Articles on Gradient methods

October 17, 2022

Contents

1 OC15

From [OC15].

They use AGM in the form: $\theta_0 = 1$ and θ_k solves

$$\begin{aligned}\theta_{k+1}^2 &= (1-\theta_{k+1})\theta_k^2 + q\theta_{k+1}\\ \beta_k &= \theta_k(1-\theta_k)/(\theta_k^2+\theta_{k+1})\\ y_{k+1} &= x_{k+1} + \beta_k(x_{k+1}-x_k)\end{aligned}$$

For $q = 1$ we have the GM.

$$\begin{aligned}\theta_{k+1}^2 + (\theta_k^2 - q)\theta_{k+1} = \theta_k^2 \quad &\Leftrightarrow \quad \left(\theta_{k+1} + \frac{\theta_k^2 - q}{2}\right)^2 = \theta_k^2 + \frac{(\theta_k^2-q)^2}{4}\\ &\Leftrightarrow \quad \theta_{k+1} = \sqrt{\theta_k^2 + \frac{(\theta_k^2-q)^2}{4}} - \frac{\theta_k^2-q}{2}\end{aligned}$$

1.1 Observation: dependence on q

... is impressive.

1.2 Restart

Restart rules:

$$
\begin{aligned}
f(x_{k+1}) &> f(x_k) \\
\langle \nabla f(y_k), x_{k+1} - x_k \rangle &> 0
\end{aligned}
$$

1.3 Linear convergence analysis

For $f(x) = \frac{1}{2}x^T A x$. And even $n = 1$, $A = \lambda$.

Suppose

$$
\begin{cases}
x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k) \\
y_{k+1} = x_{k+1} + \beta(x_{k+1} - x_k)
\end{cases}
\quad \Rightarrow \quad x_{k+1} = (1 - \frac{\lambda}{L})\left((1+\beta)x_k - \beta(x_{k-1})\right)
$$

The iteration is governed by the characteristic polynomial

$$
r^2 - (1 - \frac{\lambda}{L})\left((1+\beta)r - \beta\right)
$$

Minimizing the module of the roots $|r^*|$ gives

$$
\beta^* = \frac{1 - \sqrt{\lambda/L}}{1 + \sqrt{\lambda/L}} \quad \Rightarrow \quad |r^*| = 1 - \sqrt{\lambda/L}.
$$

For $\beta < \beta^*$ we are in the low momentum regime, and we say the system is over- damped. The convergence rate is dominated by the larger root, i.e., the system exhibits slow monotone convergence. If $\beta > \beta^*$ then the roots of the polynomial (7) are complex and we are in the high momentum regime. The system is under-damped and exhibits periodicity.

2 AZO14

From [AZO14]. They use a strongly convex distance generating function (DGF) ψ and corresponding Bregman divergence $\Delta_\psi(x, y) := \psi(x) - \psi(y) - \nabla\psi(y)(x - y)$.

We only consider the Euclidian norm. Then in our notation the considered alogorithm reads

$$
\begin{aligned}
y_0 =& z_0 = x_0 \\
y_k =& (1 - \tau_k)x_k + \tau_k z_k \\
x_{k+1} =& y_k - t_k \nabla f(y_k) \\
z_{k+1} =& z_k - \alpha_k \nabla f(y_k) \\
\tau_k =& t_k/\alpha_k \\
t_k =& 1/L, \quad \alpha_k = (k+2)/(2L)
\end{aligned}
$$

We then have

$$
z_{k+1} = z_k + \frac{\alpha_k}{t_k}(x_{k+1} - y_k) = z_k + \frac{\alpha_k}{t_k}(x_{k+1} - ((1-\tau_k)x_k + \tau_k z_k)) = x_{k+1} + \frac{1-\tau_k}{\tau_k}(x_{k+1} - x_k)
$$

and

$$
y_k = (1-\tau_k)x_k + \tau_k(x_k + \frac{1-\tau_{k-1}}{\tau_{k-1}}(x_k - x_{k-1})) = x_k + \frac{\tau_k(1-\tau_{k-1})}{\tau_{k-1}}(x_k - x_{k-1})
$$

3 DFR18

From [DFR18], inspired by [BLS15].

