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 $\beta$-strongly smooth:
  \[ \| \nabla f(x') - \nabla f(x) \| \leq \beta \| x' - x \|. \]

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

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

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

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

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

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 $t \geq 1$,

\[
f(y_{t+1}) - f(x_t) \leq -\frac{1}{2\beta} \| \nabla f(x_t) \|^2.
\]

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

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

\[
\sum_{t=1}^T \gamma_t \langle \nabla f(x_t) | z_t - x \rangle \leq \frac{\| x - z_1 \|^2}{2} + \frac{1}{2} \sum_{t=1}^T \gamma_t^2 \| \nabla f(x_t) \|^2.
\]
1.3 Proof

Let \( t \geq 1 \). 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:

\[
\begin{align*}
    f(x_t) - f(x_*) &\leq (\nabla f(x_t))|_{x_t - x_*} = \langle \nabla f(x_t), x_t - z_t \rangle + \langle \nabla f(x_t), z_t - x_* \rangle \\
    &= \frac{\lambda_t}{1 - \lambda_t} \langle \nabla f(x_t), y_t - x_t \rangle + \langle \nabla f(x_t), z_t - x_* \rangle \\
    &= \frac{\lambda_t}{1 - \lambda_t} \left( f(y_t) - f(x_t) + \langle \nabla f(x_t), z_t - x_* \rangle \right)
\end{align*}
\]

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

\[
\sum_{t=1}^T \gamma_t (f(x_t) - f(x_*)) - \sum_{t=1}^T \frac{\lambda_t \gamma_t}{1 - \lambda_t} (f(y_t) - f(x_t)) \leq \frac{\|x_* - x_1\|^2}{2} + \sum_{t=1}^T \frac{\gamma_t^2}{2} \|\nabla f(x_t)\|^2 \\
\leq \frac{\|x_* - x_1\|^2}{2} + \sum_{t=1}^T \beta \gamma_t^2 (f(x_t) - f(y_{t+1}))
\]

Reorganizing the terms, we get

\[
\sum_{t=2}^T \left( \frac{\gamma_t \lambda_t}{1 - \lambda_t} + \gamma_t - \beta \gamma_t^2 \right) f(x_t) + \sum_{t=2}^T \left( \beta \gamma_t^2 - \frac{\gamma_t \lambda_t}{1 - \lambda_t} \right) f(y_t) + \left( \frac{\gamma_t - \beta \gamma_t^2}{Y_t} \right) f(x_1) + \left( \frac{\beta \gamma_t^2}{Y_t} \right) f(y_T+1)
\]

\[
= A_t := B_t := C := D_t,
\]

\[
\leq \left( \sum_{t=1}^T \gamma_t \right) f(x_*) + \frac{\|x_* - x_1\|^2}{2}.
\]

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

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

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 \( t \geq 2 \)), \( C \) and \( D_T \) to all be nonnegative which is equivalent to having:

\[
\gamma_t \leq \frac{1}{\beta} \quad \text{and} \quad \frac{\beta \gamma_t^2}{\gamma_t} \geq \frac{\lambda_t}{1 - \lambda_t} \geq \beta \gamma_t - 1.
\]

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 \( \gamma_t \) to grow as fast as possible. In other words, for a given value of \( \gamma_t \), we want the highest possible value for \( \gamma_t \). This is equivalent to having equality in the above two inequalities. The corresponding choice of \( (\lambda_t) \) and \( (\gamma_t) \) is summarized in the following lemma.

**Lemma 1.4.** Setting \( \gamma_1 = 1/\beta, \lambda_1 = 0 \) and for \( t \geq 1 \):

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

implies:
(i) \( A_t = 0 \) for \( t \geq 2 \)

(ii) \( B_t = 0 \) for \( t \geq 2 \)

(iii) \( C = 0 \)

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

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

Proof. Easy.

Equation (1) then boils down to:

\[
f(y_{T+1}) - f(x_*) \leq 2\beta \|x_* - x_1\|^2 / T^2.
\]

2 References

References

