Acceleration

Joon Kwon

March 2, 2017

1 Journal club: handout

1.1 Introduction

We study the Nesterov accelerated gradient method. The presentation and analysis are adapted from [AZO14] and from discussions with Roberto Cominetti and Cristobal Guzman.

- Let $f : \mathbb{R}^n \to \mathbb{R}$ be differentiable, convex, with global minimizer x_* (not necessarily unique).
- We assume f to be β-strongly smooth:

$$\|\nabla f(x') - \nabla f(x)\| \leqslant \beta \|x' - x\| .$$

We consider the following algorithm. Choose $x_1 = y_1 = z_1$ in $\mathbb{R}^n$ and iterate for $t \geqslant 1$:

$$\begin{aligned} y_{t+1} &= x_t - \frac{1}{\beta}\nabla f(x_t) \\ z_{t+1} &= z_t - \gamma_t \nabla f(x_t) \\ x_{t+1} &= \lambda_{t+1} y_{t+1} + (1-\lambda_{t+1}) z_{t+1}, \end{aligned}$$

where $\gamma_t > 0$ and $\lambda_t \in [0, 1)$.

Theorem 1.1. *For good choices (see below) of sequences (λ_t) and (γ_t), we have for all $T \geqslant 1$:*

$$f(y_{T+1}) - f(x_*) \leqslant \frac{2\beta \|x_* - x_1\|^2}{T^2}.$$

1.2 Building blocks

The first building block of the analysis is the following consequence of strong smoothness which assures that the value of the objection function f decreases by $(2\beta)^{-1} \|\nabla f(x_t)\|^2$ when a gradient step with step-size $1/\beta$ is performed.

Lemma 1.2. *For all $t \geqslant 1$,*

$$f(y_{t+1}) - f(x_t) \leqslant -\frac{1}{2\beta} \|\nabla f(x_t)\|^2 .$$

The second building block is the following *regret bound*—see e.g. [SS11, Kwo16].

Lemma 1.3 (Regret minimization)**.** *For all $x \in \mathbb{R}^n$,*

$$\sum_{t=1}^{T} \gamma_t \langle \nabla f(x_t) | z_t - x \rangle \leqslant \frac{\|x - z_1\|^2}{2} + \frac{1}{2}\sum_{t=1}^{T} \gamma_t^2 \|\nabla f(x_t)\|^2 .$$

1.3 Proof

Let $t \geqslant 1$. Using convexity inequalities and the relation $x_t - z_t = \frac{\lambda_t}{1-\lambda_t}(y_t - x_t)$ (which follows from the definition of the algorithm), we write:

$$\begin{aligned} f(x_t) - f(x_*) &\leqslant \langle \nabla f(x_t) | x_t - x_* \rangle = \langle \nabla f(x_t) | x_t - z_t \rangle + \langle \nabla f(x_t) | z_t - x_* \rangle \\ &= \frac{\lambda_t}{1-\lambda_t} \langle \nabla f(x_t) | y_t - x_t \rangle + \langle \nabla f(x_t) | z_t - x_* \rangle \\ &= \frac{\lambda_t}{1-\lambda_t} \left(f(y_t) - f(x_t) \right) + \langle \nabla f(x_t) | z_t - x_* \rangle \end{aligned}$$

We multiply on both sides by γ_t, sum over $t = 1, \ldots, T$ and apply both lemmas to get:

$$\begin{aligned} \sum_{t=1}^{T} \gamma_t (f(x_t) - f(x_*)) - \sum_{t=1}^{T} \frac{\lambda_t \gamma_t}{1-\lambda_t} (f(y_t) - f(x_t)) &\leqslant \frac{\|x_* - x_1\|^2}{2} + \sum_{t=1}^{T} \frac{\gamma_t^2}{2} \|\nabla f(x_t)\|^2 \\ &\leqslant \frac{\|x_* - x_1\|^2}{2} + \sum_{t=1}^{T} \beta \gamma_t^2 \left(f(x_t) - f(y_{t+1}) \right) \end{aligned}$$