Lemma 1. *(Quadratic Averaging) Let* $Q_i(x) = Q_i^* + \frac{\alpha}{2}\|x - c_i\|^2$ *and* $Q(\lambda, x) = (1-\lambda)Q_1(x) + \lambda Q_2(x)$. *Then*

$$\begin{cases} \max\limits_{0\leqslant\lambda\leqslant 1} Q^*(\lambda) = (1-\lambda^*)Q_1^* + \lambda^* Q_2^* + \dfrac{\lambda^*(1-\lambda^*)\alpha}{2}\|c_1 - c_2\|^2, \\ \operatorname*{argmax}\limits_{0\leqslant\lambda\leqslant 1} Q^*(\lambda) = (1-\lambda^*)c_1 + \lambda^* c_2, \\ \lambda^* = P_{[0;1]}\left(\dfrac{1}{2} + \dfrac{(Q_2^* - Q_1^*)}{\alpha\|c_1 - c_2\|^2}\right). \end{cases} \tag{1}$$

If

$$\frac{|Q_2^* - Q_1^*|}{\alpha\|c_1 - c_2\|^2} \leqslant \frac{1}{2} \tag{2}$$

we have

$$Q^*(\lambda^*) = \frac{Q_1^* + Q_2^*}{2} + \frac{\alpha}{8}\|c_1 - c_2\|^2 + \frac{(Q_2^* - Q_1^*)^2}{2\alpha\|c_1 - c_2\|^2} \tag{3}$$

and the function $Q^*(\lambda^*)$ *is nondecreasing in* Q_i^*.

Proof. We have with $a^2 - 2ab = (a-b)^2 - b^2$

$$\begin{aligned} (1-\lambda)\|x - c_1\|^2 + \lambda\|x - c_2\|^2 &= \|x\|^2 - 2\langle x, (1-\lambda)c_1 + \lambda c_2\rangle + (1-\lambda)\|c_1\|^2 + \lambda\|c_2\|^2 \\ &= \|x - (1-\lambda)c_1 + \lambda c_2\|^2 - \|(1-\lambda)c_1 + \lambda c_2\|^2 + (1-\lambda)\|c_1\|^2 + \lambda\|c_2\|^2 \\ &= \|x - (1-\lambda)c_1 + \lambda c_2\|^2 + \lambda(1-\lambda)\|c_1 - c_2\|^2, \end{aligned}$$

so

$$\begin{aligned} Q(\lambda, x) &= (1-\lambda)Q_1^* + \lambda Q_2^* + \frac{(1-\lambda)\alpha}{2}\|x - c_1\|^2 + \frac{\lambda\alpha}{2}\|x - c_2\|^2 \\ &= (1-\lambda)Q_1^* + \lambda Q_2^* + \frac{\lambda(1-\lambda)\alpha}{2}\|c_1 - c_2\|^2 + \frac{\alpha}{2}\|x - (1-\lambda)c_1 + \lambda c_2\|^2, \end{aligned}$$

which gives

$$Q^*(\lambda) = (1-\lambda)Q_1^* + \lambda Q_2^* + \frac{\lambda(1-\lambda)\alpha}{2}\|c_1 - c_2\|^2, \quad \operatorname*{argmin}_x Q(\lambda, x) = (1-\lambda)c_1 + \lambda c_2$$

since

$$\frac{dQ^*(\lambda)}{d\lambda} = Q_2^* - Q_1^* + \frac{\alpha}{2}\|c_1 - c_2\|^2 - \lambda\alpha\|c_1 - c_2\|^2$$

we find

$$\lambda^* = P_{[0;1]}\left(\frac{1}{2} + \frac{(Q_2^* - Q_1^*)}{\alpha\|c_1 - c_2\|^2}\right).$$

