Linear Coupling of Gradient and Mirror Descent: Version for Composite Functions with Adaptive Estimation of the Lipschitz Constant

Anton Rodomanov

June 1, 2016

Abstract

Recently Allen-Zhu and Orecchia [2014] have proposed a new way of deriving Nesterov's fast gradient method (FGM). They showed that FGM can be viewed as a special convex combination of the primal gradient method and mirror descent. In this work we extend the method of Allen-Zhu and Orecchia [2014] in two directions: 1) we generalize the method to the class of composite convex functions; 2) we modify the method so that it does not require the knowledge of the Lipschitz constant and is able to choose it adaptively in iterations. We prove that the proposed method retains the same convergence rate as the original method of Allen-Zhu and Orecchia [2014].

1 Notation

In what follows E denotes a finite-dimensional real vector space. The dual space which is formed by all linear functions on E is denoted by E^*. The value of a function $g \in E^*$ at $x \in E$ is denoted by $\langle g, x \rangle$. The space E is endowed with a norm $\|\cdot\|$ (which can be arbitrary). The corresponding dual norm is $\|g\|_* := \max_{x \in E}\{\langle g, x\rangle : \|x\| \leq 1\}$, $g \in E^*$. The gradient of a differentiable function f at a point x is denoted by $\nabla f(x)$. For a convex function Ψ and a point x, the symbol $\partial \Psi(x)$ denotes the subdifferential of Ψ at x and $\Psi'(x) \in \partial\Psi(x)$ stands for any subgradient.

2 Main concepts

We consider the following convex composite optimization problem Nesterov [2013]:

$$\min_{x \in Q} \, [\phi(x) := f(x) + \Psi(x)].$$

Here $Q \subseteq E$ is a closed convex set, the function f is differentiable and convex on Q, and function Ψ is closed and convex on Q (not necessarily differentiable).

In what follows we assume that f is L_f-smooth on Q:

$$\|\nabla f(x) - \nabla f(y)\|_* \leq L_f \|x - y\|, \qquad \forall x,\, y \in Q. \tag{1}$$

We stress that the constant $L_f > 0$ arises only in theoretical analysis and not in the actual implementation of the proposed method.

For mirror descent, we need to introduce the Bregman divergence. Let $\omega : Q \to \mathbb{R}$ be a distance generating function, i.e. a 1-strongly convex function on Q in the $\|\cdot\|$-norm:

$$\omega(y) \geq \omega(x) + \langle \omega'(w), y - x\rangle + \frac{1}{2}\|y - x\|^2, \qquad \forall x,\, y \in Q.$$

Then the corresponding Bregman divergence is defined as

$$V_x(y) := \omega(y) - \omega(x) - \langle \omega'(x), y - x\rangle, \qquad x,\, y \in Q.$$

Finally, we generalize the Grad and Mirr operators from Allen-Zhu and Orecchia [2014] to composite functions:

$$\mathrm{Grad}_L(x) := \operatorname*{argmin}_{y\in Q}\left\{\langle \nabla f(x), y-x\rangle + \frac{L}{2}\|y-x\|^2 + \Psi(y)\right\}, \qquad x \in Q,$$

$$\mathrm{Mirr}_z^{\alpha}(g) := \operatorname*{argmin}_{y\in Q}\left\{\langle g, y-z\rangle + \frac{1}{\alpha}V_z(y) + \Psi(y)\right\}, \qquad g \in E^*,\ z \in Q.$$

3 The method

Below is the proposed scheme of the new method. The main differences between this algorithm and the algorithm of Allen-Zhu and Orecchia [2014] are as follows: 1) now the Grad and Mirr operators contain the $\Psi(y)$ term inside; 2) now the algorithm does not require the actual Lipschitz constant L_f, instead it requires an arbitrary number L_0[1] and automatically adapts the Lipschitz constant in iterations; 3) now we need to use a different formula for α_{k+1} to guarantee convergence (see next section).