Reorganinzing the terms, we get

$$\begin{aligned} \sum_{t=2}^{T} \underbrace{\left(\frac{\gamma_t \lambda_t}{1-\lambda_t} + \gamma_t - \beta \gamma_t^2 \right)}_{=:A_t} f(x_t) + \sum_{t=2}^{T} \underbrace{\left(\beta \gamma_{t-1}^2 - \frac{\gamma_t \lambda_t}{1-\lambda_t} \right)}_{=:B_t} f(y_t) + \underbrace{(\gamma_1 - \beta \gamma_1^2)}_{=:C} f(x_1) + \underbrace{\beta \gamma_T^2}_{=:D_T} f(y_{T+1}) \\ \leqslant \left(\sum_{t=1}^{T} \gamma_t \right) f(x_*) + \frac{\|x_* - x_1\|^2}{2}. \end{aligned}$$

We can easily check that the sum of all the coefficients A_t, B_t, C and D is equal to $\sum_{t=1}^{T} \gamma_t$. Let us divide on both sides by the latter quantity:

$$\frac{1}{\sum_{t=1}^{T} \gamma_t} \left(\sum_{t=2}^{T} A_t f(x_t) + \sum_{t=2}^{T} B_t f(y_t) + C f(x_1) + D_T f(y_{T+1}) \right) \leqslant f(x_*) + \frac{\|x_1 - x_*\|^2}{2 \left(\sum_{t=1}^{T} \gamma_t \right)}. \tag{1}$$

We want the above left-hand side to be a convex combination of different values of f. For this to be the case, we want coefficients A_t, B_t (for $t \geqslant 2$), C and D_T to all be nonnegative which is equivalent to having:

$$\gamma_1 \leqslant \frac{1}{\beta} \quad \text{and} \quad \frac{\beta \gamma_{t-1}^2}{\gamma_t} \geqslant \frac{\lambda_t}{1-\lambda_t} \geqslant \beta \gamma_t - 1.$$

Besides, we see in the right-hand side of Equation (1) that **the speed of convergence will be given by $\sum_{t=1}^{T} \gamma_t$**. We therefore want (γ_t) to grow **as fast as possible**. In other words, for a given value of γ_{t-1}, we want the highest possible value for γ_t. This is equivalent to having equality in the above two inequalities. The corresponding choice of (λ_t) and (γ_t) is summarized in the following lemma.

Lemma 1.4. *Setting $\gamma_1 = 1/\beta$, $\lambda_1 = 0$ and for $t \geqslant 1$:*

$$\gamma_{t+1} = \frac{1 + \sqrt{1 + 4\beta^2 \gamma_t^2}}{2\beta} \quad \textit{and} \quad \lambda_t = 1 - \frac{1}{\beta \gamma_t},$$

implies:

(i) $A_t = 0$ *for* $t \geqslant 2$

(ii) $B_t = 0$ *for* $t \geqslant 2$

(iii) $C = 0$

(iv) $D_T = \sum_{t=1}^{T} \gamma_t$

(v) $\sum_{t=1}^{T} \gamma_t \geqslant T^2/4\beta$

Proof. Easy. □

Equation (1) then boils down to:

$$f(y_{T+1}) - f(x_*) \leqslant \frac{2\beta \left\| x_* - x_1 \right\|^2}{T^2}.$$

2 References

References

[AZO14] Zeyuan Allen-Zhu and Lorenzo Orecchia. Linear coupling: An ultimate unification of gradient and mirror descent. *arXiv preprint arXiv:1407.1537*, 2014.

[Kwo16] Joon Kwon. *Mirror descent strategies for regret minimization and approachability.* PhD thesis, Université Pierre-et-Marie-Curie, 2016.

[SS11] Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.