If (2)

$$\begin{aligned}
\lambda^* =& \frac{1}{2} + \frac{Q_2^* - Q_1^*}{\alpha \|c_1 - c_2\|^2} \\
Q^*(\lambda^*) =& \frac{Q_1^* + Q_2^*}{2} + \frac{(Q_2^* - Q_1^*)}{\alpha \|c_1 - c_2\|^2}(Q_2^* - Q_1^*) + \frac{\alpha\left(\frac{1}{4} - \frac{|Q_2^* - Q_1^*|^2}{\alpha^2 \|c_1 - c_2\|^4}\right)}{2} \|c_1 - c_2\|^2 \\
=& \frac{Q_1^* + Q_2^*}{2} + \frac{\alpha}{8}\|c_1 - c_2\|^2 + \frac{(Q_2^* - Q_1^*)^2}{2\alpha \|c_1 - c_2\|^2} \\
=& Q_2^* + \frac{Q_1^* - Q_2^*}{2} + \frac{\alpha}{8}\|c_1 - c_2\|^2 + \frac{(Q_2^* - Q_1^*)^2}{2\alpha \|c_1 - c_2\|^2} \\
=& Q_2^* + \frac{\left((Q_2^* - Q_1^*) + \frac{\alpha}{2}\|c_1 - c_2\|^2\right)^2}{2\alpha \|c_1 - c_2\|^2}
\end{aligned}$$

Finally we have

$$\frac{\partial Q^*(\lambda^*)}{\partial Q_1^*} = \frac{1}{2} - \frac{Q_2^* - Q_1^*}{\alpha \|c_1 - c_2\|^2} \geqslant 0.$$

□

Let

$$x^+ := x - \frac{1}{L}\nabla f(x), \quad x^{++} := x - \frac{1}{\mu}\nabla f(x).$$

Algorithm 1: Quadratic averaging

Inputs: $x_0 \in X$. Set $k = 0$, $v_0 := f(x_0) - \frac{\|\nabla f(x_0)\|^2}{2\mu}$, $c_0 := x_0^{++}$,
$Q_0(x) = v_0 + \frac{\mu}{2}\|x - c_0\|^2$

(1) $x_{k+1} := \min_{0 \leqslant t \leqslant 1}(c_k + t(x_k^+ - c_k))$.

(2) $\widetilde{v} := f(x_{k+1}) - \frac{\|\nabla f(x_{k+1})\|^2}{2\mu}$, $\lambda_k := P_{[0;1]}\left(\frac{1}{2} + \frac{v_k - \widetilde{v}}{\mu\|c_k - x_{k+1}^{++}\|^2}\right)$,
$c_{k+1} := (1 - \lambda_k)x_{k+1}^{++} + \lambda_k c_k$, $v_{k+1} = (1 - \lambda_k)\widetilde{v} + \lambda_k v_k + \frac{\lambda_k(1-\lambda_k)\mu}{2}\left\|x_{k+1}^{++} - c_k\right\|^2$

(3) Increment k and go to (1).

Theorem 1. *(2.3) We have*

$$v_k \leqslant f^* \leqslant f(x_k^+), \quad f(x_k^+) - v_k \leqslant \rho^k \left(f(x_0^+) - v_0\right), \quad \rho := 1 - 1/\sqrt{\kappa}. \tag{4}$$

Proof. Let $r_k := \rho^k\left(f(x_0^+) - v_0\right)$. By induction we show $f(x_k^+) \leqslant v_k + r_k$. For $k = 0$ this evident. Let the induction hypothesis be true. We want to show

$$f(x_{k+1}^+) \leqslant v_{k+1} + r_{k+1}.$$

We have

$$\begin{aligned} f(x_{k+1}^+) \leqslant & f(x_{k+1}) - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2 && \text{(Lipschitz)} \\ \leqslant & f(x_k^+) - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2 && \text{(Line-search)} \\ \leqslant & \nu_k + r_k - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2 && \text{(Induction)} \end{aligned}$$

Now suppose that

$$\|\nabla f(x_{k+1})\|^2 \geqslant 2\sqrt{L\mu}r_k. \tag{5}$$

Then

$$\begin{aligned} f(x_{k+1}^+) \leqslant & \nu_k + \left(1 - \frac{\sqrt{\mu}}{\sqrt{L}}\right) r_k && \text{(5)} \\ \leqslant & \nu_{k+1} + r_{k+1} && (\nu_k \text{ increasing}) \end{aligned}$$

Let

$$\frac{\|\nabla f(x_{k+1})\|^2}{\mu} \leqslant 2\frac{\sqrt{\kappa}}{\sqrt{\kappa}+1} r_k \tag{6}$$