Algorithm 1 Accelerated gradient method.

```
Require: $x_0 \in Q$: initial point; $T$: number of iterations; $L_0$: initial estimate of $L_f$.
  $y_0 \leftarrow x_0,\ z_0 \leftarrow x_0,\ \alpha_0 \leftarrow 0$
  for $k = 0, \dots, T-1$ do
      $L_{k+1} \leftarrow \max\{L_0, L_k/2\}$
      while True do
          $\alpha_{k+1} \leftarrow \sqrt{\alpha_k^2 \frac{L_k}{L_{k+1}} + \frac{1}{4L_{k+1}^2}} + \frac{1}{2L_{k+1}}$, and $\tau_k \leftarrow \frac{1}{\alpha_{k+1}L_{k+1}}$.
          $x_{k+1} \leftarrow \tau_k z_k + (1-\tau_k)y_k$
          $y_{k+1} \leftarrow \mathrm{Grad}_{L_{k+1}}(x_{k+1})$
          if $f(y_{k+1}) \le f(x_{k+1}) + \langle \nabla f(x_{k+1}), y_{k+1} - x_{k+1}\rangle + \frac{L_{k+1}}{2}\|y_{k+1} - x_{k+1}\|^2$ then break
          $L_{k+1} \leftarrow 2L_{k+1}$
      end while
      $z_{k+1} \leftarrow \mathrm{Mirr}_{z_k}^{\alpha_{k+1}}(\nabla f(x_{k+1}))$
  end forreturn $y_T$
```

Note that Algorihtm 1 if well-defined in the sense that it is always guaranteed that $\tau_k \in [0,1]$ and so $x_{k+1} \in Q$ as a convex combination of points from Q. Indeed, from the formula for α_{k+1} we have

$$\alpha_{k+1}L_{k+1} \ge \left(\sqrt{\frac{1}{4L_{k+1}^2}} + \frac{1}{2L_{k+1}}\right)L_{k+1} = 1,$$

therefore $\tau_k = \frac{1}{\alpha_{k+1}L_{k+1}} \le 1$.

4 Convergence rate

First we prove the analogues of Lemma 4.2 and Lemma 4.3 from Allen-Zhu and Orecchia [2014].

Lemma 1. *For any $u \in Q$ and $\tau_k = \frac{1}{\alpha_{k+1}L_{k+1}}$ we have*

$$\begin{aligned}\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u\rangle \le\ & \alpha_{k+1}^2 L_{k+1}(\phi(x_{k+1}) - \phi(y_{k+1})) + (V_{z_k}(u) - V_{z_{k+1}}(u)) \\ & + \alpha_{k+1}\Psi(u) - (\alpha_{k+1}^2 L_{k+1})\Psi(x_{k+1}) + (\alpha_{k+1}^2 L_{k+1} - \alpha_{k+1})\Psi(y_k).\end{aligned}$$

Proof. From the first order optimality condition for $z_{k+1} = \mathrm{Mirr}_{z_k}^{\alpha_{k+1}}(\nabla f(x_{k+1}))$ we get

$$\left\langle \nabla f(x_{k+1}) + \frac{1}{\alpha_k}V'_{z_k}(z_{k+1}) + \Psi'(z_{k+1}), z_{k+1} - u\right\rangle \le 0, \qquad \forall u \in Q.$$

[1] The number L_0 can be always set to 1 with virtually no harm to the convergence rate of the method.

Therefore

$$\begin{aligned}
&\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u\rangle \\
&\qquad = \alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - z_{k+1}\rangle + \alpha_{k+1}\langle \nabla f(x_{k+1}), z_{k+1} - u\rangle \\
&\qquad \le \alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - z_{k+1}\rangle + \langle V'_{z_k}(z_{k+1}), u - z_{k+1}\rangle + \alpha_{k+1}\langle \Psi'(z_{k+1}), u - z_{k+1}\rangle \\
&\qquad \le (\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - z_{k+1}\rangle - \alpha_{k+1}\Psi(z_{k+1})) + \langle V'_{z_k}(z_{k+1}), u - z_{k+1}\rangle + \alpha_{k+1}\Psi(u),
\end{aligned}$$

where the inequality follows from the convexity of Ψ.

