_{k-1})\lambda_{k-1}^2, \\ \lambda_k \le \frac{\beta||x^k - x^{k-1}||}{||\nabla f(x^k) - \nabla f(x^{k-1})||}, \end{cases}$$

And we want to find maximum possible β for which algorithm will converge and check whether $max(\beta) > \frac{1}{2}$.

Lemma 1

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable and let $x^* = \min\limits_{x\in\mathbb{R}^d} f(x) = 0$. *Then for* x^k*generated by the Algorithm it holds*

$$\frac{1}{2}||x^{k+1}||^2 + \left(\frac{3}{2} - \beta\right)||x^{k+1} - x^k||^2 + 2\lambda_k(1+\theta_k)f(x^k) \le \frac{1}{2}||x^k||^2 + \beta||x^k - x^{k-1}||^2 + 2\lambda_k\theta_k f(x^{k-1}). \quad (4)$$

Proof.

Let $k \ge 1$.

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

Using second-order Taylor expansion for our quadratic f, we have:

$$2\lambda_k\langle\nabla f(x^k), x^* - x^k\rangle = 2\lambda_k(f_* - f(x^k) - \frac{1}{2}\langle x^* - x^k, A(x^* - x^k)\rangle) \quad (5)$$

Which, since $Ax_* = 0$, simplifies to (using the definition of the GD step):

$$2\lambda_k\langle\nabla f(x^k), x^* - x^k\rangle = \langle x^{k+1} - x^k, x^k\rangle - 2\lambda_k f(x^k) \quad (6)$$

Thus, we can rewrite the initial equation as

$$||x^{k+1}||^2 = ||x^k||^2 + \langle x^{k+1} - x^k, x^k\rangle - 2\lambda_k f(x^k) + ||x^{k+1} - x^k||^2. \quad (7)$$

Next, let us examine $||x^{k+1} - x^k||^2$:

$$\begin{aligned}||x^{k+1} - x^k||^2 &= 2||x^{k+1} - x^k||^2 - ||x^{k+1} - x^k||^2 \text{ add and subtract} \\ &= -2\lambda_k\langle\nabla f(x^k), x^{k+1} - x^k\rangle - ||x^{k+1} - x^k||^2 \text{ apply G.D step def.} \\ &= 2\lambda_k\langle\nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k+1}\rangle \\ &\quad + 2\lambda_k\langle\nabla f(x^{k-1}), x^k - x^{k+1}\rangle - ||x^{k+1} - x^k||^2.\end{aligned}$$

Where in the last row, we added and subtracted $2\lambda_k\langle f(x^{k-1}), x^k - x^{k+1}\rangle$.

We can analyze parts of the above equation separately:

$$\begin{aligned}2\lambda_k\langle\nabla f(x^k) - \nabla f(x^{k-1}), x^k - x^{k+1}\rangle &\le 2\lambda_k||\nabla f(x^k) - \nabla f(x^{k-1})||||x^k - x^{k+1}|| \\ &\le 2\beta||x^k - x^{k-1}||||x^k - x^{k+1}|| \text{ as } \lambda_k \le \frac{\beta||x^k - x^{k-1}||}{||\nabla f(x^k) - \nabla f(x^{k-1})||} \\ &\le \beta(||x^k - x^{k-1}||^2 + ||x^{k+1} - x^k||^2)\end{aligned}$$

Where in the last inequality we used that $a^2 + b^2 \ge 2ab$.

$$\begin{aligned}\lambda_k\langle\nabla f(x^{k-1}), x^k - x^{k+1}\rangle &\overset{\text{G.D. step}}{=} \frac{2\lambda_k}{\lambda_{k-1}}\langle x^{k-1} - x^k, x^k - x^{k+1}\rangle \\ &\overset{\text{G.D. step}}{=} 2\lambda_k\theta_k\langle x^{k-1} - x^k, \nabla f(x^k)\rangle \\ &\overset{\text{convexity}}{\le} 2\lambda_k\theta_k(f(x^{k-1}) - f(x^k)).\end{aligned}$$

Combining these bounds we get

$$||x^{k+1}-x^k||^2 \le \beta||x^k-x^{k-1}||^2 + 2\lambda_k\theta_k(f(x^{k-1})-f(x^k)) + (\beta-1)||x^{k+1}-x^k||^2$$

Putting it into (7) we get

$$||x^{k+1}||^2 \le ||x^k||^2 + \langle x^{k+1}-x^k, x^k\rangle - 2\lambda_k f(x^k) + \beta||x^k-x^{k-1}||^2 + 2\lambda_k\theta_k(f(x^{k-1})-f(x^k)) + (\beta-1)||x^{k+1}-x^k||^2 \tag{8}$$

Note also, by cosine theorem:

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

Therefore, (8) can be reordered to:

$$\frac{1}{2}||x^{k+1}||^2 + \left(\frac{3}{2}-\beta\right)||x^{k+1}-x^k||^2 + 2\lambda_k(1+\theta_k)f(x^k) \le \frac{1}{2}||x^k||^2 + \beta||x^k-x^{k-1}||^2 + 2\lambda_k\theta_k f(x^{k-1}). \tag{9}$$

Which is the desired inequality $\blacksquare$.

To ensure that inequality from Lemma 1 represents a proper Lyapunov energy dissipation, we require:

$$\frac{3}{2}-\beta \ge \beta \Rightarrow \beta \le \frac{3}{4}$$

So, $\max(\beta)$ is indeed larger than $\frac{1}{2}$.