We then have

$$\begin{aligned} f(x_{k+1}^+) \leqslant & f(x_{k+1}) - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2 && \text{(Lipschitz)} \\ \leqslant & f(x_{k+1}) - \frac{1}{2\mu}\|\nabla f(x_{k+1})\|^2 + \frac{1}{2\mu}\left(1 - \frac{1}{\kappa}\right)\|\nabla f(x_{k+1})\|^2 \\ = & \tilde{\nu} + \frac{1}{2\mu}\left(1 - \frac{1}{\kappa}\right)\|\nabla f(x_{k+1})\|^2 \\ \leqslant & \nu_{k+1} + \left(1 - \frac{1}{\kappa}\right)\frac{\|\nabla f(x_{k+1})\|^2}{2\mu} && \text{(QA)} \\ \leqslant & \nu_{k+1} + \frac{\kappa - 1}{\kappa}\frac{\sqrt{\kappa}}{\sqrt{\kappa}+1} r_k && \text{(6)} \\ \leqslant & \nu_{k+1} + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}} r_k = \nu_{k+1} + r_{k+1} \end{aligned}$$

Now we suppose that (6) is false, so

$$\frac{\|\nabla f(x_{k+1})\|^2}{\mu} \geqslant 2\frac{\sqrt{\kappa}}{\sqrt{\kappa}+1} r_k \tag{7}$$

From the previous computation we have

$$\tilde{\nu} \geqslant f(x_{k+1}^+) - \frac{1}{2\mu}\left(1 - \frac{1}{\kappa}\right)\|\nabla f(x_{k+1})\|^2 =: \tilde{\nu}_A$$

We also have

$$\begin{aligned} f(x_{k+1}^+) \leqslant & f(x_{k+1}) - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2 && \text{(Lipschitz)} \\ \leqslant & f(x_k^+) - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2 && \text{(Line-search)} \\ \leqslant & \nu_k + r_k - \frac{1}{2L}\|\nabla f(x_{k+1})\|^2 && \text{(Induction)} \end{aligned}$$

such that

$$\nu_k \geqslant f(x_{k+1}^+) - r_k + \frac{1}{2\mu\kappa} \|\nabla f(x_{k+1})\|^2 =: \widehat{\nu}_B.$$

By the line-search we have (!)

$$\langle \nabla f(x_{k+1}), x_{k+1} - c_k \rangle \leqslant 0,$$

such that

$$\left\|x_{k+1}^{++} - c_k\right\|^2 = \left\|x_{k+1} - c_k - \frac{1}{\mu}\nabla f(x_{k+1})\right\|^2 \geqslant \|x_{k+1} - c_k\|^2 + \frac{1}{\mu^2} \|\nabla f(x_{k+1})\|^2 \geqslant \frac{1}{\mu^2} \|\nabla f(x_{k+1})\|^2$$

We have

$$|\widehat{\nu}_A - \widehat{\nu}_B| = \left| r_k - \frac{1}{2\mu} \|\nabla f(x_{k+1})\|^2 \right|,$$

such that

$$\frac{|\widehat{\nu}_A - \widehat{\nu}_B|}{\mu \left\|x_{k+1}^{++} - c_k\right\|^2} \leqslant \mu \frac{\left| r_k - \frac{1}{2\mu} \|\nabla f(x_{k+1})\|^2 \right|}{\|\nabla f(x_{k+1})\|^2} = \left| \frac{\mu r_k}{\|\nabla f(x_{k+1})\|^2} - \frac{1}{2} \right| \leqslant \frac{1}{2}$$

since by (7) we have $0 \leqslant \frac{\mu r_k}{\|\nabla f(x_{k+1})\|^2} \leqslant \frac{\sqrt{\kappa}+1}{2\sqrt{\kappa}} \leqslant 1$ (and $\kappa \geqslant 1$).