Using the triangle equality of the Bregman divergence, $\langle V'_x(y), u - y\rangle = V_x(u) - V_y(u) - V_x(y)$, we get

$$\begin{aligned}
\langle V'_{z_k}(z_{k+1}), u - z_{k+1}\rangle &= V_{z_k}(u) - V_{z_{k+1}}(u) - V_{z_k}(z_{k+1}) \\
&\le V_{z_k}(u) - V_{z_{k+1}}(u) - \frac{1}{2}\left\|z_{k+1} - z_k\right\|^2,
\end{aligned}$$

where we have used $V_{z_k}(z_{k+1}) \ge \frac{1}{2}\left\|z_{k+1} - z_k\right\|^2$ in the last inequality.

So we have

$$\begin{aligned}
\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u\rangle \le &\left(\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - z_{k+1}\rangle - \frac{1}{2}\left\|z_{k+1} - z_k\right\|^2 - \alpha_{k+1}\Psi(z_{k+1})\right) \\
&+ (V_{z_k}(u) - V_{z_{k+1}}(u)) + \alpha_{k+1}\Psi(u)
\end{aligned}$$

Define $v := \tau_k z_{k+1} + (1-\tau_k)y_k \in Q$. Then we have $x_{k+1} - v = \tau_k(z_k - z_{k+1})$ and $\tau_k\Psi(z_{k+1}) + (1-\tau_k)\Psi(y_k) \ge \Psi(v)$ due to convexity of Ψ. Using this and the formula for τ_k, we get

$$\begin{aligned}
&\left(\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - z_{k+1}\rangle - \frac{1}{2}\left\|z_{k+1} - z_k\right\|^2 - \Psi(z_{k+1})\right) \\
&\quad \le -\left(\frac{\alpha_{k+1}}{\tau_k}\langle \nabla f(x_{k+1}), v - x_{k+1}\rangle + \frac{1}{2\tau_k^2}\left\|v - x_{k+1}\right\|^2 + \frac{\alpha_{k+1}}{\tau_k}\Psi(v)\right) + \frac{\alpha_{k+1}(1-\tau_k)}{\tau_k}\Psi(y_k) \\
&\quad \le -(\alpha_{k+1}^2 L_{k+1})\left(\langle \nabla f(x_{k+1}), v - x_{k+1}\rangle + \frac{L_{k+1}}{2}\left\|v - x_{k+1}\right\|^2 + \Psi(v)\right) + (\alpha_{k+1}^2 L_{k+1} - \alpha_{k+1})\Psi(y_k) \\
&\quad \le -(\alpha_{k+1}^2 L_{k+1})\left(\langle \nabla f(x_{k+1}), y_{k+1} - x_{k+1}\rangle + \frac{L_{k+1}}{2}\left\|y_{k+1} - x_{k+1}\right\|^2 + \Psi(y_{k+1})\right) + (\alpha_{k+1}^2 L_{k+1} - \alpha_{k+1})\Psi(y_k)
\end{aligned}$$

Here the last inequality follows from the definition of y_{k+1}.

Note that by the termination condition for choosing L_{k+1} we have

$$\begin{aligned}
&\phi(y_{k+1}) = f(y_{k+1}) + \Psi(y_{k+1}) \\
&\qquad \le f(x_{k+1}) + \langle \nabla f(x_{k+1}), y_{k+1} - x_{k+1}\rangle + \frac{L_{k+1}}{2}\left\|y_{k+1} - x_{k+1}\right\|^2 + \Psi(y_{k+1}) \\
&\qquad = \phi(x_{k+1}) + \langle \nabla f(x_{k+1}), y_{k+1} - x_{k+1}\rangle + \frac{L_{k+1}}{2}\left\|y_{k+1} - x_{k+1}\right\|^2 + \Psi(y_{k+1}) - \Psi(x_{k+1}).
\end{aligned}$$