Let us now continue our analysis for $\beta = \frac{3}{4}$:

Theorem 1

For $f(x) = \frac{\langle x, Ax\rangle}{2}$, (x^k) generated by Algorithm 1 converges to a solution of (1) and we have that

$$f(\hat{x}^k) \le \frac{D}{2S_k} = \mathcal{O}\left(\frac{1}{k}\right)$$

where

$$\hat{x}^k = \frac{\lambda_k(1+\theta_k)x^k + \sum_{i=1}^{k-1} w_i x^i}{S_k},$$
$$w_i = \lambda_i(1+\theta_i) - \lambda_{i+1}\theta_{i+1},$$
$$S_k = \lambda_k(1+\theta_k) + \sum_{i=1}^{k-1} w_i = \sum_{i=1}^{k} \lambda_i + \lambda_1\theta_1$$

and D is a constant that explicitly depends on the initial data and the solution set.

Proof. Fix any $x^* = \min_{x\in\mathbb{R}^d} f(x) = 0$. Telescoping inequality from Lemma 1 we get

$$\sum_{i=1}^{k}\frac{1}{2}||x^{i+1}||^2 + \sum_{i=1}^{k}\frac{3}{4}||x^{i+1}-x^i||^2 + \sum_{i=1}^{k} 2\lambda_i(1+\theta_i)f(x^i) \tag{10}$$

$$\le \sum_{i=1}^{k}\frac{1}{2}||x^i||^2 + \sum_{i=1}^{k}\frac{3}{4}||x^i-x^{i-1}||^2 + \sum_{i=1}^{k} 2\lambda_i\theta_i f(x^{i-1}) \tag{11}$$

$$\frac{1}{2}||x^{k+1}||^2 + \frac{3}{4}||x^{k+1} - x^k||^2 + 2\lambda_k(1+\theta_k)f(x^k) \tag{12}$$

$$+ 2\sum_{i=1}^{k-1}\left[\lambda_i(1+\theta_i) - \lambda_{i+1}\theta_{i+1}\right]f(x^i) \tag{13}$$

$$\leq \frac{3}{4}||x^1 - x^0||^2 + \frac{1}{2}||x^1||^2 + 2\lambda_1\theta_1 f(x^0) = D \tag{14}$$

Notice that $2\sum_{i=1}^{k-1}\left[\lambda_i(1+\theta_i) - \lambda_{i+1}\theta_{i+1}\right](f(x^i) - f_*)$ is non-negative, as, $f(x^i) \geq 0$, and $\lambda_i(1+\theta_i) - \lambda_{i+1}\theta_{i+1} \geq 0 \iff \lambda_{i+1}^2 \leq \lambda_i^2(1+\theta_i)$, which is condition required by algorithm. Thus, sequence of x^k is bounded ($||x^{k+1}||^2 + \text{positive term} \leq D$). For a quadratic f, we know that is is L-Lipschitz, where L is the max eigenvalue of A:

$$||\nabla f(x) - \nabla f(y)|| \leq L||x-y|| \quad \forall x, y \in \mathbb{R}$$

Clearly, $\lambda_1 = \frac{3||x^1 - x^0||}{4||\nabla f(x^1) - \nabla f(x^0)||} \geq \frac{3}{4L}$ thus, by induction one can prove that $\lambda_k \geq \frac{3}{4L}$ in other words, the sequence (λ_k) is separated from zero.

Notice, that the total sum of coefficients at these terms is

$$\lambda_k(1+\theta_k) + \sum_{i=1}^{k-1}[\lambda_i(1+\theta_i) - \lambda_{i+1}\theta_{i+1}] = \sum_{i=1}^{k}\lambda_i + \lambda_1\theta_1 = S_k$$

Then apply Jensen inequality

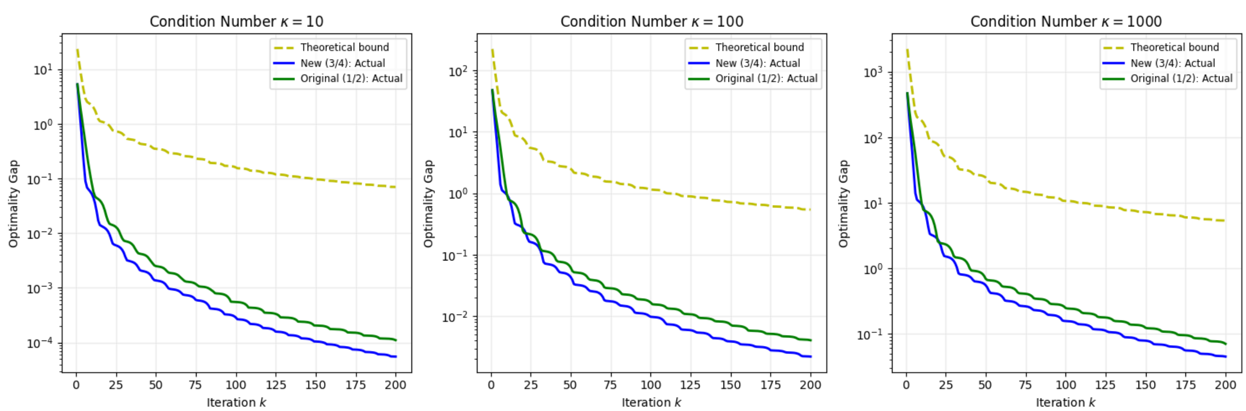
$$\lambda_k(1+\theta_k)f(x^k) + \sum_{i=1}^{k-1}\left[\lambda_i(1+\theta_i) - \lambda_{i+1}\theta_{i+1}\right]f(x^i) \geq S_k f(\hat{x}^k)$$

Consequently

$$\frac{D}{2} \geq S_k f(\hat{x}^k)$$

It proves convergence rate for the algorithm and boundeness of iterations.

Experiments



Here you can see comparison of original algorithm from paper, updated version for quadratic function and overimposed theoretical bound for updated algorithm