Then we have by Lemma 1 and $d^2 := \left\|x_{k+1}^{++} - c_k\right\|^2$ and $h^2 := \frac{\|\nabla f(x_{k+1})\|^2}{\mu}$

$$\begin{aligned}
\nu_{k+1} \geqslant& \frac{\widehat{\nu}_A + \widehat{\nu}_B}{2} + \frac{\mu}{8} \left\|x_{k+1}^{++} - c_k\right\|^2 + \frac{(\widehat{\nu}_B - \widehat{\nu}_A)^2}{2\mu \left\|x_{k+1}^{++} - c_k\right\|^2} \\
\nu =& f(x_{k+1}^+) + \frac{1}{2}\left(\frac{h^2}{\kappa} - \frac{h^2}{2} - r_k\right) + \frac{\mu}{8} d^2 + \frac{\left(r_k - \frac{h^2}{2}\right)^2}{2\mu d^2} \\
=& f(x_{k+1}^+) - r_k + \frac{h^2}{2\kappa} + \frac{\left(\frac{\mu}{2} d^2 + (r_k - \frac{h^2}{2})\right)^2}{2\mu d^2} \\
=& f(x_{k+1}^+) - r_k + \frac{h^2}{2\kappa} + \frac{\mu}{8}\left(d + \frac{2}{\mu}(r_k - \frac{h^2}{2})/d\right)^2
\end{aligned}$$

$$\begin{aligned}
f(x_{k+1}^+) - r_k + X \geqslant& f(x_{k+1}^+) - r_{k+1} = f(x_{k+1}^+) - (1 - 1\sqrt{\kappa}) r_k \\
\Leftrightarrow \quad \sqrt{\kappa} X \geqslant& r_k
\end{aligned}$$

Let $\phi(s) = s + a/s$ on $[b; +\infty[$. Then $\phi'(s) = 1 - a/s^2$, $\phi''(s) = 2a/s^3$. If $a \leqslant 0$, ϕ is strictly increasing and $\phi(s) \geqslant \phi(b) = b + a/b$. Otherwise, ϕ is strictly convex with global minimum $s = \sqrt{a}$, so $\phi(s) \geqslant 2\sqrt{a}$ if $\sqrt{a} \geqslant b$. This gives with $a = \frac{2}{\mu}(r_k - \frac{h^2}{2})$ and $b = h^2/\mu$.

If $a \leqslant 0$ we have $2\frac{\sqrt{\kappa}}{\sqrt{\kappa}+1} r \leqslant h^2 \leqslant 2r$

If $a \leqslant 0$ we have $h^2 \geqslant 2r$ and $b + a/b = h^2/\mu + 2r_k/h^2 - 1$

$$\begin{aligned}
\nu_{k+1} \geqslant& f(x_{k+1}^+) - r_k + \frac{h^2}{2\kappa} + \frac{\mu}{8}\left(h^2/\mu + 2r_k/h^2 - 1\right)^2 \\
=& f(x_{k+1}^+) - r_k +
\end{aligned}$$

$$
\begin{aligned}
-r_k + \frac{h^2}{2\kappa} + \frac{\mu}{4}\sqrt{\frac{2}{\mu}(r_k - \frac{h^2}{2})} &\geqslant -r_{k+1} = -r_k(1 - 1/\sqrt{\kappa}) \\
\Leftrightarrow \quad \frac{h^2}{2}\frac{\mu}{L} + \frac{\mu}{4}\sqrt{\frac{2}{\mu}(r_k - \frac{h^2}{2})} &\geqslant r_k/\sqrt{\kappa} = r_k\frac{\sqrt{\mu}}{\sqrt{L}} \\
\Leftrightarrow \quad \frac{h^2}{2}\frac{\sqrt{\mu}}{\sqrt{L}} + \frac{\sqrt{2L}}{4}\sqrt{(r_k - \frac{h^2}{2})} &\geqslant r_k
\end{aligned}
$$

Let $\phi(s) = as + b\sqrt{c-s}$ on $[0;c]$. Then $\phi'(s) = a - b(c-s)^{-1/2}$, $s^* = c - b^2/a^2$, $\phi(s^*) = ac - b^2/a + b^2 + a = ac$, so

$$
\frac{h^2}{2}\frac{\sqrt{\mu}}{\sqrt{L}} + \frac{\sqrt{2L}}{4}\sqrt{(r_k - \frac{h^2}{2})} \geqslant \frac{\sqrt{\mu}}{\sqrt{L}} r_k
$$

□

4 JGMTRT21

From [Jah+21], inspired by [DFR18].

