The University of Texas at Austin
Department of Electrical and Computer Engineering

ECE 381V: Large-Scale Optimization II — Spring 2022

Lecture 6

Caramanis & Mokhtari Monday, February 7, 2022

Goal: In this lecture, we talk about the Nesterov's accelerated gradient descent method and analyze it for both smooth convex and strongly-convex smooth settings.

1 Nesterov's Accelerated Gradient (NAG) Algorithm

The general form of NAG update is given by

$$\mathbf{y}_k = \mathbf{x}_k + \gamma_k(\mathbf{x}_k - \mathbf{x}_{k-1})$$

$$\mathbf{x}_{k+1} = \mathbf{y}_k - \eta_k \nabla f(\mathbf{y}_k)$$

where $\mathbf{x}_k$ are the main iterates and $\mathbf{y}_k$ are some intermediate iterates that we need for generating $\mathbf{x}_k$. Note that we often use the initialization $\mathbf{x}_0 = \mathbf{y}_0$.
If we simplify these two expressions and eliminate intermediate variable $\mathbf{y}$, then we obtain that NAG can be written as

$$\mathbf{x}_{k+1} = \mathbf{x}_k + \gamma_k(\mathbf{x}_k - \mathbf{x}_{k-1}) - \eta_k \nabla f(\mathbf{x}_k + \gamma(\mathbf{x}_k - \mathbf{x}_{k-1}))$$

There are two important observations here:

- The term $\gamma_k(\mathbf{x}_k - \mathbf{x}_{k-1})$ is like a momentum term that adds the drift direction to the update
- The gradient is evaluated at a point that is not even in the convex combination of $\mathbf{x}_k$ and $\mathbf{x}_{k-1}$.

2 Smooth and Strongly-Convex Setting

Theorem 1. *Let $f \in \mathcal{F}^{1,1}_{\mu,L}(\mathbb{R}^n)$ and suppose the iterates of NAG are generated according the updates*

$$\mathbf{y}_k = \mathbf{x}_k + \frac{\sqrt{\kappa}-1}{\sqrt{\kappa}+1}(\mathbf{x}_k - \mathbf{x}_{k-1})$$

$$\mathbf{x}_{k+1} = \mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k)$$

Then, the iterates of NAG satisfy

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k (f(\mathbf{x}_0) - f(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_0\|^2)$$

Consider the sequence of functions $\phi_k(\mathbf{x})$ defined as following:

$$\phi_0(\mathbf{x}) = f(\mathbf{y}_0) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}_0\|^2$$

$$\phi_{s+1}(\mathbf{x}) = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\phi_s(\mathbf{x}) + \frac{1}{\sqrt{\kappa}}\left(f(\mathbf{y}_s) + \nabla f(\mathbf{y}_s)^\top(\mathbf{x} - \mathbf{y}_s) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}_s\|^2\right)$$

We can consider $\phi_s(\mathbf{x})$ as an approximation of $f(\mathbf{x})$ that becomes better and better in the following sense:

$$\phi_k(\mathbf{x}) \leq f(\mathbf{x}) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k (\phi_0(\mathbf{x}) - f(\mathbf{x}))$$

Let's prove the above inequality by induction. Indeed, it holds for $k = 0$. Now suppose it holds for $k = j$ and we have

$$\phi_j(\mathbf{x}) \leq f(\mathbf{x}) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)^j (\phi_0(\mathbf{x}) - f(\mathbf{x}))$$

For $\phi_{j+1}(\mathbf{x})$ we have

$$\begin{aligned}
\phi_{j+1}(\mathbf{x}) &= \left(1 - \frac{1}{\sqrt{\kappa}}\right)\phi_j(\mathbf{x}) + \frac{1}{\sqrt{\kappa}}\left(f(\mathbf{y}_j) + \nabla f(\mathbf{y}_j)^\top(\mathbf{x} - \mathbf{y}_j) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}_j\|^2\right) \\
&\leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)\left(f(\mathbf{x}) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)^j (\phi_0(\mathbf{x}) - f(\mathbf{x}))\right) + \frac{1}{\sqrt{\kappa}}\left(f(\mathbf{y}_j) + \nabla f(\mathbf{y}_j)^\top(\mathbf{x} - \mathbf{y}_j) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}_j\|^2\right) \\
&\leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)\left(f(\mathbf{x}) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)^j (\phi_0(\mathbf{x}) - f(\mathbf{x}))\right) + \frac{1}{\sqrt{\kappa}} f(\mathbf{x}) \\
&\leq f(\mathbf{x}) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)^{j+1} (\phi_0(\mathbf{x}) - f(\mathbf{x}))
\end{aligned}$$