After rearranging:

$$-\left(\langle \nabla f(x_{k+1}), y_{k+1} - x_{k+1}\rangle + \frac{L_{k+1}}{2}\left\|y_{k+1} - x_{k+1}\right\|^2 + \Psi(y_{k+1})\right) \le \phi(x_{k+1}) - \phi(y_{k+1}) - \Psi(x_{k+1}).$$

Hence,

$$\begin{aligned}
&\left(\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - z_{k+1}\rangle - \frac{1}{2}\left\|z_{k+1} - z_k\right\|^2 - \Psi(z_{k+1})\right) \\
&\qquad \le (\alpha_{k+1}^2 L_{k+1})(\phi(x_{k+1}) - \phi(y_{k+1})) - (\alpha_{k+1}^2 L_{k+1})\Psi(x_{k+1}) + (\alpha_{k+1}^2 L_{k+1} - \alpha_{k+1})\Psi(y_k).
\end{aligned}$$

Finally, combining the previous estimates, we get

$$\begin{aligned}
\alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u\rangle \le &(\alpha_{k+1}^2 L_{k+1})(\phi(x_{k+1}) - \phi(y_{k+1})) + (V_{z_k}(u) - V_{z_{k+1}}(u)) \\
&- (\alpha_{k+1}^2 L_{k+1})\Psi(x_{k+1}) + (\alpha_{k+1}^2 L_{k+1} - \alpha_{k+1})\Psi(y_k) + \alpha_{k+1}\Psi(u).
\end{aligned}$$

□

Lemma 2. *For any $u \in Q$ and $\tau_k = \frac{1}{\alpha_{k+1} L_{k+1}}$ we have*

$$(\alpha_{k+1}^2 L_{k+1})\phi(y_{k+1}) - (\alpha_{k+1}^2 L_{k+1} - \alpha_{k+1})\phi(y_k) + (V_{z_{k+1}}(u) - V_{z_k}(u)) \leq \alpha_{k+1}\phi(u). \qquad (2)$$

Proof. Using convexity of f and relation $\tau_k(x_{k+1} - z_k) = (1 - \tau_k)(y_k - x_{k+1})$, we obtain

$$\begin{aligned}
&\alpha_{k+1}(\phi(x_{k+1}) - \phi(u)) \\
&\quad = \alpha_{k+1}(\Psi(x_{k+1}) - \Psi(u)) + \alpha_{k+1}(f(x_{k+1}) - f(u)) \\
&\quad \leq \alpha_{k+1}(\Psi(x_{k+1}) - \Psi(u)) + \alpha_{k+1}\langle \nabla f(x_{k+1}), x_{k+1} - u \rangle \\
&\quad = \alpha_{k+1}(\Psi(x_{k+1}) - \Psi(u)) + \alpha_{k+1}\langle \nabla f(x_{k+1}), x_{k+1} - z_k \rangle + \alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u \rangle \\
&\quad \leq \alpha_{k+1}(\Psi(x_{k+1}) - \Psi(u)) + \frac{\alpha_{k+1}(1 - \tau_k)}{\tau_k}\langle \nabla f(x_{k+1}), y_k - x_{k+1} \rangle + \alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u \rangle \\
&\quad \leq \alpha_{k+1}(\Psi(x_{k+1}) - \Psi(u)) + (\alpha_{k+1}^2 L_{k+1} - \alpha_{k+1})(f(y_k) - f(x_{k+1})) + \alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u \rangle \\
&\quad \leq \alpha_{k+1}\phi(x_{k+1}) - \alpha_{k+1}\Psi(u) + (\alpha_{k+1}^2 L_{k+1} - \alpha_{k+1})f(y_k) - (\alpha_{k+1}^2 L_{k+1})f(x_{k+1}) + \alpha_{k+1}\langle \nabla f(x_{k+1}), z_k - u \rangle.
\end{aligned}$$