Algorithm 2: Accelerated Smooth Underestimate Sequence Algorithm (ASUESA)

Inputs: $x_0 \in X$, $\varepsilon > 0$. Set $k = 0$, $v_0 := x_0^{++}$, $\phi_0^* := f(x_0^+) + \left(1 - \frac{1}{\kappa}\right)\frac{\|\nabla f(x_0)\|^2}{2\mu}$,
$\alpha_k = 1/\sqrt{\kappa}$, $\beta_k = 1/(1+\alpha_k) = \sqrt{\kappa}/(\sqrt{\kappa}+1)$

(1) $y_k := \beta_k x_k + (1-\beta_k)v_k$.

(2) $x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k)$

(3) $v_{k+1} = (1-\alpha_k)v_k + \alpha_k y_k^{++}$,
$\phi_{k+1}^* = (1-\alpha_k)\left(\phi_k^* + \frac{\alpha_k\mu}{2}\|y_k^{++} - v_k\|^2\right) + \alpha_k\left(f(y_k) - \frac{\|\nabla f(y_k)\|^2}{2\mu}\right)$

(4) If $f(x_{k+1}) - \phi_{k+1}^* \leqslant \varepsilon$: quit.

(5) Increment k and go to (1).

Theorem 2. *(Corollary 4 in [Jah+21]) We have*

$$
\phi_k^* \leqslant f^* \leqslant f(x_k^+), \quad f(x_k) - \phi_k^* \leqslant \rho^k\left(f(x_0) - \phi_0^*\right), \quad \rho := 1 - 1/\sqrt{\kappa}. \tag{8}
$$

The idea (underestimate sequence) is to show that

$$
\phi_k^* \leqslant f(x^*), \quad f(x_{k+1}) - \phi_{k+1}^* \leqslant (1-\alpha_k)\left(f(x_k) - \phi_k^*\right) \tag{9}
$$

which implies $f(x_{k+1}) - \phi_k^* \leqslant \prod_{m=0}^{k}(1-\alpha_k)\left(f(x_0) - \phi_0^*\right)$.

The sequence is constructed recursively by

$$
\begin{cases}
\phi_0(x) = \phi_0^* + \frac{\mu}{2}\|v_0\|^2, \quad \phi_{k+1} = (1-\alpha_k)\phi_k + \alpha_k\psi(x, y_k) \\
\qquad\qquad \psi(x,y) := f(y) + \langle\nabla f(y), x - y\rangle + \frac{\mu}{2}\|x-y\|^2
\end{cases} \tag{10}
$$

Lemma 2. *We have*

$$\psi(x,y) \leqslant f(x), \quad \psi(x,y) = f(y) + \frac{\mu}{2}\left\|x - y^{++}\right\|^2 - \frac{\|\nabla f(y)\|^2}{2\mu} \tag{11}$$

Remark 1. *For composite function, the authors use instead*

$$\psi(x,y) := f(y^+) + \langle \nabla f(y), x - y\rangle + \frac{\mu}{2}\|x-y\|^2 + \frac{1}{2L}\|\nabla f(y)\|^2,$$

giving

$$\psi(x,y) = f(y^+) - \left(1 - \frac{1}{\kappa}\right)\frac{\|\nabla f(y)\|^2}{2\mu} + \frac{\mu}{2}\left\|x - y^{++}\right\|^2$$

Proof.

$$f(y) \leqslant f(x) - \langle \nabla f(y), x - y\rangle - \frac{\mu}{2}\|x-y\|^2$$

which gives the first assertion. With $ab + b^2/2 = (a+b)^2/2 - a^2/2$ it follows also that

$$\begin{aligned}
\psi(x,y) =& f(y) + \langle \nabla f(y), x - y\rangle + \frac{\mu}{2}\|x-y\|^2 \\
=& f(y) + \frac{\mu}{2}\left\|x - y + \frac{1}{2\mu}\nabla f(y)\right\|^2 - \frac{1}{2\mu}\|\nabla f(y)\|^2 \\
=& f(y) + \frac{\mu}{2}\left\|x - y^{++}\right\|^2 - \frac{\|\nabla f(y)\|^2}{2\mu}
\end{aligned}$$