and the proof is complete.
This result shows that as we run the iterates, our function approximation becomes more accurate. Now we proceed to show that $f(\mathbf{x}_k) \leq \min \phi_k(\mathbf{x}) := \phi_k^*$. We prove this by induction. Indeed, it holds for time 0 as $\mathbf{x}_0 = \mathbf{y}_0$. Suppose it holds for k and we have $f(\mathbf{x}_k) \leq \min \phi_k(\mathbf{x}) := \phi_k^*$. Then,

$$\begin{aligned}
f(\mathbf{x}_{k+1}) &\leq f(\mathbf{y}_k) - \frac{1}{2L}\|\nabla f(\mathbf{y}_k)\|^2 \\
&= \left(1 - \frac{1}{\sqrt{\kappa}}\right) f(\mathbf{x}_k) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)(f(\mathbf{y}_k) - f(\mathbf{x}_k)) + \frac{1}{\sqrt{\kappa}} f(\mathbf{y}_k) - \frac{1}{2L}\|\nabla f(\mathbf{y}_k)\|^2 \\
&\leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)\phi_k^* + \left(1 - \frac{1}{\sqrt{\kappa}}\right)(f(\mathbf{y}_k) - f(\mathbf{x}_k)) + \frac{1}{\sqrt{\kappa}} f(\mathbf{y}_k) - \frac{1}{2L}\|\nabla f(\mathbf{y}_k)\|^2 \\
&\leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)\phi_k^* + \left(1 - \frac{1}{\sqrt{\kappa}}\right)(\nabla f(\mathbf{y}_k)^\top(\mathbf{y}_k - \mathbf{x}_k)) + \frac{1}{\sqrt{\kappa}} f(\mathbf{y}_k) - \frac{1}{2L}\|\nabla f(\mathbf{y}_k)\|^2
\end{aligned}$$

Now we establish a lower bound for ϕ_{k+1}^*. To do so, first note that the Hessian of $\phi_{k+1}(\mathbf{x})$ satisfies

$$\nabla^2 \phi_{k+1}(\mathbf{x}) = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla^2 \phi_k(\mathbf{x}) + \mu\frac{1}{\sqrt{\kappa}}\mathbf{I}$$

and hence by induction we have

$$\nabla^2 \phi_k(\mathbf{x}) = \mu \mathbf{I}$$

Hence, we have

$$\phi_k(\mathbf{x}) = \phi_k^* + \frac{\mu}{2}\|\mathbf{x} - \mathbf{v}_k\|^2$$

for some $\mathbf{v}_k$. Now using the fact that

$$\phi_{s+1}(\mathbf{x}) = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\phi_s(\mathbf{x}) + \frac{1}{\sqrt{\kappa}}\left(f(\mathbf{y}_s) + \nabla f(\mathbf{y}_s)^\top(\mathbf{x} - \mathbf{y}_s) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}_s\|^2\right)$$

we have

$$\nabla \phi_{k+1}(\mathbf{x}) = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla \phi_k(\mathbf{x}) + \frac{1}{\sqrt{\kappa}}\left(\nabla f(\mathbf{y}_k) + \mu(\mathbf{x} - \mathbf{y}_k)\right)$$

which implies that

$$\nabla \phi_{k+1}(\mathbf{x}) = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\mu(\mathbf{x} - \mathbf{v}_k) + \frac{1}{\sqrt{\kappa}}\left(\nabla f(\mathbf{y}_k) + \mu(\mathbf{x} - \mathbf{y}_k)\right)$$

Since $\nabla \phi_{k+1}(\mathbf{v}_{k+1}) = \mathbf{0}$ we have

$$\mathbf{0} = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\mu(\mathbf{v}_{k+1} - \mathbf{v}_k) + \frac{1}{\sqrt{\kappa}}\left(\nabla f(\mathbf{y}_k) + \mu(\mathbf{v}_{k+1} - \mathbf{y}_k)\right)$$

which implies that

$$\mathbf{v}_{k+1} = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\mathbf{v}_k - \frac{1}{\mu\sqrt{\kappa}}\nabla f(\mathbf{y}_k) + \frac{1}{\sqrt{\kappa}}\mathbf{y}_k$$

Hence, we can write

$$\begin{aligned}
\phi_{k+1}^* + \frac{\mu}{2}\|\mathbf{x} - \mathbf{v}_{k+1}\|^2 &= \phi_{k+1}(\mathbf{x}) \\
&= \left(1 - \frac{1}{\sqrt{\kappa}}\right)\phi_k(\mathbf{x}) + \frac{1}{\sqrt{\kappa}}\left(f(\mathbf{y}_k) + \nabla f(\mathbf{y}_k)^\top(\mathbf{x} - \mathbf{y}_k) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}_k\|^2\right) \\
&= \left(1 - \frac{1}{\sqrt{\kappa}}\right)(\phi_k^* + \frac{\mu}{2}\|\mathbf{x} - \mathbf{v}_k\|^2) + \frac{1}{\sqrt{\kappa}}\left(f(\mathbf{y}_k) + \nabla f(\mathbf{y}_k)^\top(\mathbf{x} - \mathbf{y}_k) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}_k\|^2\right)
\end{aligned}$$