Now we apply Lemma 1 to bound the last term, group the terms and get

$$\begin{aligned}
\alpha_{k+1}(\phi(x_{k+1}) - \phi(u)) \leq \alpha_{k+1}\phi(x_{k+1}) &- (\alpha_{k+1}^2 L_{k+1})\phi(y_{k+1}) + (\alpha_{k+1}^2 L_{k+1} - \alpha_{k+1})\phi(y_k) \\
&+ (V_{z_k}(u) - V_{z_{k+1}}(u)).
\end{aligned}$$

After rearranging, we obtain (2). □

Now we need to use Lemma 2 for obtaining the convergence rate. Note that the special choice of $\{\alpha_k\}_{k \geq 0}$ in Algorithm 1 gives us

$$\alpha_{k+1}^2 L_{k+1} - \alpha_{k+1} = \alpha_k^2 L_k, \qquad k \geq 0. \qquad (3)$$

Therefore, taking the sum over $k = 0, \ldots, T-1$ in (2) and using that $\alpha_0 = 0$, $V_{z_T}(u) \geq 0$ we get

$$(\alpha_T^2 L_T)\phi(y_T) \leq \left(\sum_{k=1}^{T} \alpha_k\right)\phi(u) + V_{z_0}(u).$$

From (3) it follows that $\sum_{k=1}^{T} \alpha_k = \alpha_T^2 L_T$, so

$$\phi(y_T) \leq \phi(u) + \frac{1}{\alpha_T^2 L_T} V_{z_0}(u). \qquad (4)$$

Now it remains to estimate the rate of growth of coefficients $A_k := \alpha_k^2 L_k$. For this we use the technique from Nesterov [2013]. Note that from (3) we have

$$A_{k+1} - A_k = \sqrt{\frac{A_{k+1}}{L_{k+1}}}$$

Rearranging and using $(a+b)^2 \leq 2a^2 + 2b^2$ and $A_k \leq A_{k+1}$, we get

$$\begin{aligned}
A_{k+1} = L_{k+1}(A_{k+1} - A_k)^2 &= L_{k+1}\left(\sqrt{A_{k+1}} + \sqrt{A_k}\right)^2\left(\sqrt{A_{k+1}} - \sqrt{A_k}\right)^2 \\
&\leq 4L_{k+1}A_{k+1}\left(\sqrt{A_{k+1}} - \sqrt{A_k}\right)^2
\end{aligned}$$

From this it follows that

$$\sqrt{A_{k+1}} \geq \frac{1}{2}\sum_{i=0}^{k} \frac{1}{\sqrt{L_i}}.$$

Note that according to (1) and the stopping criterion for choosing L_{k+1} in Algorithm (1), we always have $L_i \leq 2L_f$. Hence,

$$\sqrt{A_{k+1}} \geq \frac{k+1}{2\sqrt{2L_f}} \qquad \Longleftrightarrow \qquad A_{k+1} \geq \frac{(k+1)^2}{8L_f}. \qquad (5)$$

Thus, combining (5) and (4) with $u = x^* = \operatorname{argmin}_{x \in Q} \phi(x)$ and $V_{z_0}(x^*) =: \frac{R^2}{2}$, we have proved the following theorem:

Theorem 1. *For the sequence $\{y_k\}_{k\geq 0}$ in Algorithm 1 we have the following rate of convergence:*

$$\phi(y_T) - \phi(x^*) \leq \frac{16 L_f R^2}{T^2}.$$

Using absolutely identical arguments to Nesterov [2013], it is also possible to prove that the average number of evaluations of the function f per iteration in Algorithm 1 equals 4.

Theorem 2. *Let N_k be the total number of evaluations of the function f in Algorithm 1 after the first k iterations. Then for any $k \geq 0$ we have*

$$N_k \leq 4(k+1) + 2\log_2 \frac{L_f}{L_0}.$$

References

Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

Yu Nesterov. Gradient methods for minimizing composite functions. *Mathematical Programming*, 140(1):125–161, 2013.