□

Lemma 3. *We have for $(\phi_k)_{k\in\mathbb{N}}$ defined by (10)*

$$\begin{cases}
\phi_{k+1}(x) = \phi^*_{k+1} + \frac{\mu}{2}\|x - v_{k+1}\|^2, \quad v_{k+1} = (1-\alpha_k)v_k + \alpha_k y_k^{++} \\
\phi^*_{k+1} = (1-\alpha_k)\left(\phi^*_k + \frac{\alpha_k\mu}{2}\left\|v_k - y_k^{++}\right\|^2\right) + \alpha_k\left(f(y_k) - \frac{\|\nabla f(y_k)\|^2}{2\mu}\right).
\end{cases} \tag{12}$$

Proof. By induction. $k = 0$ is trivial.

$$\begin{aligned}
\phi_{k+1}(x) =& (1-\alpha_k)\phi_k(x) + \alpha_k\psi(x,y_k) && \text{(Definition of } (\phi \\
=& (1-\alpha_k)\left(\phi^*_k + \frac{\mu}{2}\|x - v_k\|^2\right) + \alpha_k\psi(x,y_k) && \text{(Induction)} \\
=& (1-\alpha_k)\left(\phi^*_k + \frac{\mu}{2}\|x - v_k\|^2\right) + \alpha_k\left(f(y_k) - \frac{\|\nabla f(y_k)\|^2}{2\mu} + \frac{\mu}{2}\left\|x - y_k^{++}\right\|^2\right) && (11)
\end{aligned}$$

Now we have

$$(1-\alpha)\|a-b\|^2 + \alpha\|a-c\|^2 = \|a - (1-\alpha)b - \alpha c\|^2 + \alpha(1-\alpha)\|c-b\|^2 \tag{13}$$

This is true for $\alpha = 0$. Since the right hand side is equal to

$$\|a - b + \alpha(b-c)\|^2 + \alpha(1-\alpha)\|c-b\|^2 = \|a-b\|^2 + 2\alpha\langle a-b, b-c\rangle + \alpha\|c-b\|^2$$

its derivative with respect to α is

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

which equals the derivative of the left hand side of (13). □

It remains to check (9). We have

$$f(x_{k+1}) \leqslant f(y_k) - \frac{1}{2L} \|\nabla f(y_k)\|^2 \leqslant f(x_k) + \langle \nabla f(y_k), y_k - x_k \rangle - \frac{1}{2L} \|\nabla f(y_k)\|^2 \tag{14}$$

such that

$$\begin{aligned}
f(x_{k+1}) - \phi_{k+1}^* =& f(x_{k+1}) - (1-\alpha_k)\left(\phi_k^* + \frac{\alpha_k \mu}{2}\|v_k - y_k^{++}\|^2\right) - \alpha_k \left(f(y_k) - \frac{\|\nabla f(y_k)\|^2}{2\mu}\right) && (12)\\
=& (1-\alpha_k)\left(f(x_{k+1}) - \phi_k^* - \frac{\alpha_k \mu}{2}\|v_k - y_k^{++}\|^2\right) + \alpha_k \left(\frac{\|\nabla f(y_k)\|^2}{2\mu} - \frac{\|\nabla f(y_k)\|^2}{2L}\right) && (14)\\
\leqslant& (1-\alpha_k)(f(x_k) - \phi_k^*) + \alpha_k \left(\frac{\|\nabla f(y_k)\|^2}{2\mu} - \frac{\|\nabla f(y_k)\|^2}{2L}\right) - (1-\alpha_k)\frac{\|\nabla f(y_k)\|^2}{2L}\\
&+ (1-\alpha_k)\left(\langle \nabla f(y_k), y_k - x_k\rangle - \frac{\alpha_k \mu}{2}\|v_k - y_k^{++}\|^2\right)
\end{aligned}$$