Now set $\mathbf{x} = \mathbf{y}_k$ to obtain

$$\phi_{k+1}^* + \frac{\mu}{2}\|\mathbf{y}_k - \mathbf{v}_{k+1}\|^2 = \left(1 - \frac{1}{\sqrt{\kappa}}\right)(\phi_k^* + \frac{\mu}{2}\|\mathbf{y}_k - \mathbf{v}_k\|^2) + \frac{1}{\sqrt{\kappa}}f(\mathbf{y}_k)$$

Now using the expression for $\mathbf{v}_{k+1}$ we have

$$\|\mathbf{y}_k - \mathbf{v}_{k+1}\|^2 = \left(1 - \frac{1}{\sqrt{\kappa}}\right)^2\|\mathbf{v}_k - \mathbf{y}_k\|^2 - \frac{2}{\mu\sqrt{\kappa}}\left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla f(\mathbf{y}_k)^\top(\mathbf{v}_k - \mathbf{y}_k) + \frac{1}{\mu^2\kappa}\|\nabla f(\mathbf{y}_k)\|^2$$

which implies that

$$\begin{aligned}
\phi_{k+1}^* = &\left(1 - \frac{1}{\sqrt{\kappa}}\right)\phi_k^* + \left(1 - \frac{1}{\sqrt{\kappa}}\right)\left(\frac{1}{\sqrt{\kappa}}\right)\frac{\mu}{2}\|\mathbf{y}_k - \mathbf{v}_k\|^2 - \frac{1}{2L}\|\nabla f(\mathbf{y}_k)\|^2 \\
&+ \frac{1}{\sqrt{\kappa}}f(\mathbf{y}_k) + \frac{1}{\sqrt{\kappa}}\left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla f(\mathbf{y}_k)^\top(\mathbf{v}_k - \mathbf{y}_k)
\end{aligned}$$

Note that by induction we can show that $\mathbf{v}_k - \mathbf{y}_k = \sqrt{\kappa}(\mathbf{y}_k - \mathbf{x}_k)$ and hence we have

$$\phi^*_{k+1} = \left(1 - \frac{1}{\sqrt{\kappa}}\right)\phi^*_k + \left(1 - \frac{1}{\sqrt{\kappa}}\right)\frac{\sqrt{\kappa}\mu}{2}\|\mathbf{y}_k - \mathbf{x}_k\|^2 - \frac{1}{2L}\|\nabla f(\mathbf{y}_k)\|^2 + \frac{1}{\sqrt{\kappa}}f(\mathbf{y}_k) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)\nabla f(\mathbf{y}_k)^\top(\mathbf{y}_k - \mathbf{x}_k)$$

Hence,

$$f(\mathbf{x}_{k+1}) \leq \phi^*_{k+1}$$

and the proof of $f(\mathbf{x}_k) \leq \phi^*_k$ is complete by induction.
Now if in ?? we set $\mathbf{x} = \mathbf{x}^*$ then we have

$$\phi_k(\mathbf{x}^*) \leq f(\mathbf{x}^*) + \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k (\phi_0(\mathbf{x}^*) - f(\mathbf{x}^*))$$

which leads to

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \phi_k(\mathbf{x}^*) - f(\mathbf{x}^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k (\phi_0(\mathbf{x}^*) - f(\mathbf{x}^*))$$

Leading to

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \left(1 - \frac{1}{\sqrt{\kappa}}\right)^k \left(f(\mathbf{x}_0) - f(\mathbf{x}^*) + \frac{\mu}{2}\|\mathbf{x}^* - \mathbf{x}_0\|^2\right)$$

3 Smooth and convex setting

Theorem 2. *Let $f \in \mathcal{F}^{1,1}_L(\mathbb{R}^n)$ and suppose the iterates of NAG are generated according the updates*

$$\mathbf{x}_{k+1} = \mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k)$$

$$\mathbf{y}_{k+1} = \mathbf{x}_{k+1} + \gamma_k(\mathbf{x}_{k+1} - \mathbf{x}_k)$$

and further suppose $\mathbf{y}_0 = \mathbf{x}_0$ and we have

$$\lambda_0 = 0, \qquad \lambda^2_{k-1} = \lambda^2_k - \lambda_k, \qquad \gamma_k = \frac{\lambda_k - 1}{\lambda_{k+1}}$$

Then, the iterates of NAG satisfy

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{L}{2}\frac{\|\mathbf{x}_0 - \mathbf{x}^*\|^2}{k^2}$$

We start the proof by exploiting the smoothness property:

$$\begin{aligned}
f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) &= f\left(\mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k)\right) - f(\mathbf{x}_k) \\
&= f\left(\mathbf{y}_k - \frac{1}{L}\nabla f(\mathbf{y}_k)\right) - f(\mathbf{y}_k) - \nabla f(\mathbf{y}_k)^\top(\mathbf{x}_k - \mathbf{y}_k) \\
&\leq -\frac{1}{2L}\|\nabla f(\mathbf{y}_k)\|^2 - \nabla f(\mathbf{y}_k)^\top(\mathbf{x}_k - \mathbf{y}_k) \\
&= \nabla f(\mathbf{y}_k)^\top(\mathbf{y}_k - \mathbf{x}_k) - \frac{1}{2L}\|\nabla f(\mathbf{y}_k)\|^2 \\
&= -\frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{y}_k\|^2 - L(\mathbf{x}_{k+1} - \mathbf{y}_k)^\top(\mathbf{y}_k - \mathbf{x}_k)
\end{aligned}$$

Hence we have

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}_k) \leq -\frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{y}_k\|^2 - L(\mathbf{x}_{k+1} - \mathbf{y}_k)^\top(\mathbf{y}_k - \mathbf{x}_k)$$

Using a similar argument we can show that

$$f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) \leq -\frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{y}_k\|^2 - L(\mathbf{x}_{k+1} - \mathbf{y}_k)^\top(\mathbf{y}_k - \mathbf{x}^*)$$

Now if we multiply the first inequality by $\lambda_k - 1$ and add to the second one use the notation $\delta_k = f(\mathbf{x}_k) - f(\mathbf{x}^*)$ we have

$$\lambda_k\delta_{k+1} - (\lambda_k - 1)\delta_k \leq -\lambda_k\frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{y}_k\|^2 - L(\mathbf{x}_{k+1} - \mathbf{y}_k)^\top(\lambda\mathbf{y}_k - (\lambda - 1)\mathbf{x}_k - \mathbf{x}^*)$$

Multiplying both sides by λ_k and using the fact that $\lambda_{k-1}^2 = \lambda_k^2 - \lambda_k$ we obtain

$$\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \leq -\frac{L}{2}\|\lambda_k(\mathbf{x}_{k+1} - \mathbf{y}_k)\|^2 - L\lambda_k(\mathbf{x}_{k+1} - \mathbf{y}_k)^\top(\lambda_k\mathbf{y}_k - (\lambda_k - 1)\mathbf{x}_k - \mathbf{x}^*)$$

Now using the fact that $2\mathbf{a}^\top\mathbf{b} - \|\mathbf{a}\|^2 = \|\mathbf{b}\|^2 - \|\mathbf{b} - \mathbf{a}\|^2$ we have

$$\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \leq -\frac{L}{2}\left(\|\lambda_k\mathbf{x}_{k+1} - (\lambda_k - 1)\mathbf{x}_k - \mathbf{x}^*\|^2 - \|\lambda_k\mathbf{y}_k - (\lambda_k - 1)\mathbf{x}_k - \mathbf{x}^*\|^2\right)$$

Now using the fact that $\mathbf{y}_{k+1} = \mathbf{x}_{k+1} + \gamma_{k+1}(\mathbf{x}_{k+1} - \mathbf{x}_k)$ we have

$$\lambda_{k+1}\mathbf{y}_{k+1} - (\lambda_{k+1} - 1)\mathbf{x}_{k+1} = \lambda_k\mathbf{x}_{k+1} - (\lambda_k - 1)\mathbf{x}_k$$

Now if we define $\mathbf{u}_k := \lambda_k\mathbf{y}_k - (\lambda_k - 1)\mathbf{x}_k - \mathbf{x}^*$ then we have

$$\lambda_k^2\delta_{k+1} - \lambda_{k-1}^2\delta_k \leq \frac{L}{2}\left(\|\mathbf{u}_k\|^2 - \|\mathbf{u}_{k+1}\|^2\right)$$

Summing up we obtain

$$\delta_k \leq \frac{L}{2}\frac{\|\mathbf{u}_0\|^2}{\lambda_{k-1}^2}$$

By induction one can show that $\lambda_{k-1} \geq k/2$ and the proof is complete.

Remark 1. *As we observe, for both considered settings, the NAG method achieves the optimal iteration complexity (order-wise), and for that reason, it is also known as the optimal first-order algorithm.*