By the scheme we have

$$v_k = y_k + \frac{\beta_k}{1-\beta_k}(y_k - x_k),$$

so with $\frac{\beta_k}{1-\beta_k} = \frac{1}{1+\alpha_k}\frac{1+\alpha_k}{\alpha_k} = 1/\alpha_k$

$$\begin{aligned}
\|v_k - y_k^{++}\|^2 =& \left\|\frac{\beta_k}{1-\beta_k}(y_k - x_k) + \frac{1}{\mu}\nabla f(y_k)\right\|^2\\
=& \frac{1}{\alpha_k^2}\|y_k - x_k\|^2 + \frac{2}{\mu\alpha_k}\langle y_k - x_k, \nabla f(y_k)\rangle + \frac{1}{\mu^2}\|\nabla f(y_k)\|^2
\end{aligned}$$

and

$$\langle \nabla f(y_k), y_k - x_k\rangle - \frac{\alpha_k \mu}{2}\|v_k - y_k^{++}\|^2 = -\frac{\mu}{2\alpha_k}\|y_k - x_k\|^2 - \frac{\alpha_k}{2\mu}\|\nabla f(y_k)\|^2,$$

so

$$\begin{aligned}
f(x_{k+1}) - \phi_{k+1}^* \leqslant& (1-\alpha_k)(f(x_k) - \phi_k^*) + \alpha_k \frac{\|\nabla f(y_k)\|^2}{2\mu} - \frac{\|\nabla f(y_k)\|^2}{2L}\\
&- (1-\alpha_k)\left(\frac{\mu}{2\alpha_k}\|y_k - x_k\|^2 + \frac{\alpha_k}{2\mu}\|\nabla f(y_k)\|^2\right)\\
=& (1-\alpha_k)(f(x_k) - \phi_k^*) - (1-\alpha_k)\left(\frac{\mu}{2\alpha_k}\|y_k - x_k\|^2\right)\\
&+ \left(\frac{\alpha_k}{2\mu} - \frac{(1-\alpha_k)}{2L} - \frac{\alpha_k(1-\alpha_k)}{2\mu}\right)\|\nabla f(y_k)\|^2
\end{aligned}$$

But

$$\frac{\alpha_k}{2\mu} - \frac{1}{2L} - \frac{\alpha_k(1-\alpha_k)}{2\mu} = \frac{\alpha_k^2}{2\mu} - \frac{1}{2L} = 0.$$

5 PSW21

From [PSW21].

Algorithm 3: Accelerated GM

Inputs: $x_0 \in X$, $\eta > 0$. Set $k = 0$, $x_{-1} := x_0$, $\beta = \frac{1-\eta\sqrt{s}}{1+\eta\sqrt{s}}$

(1) $y_k := x_k + \beta(x_k - x_{k-1})$.

(2) $x_{k+1} = y_k - t\nabla f(y_k)$

(5) Increment k and go to (1).

Lemma 4. *Let* $\theta = \eta\sqrt{s}$ *and*

$$v_{k+1} = x_k - \frac{1}{\theta}(x_{k+1} - x_k)$$

References

[OC15] B. O'Donoghue and E. Candès. "Adaptive restart for accelerated gradient schemes". In: *Found. Comput. Math.* 15.3 (2015), pp. 715–732.

[AZO14] Z. Allen-Zhu and L. Orecchia. *A Novel, Simple Interpretation of Nesterov's Accelerated Method as a Combination of Gradient and Mirror Descent*. ArXiv. 2014.

[DFR18] D. Drusvyatskiy, M. Fazel, and S. Roy. "An optimal first order method based on optimal quadratic averaging". In: *SIAM J. Optim.* 28.1 (2018), pp. 251–271.

[BLS15] S. Bubeck, Y. T. Lee, and M. Singh. *A geometric alternative to Nesterov's accelerated gradient descent*. ArXiv. 2015.

[Jah+21] M. Jahani, N. V. C. Gudapati, C. Ma, R. Tappenden, and M. Takáč. "Fast and safe: accelerated gradient methods with optimality certificates and underestimate sequences". In: *Comput. Optim. Appl.* 79.2 (2021), pp. 369–404.

[PSW21] J.-H. Park, A. J. Salgado, and S. M. Wise. "Preconditioned accelerated gradient descent methods for locally Lipschitz smooth objectives with applications to the solution of nonlinear PDEs". In: *J. Sci. Comput.* 89.1 (2021), Paper No. 17, 37.