Nesterov Accelerated Gradient Descent

$d_0 = 0$

$\lambda_0 = 1$

for $t = 0, \dots, T-1$

$$\begin{cases} \beta_t = \theta_t + d_t \\ \theta_{t+1} = \beta_t - \frac{1}{L} \nabla F(\beta_t) \end{cases}$$

λ_{t+1} = largest solution of $\lambda_{t+1}^2 - \lambda_{t+1} = \lambda_t^2$

$d_{t+1} = \dfrac{\lambda_t - 1}{\lambda_{t+1}} (\theta_{t+1} - \theta_t)$ "momentum term" $\dfrac{\lambda_t - 1}{\lambda_{t+1}} \lessapprox 1$

return θ_T.

by def$^\circ$ of ξ_t

Remark: $\lambda_t \gtrapprox \frac{t}{2} + 1$.

Let $\xi_t = \lambda_t - \lambda_{t-1} \geq 0$ and observe that $\lambda_t^2 - \lambda_{t-1}^2 = \xi_t (2\lambda_t - \xi_t)$

by def$^\circ$ of (λ_t) $= \lambda_t$

Check that $\forall t$ $\lambda_t \geq 0$ and $\lambda_t \nearrow$, so that $\xi_t \geq 0$ (indeed $\lambda_{t+1}^2 = \lambda_t^2 + \lambda_{t+1}$) so $\lambda_{t+1} \geq \lambda_t$.

Therefore, $1/2 \leq \xi_t = \dfrac{1}{2 - \xi_t/\lambda_t} \leq 1$

$\Rightarrow$ $1/2 \leq \lambda_t - \lambda_{t-1} \leq 1$ $\Rightarrow$ $\frac{1}{2} + \lambda_{t-1} \leq \lambda_t \leq 1 + \lambda_{t-1}$

By recurrence, $\frac{t}{2} + \underset{\lambda_0}{1} \leq \lambda_t \leq t + \underset{\lambda_0}{1}$ $\square$

$\xi_t \leq \dfrac{1}{2 - 1/(1+t/2)} \leq \dfrac{1}{2} + \dfrac{1}{t+1}$ $\qquad 1 + \frac{t}{2} \leq \lambda_t \leq \frac{t}{2} + \log(t+1) + 1$.

Exercise [Nesterov's acceleration]

Show that when F is cvx and L-smooth, the Nesterov's method satisfies

$$F(\theta_T) - F^* \leqslant \frac{2L \|\theta_0 - \theta^*\|_2^2}{T^2}$$

Solution: Call $\delta_t := F(\theta_t) - F^*$.

Nesterov iterates:
$$\begin{cases} \beta_k = \theta_k + (1-\alpha_k)(\theta_k - \theta_{k-1}) = \theta_k + d_k \\ \theta_{k+1} = \beta_k - \gamma \nabla F(\beta_k) = \theta_k + d_k - \gamma \nabla F(\theta_k + d_k) \end{cases}$$

$$\delta_{t+1} - \delta_t = F(\theta_{t+1}) - F(\theta_t)$$

$$= F(\theta_{t+1}) - F(\theta_t + \underbrace{(1-\alpha_t)(\theta_t - \theta_{t-1})}_{=:d_t}) + F(\theta_t + \underbrace{(1-\alpha_t)(\theta_t - \theta_{t-1})}_{=:d_t}) - F(\theta_t)$$

$$\leqslant \langle \nabla F(\theta_t + d_t), \theta_{t+1} - \theta_t - d_t \rangle + \frac{L}{2} \|\theta_{t+1} - \theta_t - d_t\|_2^2 + \text{—}//\text{—}$$

$$\leqslant \langle \nabla F(\theta_t + d_t), -\gamma \nabla F(\theta_t + d_t) \rangle + \frac{L}{2}\gamma^2 \|\nabla F(\theta_t + d_t)\|_2^2 + \text{—}//\text{—}$$

$$= \left(\frac{L\gamma^2}{2} - \gamma\right) \|\nabla F(\theta_t + d_t)\|_2^2 + F(\theta_t + d_t) - F(\theta_t)$$

$$F(\theta_t) \geqslant F(\theta_t + d_t) + \langle \nabla F(\theta_t + d_t), -d_t \rangle$$

$$\Rightarrow \langle \nabla F(\theta_t + d_t), d_t \rangle \geqslant F(\theta_t + d_t) - F(\theta_t)$$

$$\leqslant -\frac{1}{2L}\|\nabla F(\theta_t+d_t)\|_2^2 + \langle \nabla F(\theta_t+d_t), d_t\rangle$$

by convexity
+ $\gamma = 1/L$

$$= -\frac{L}{2}\left\|\frac{1}{L}\nabla F(\theta_t+d_t)\right\|_2^2 + \frac{L}{2}\cdot\left\langle \frac{2}{L}\nabla F(\theta_t+d_t), d_t\right\rangle$$

$$= -\frac{L}{2}\left(\|g_t\|_2^2 - 2\langle g_t, d_t\rangle\right) \qquad \text{with } g_t := \frac{1}{L}\nabla F(\theta_t+d_t)$$

$F(\theta^*) \geqslant F(\theta_t+d_t) + \langle \nabla F(\theta_t+d_t), \theta^* - \theta_t - d_t\rangle$

Similarly,

$$\delta_{t+1} = F(\theta_{t+1}) - F(\theta^*) = \underbrace{F(\theta_{t+1}) - F(\theta_t+d_t)} + \underbrace{F(\theta_t+d_t) - F(\theta^*)}$$

$$\leqslant -\frac{1}{2L}\|\nabla F(\theta_t+d_t)\|_2^2 + \langle \nabla F(\theta_t+d_t), \theta_t+d_t-\theta^*\rangle$$

$$= -\frac{L}{2}\left(\|g_t\|_2^2 - 2\langle g_t, \theta_t+d_t-\theta^*\rangle\right)$$

Hence,

$$(\lambda_t - 1)(\delta_{t+1} - \delta_t) + \delta_{t+1}$$

$$\leqslant (\lambda_t \cancel{-1})\left(-\frac{L}{2}\right)\left(\|g_t\|_2^2 - 2\langle g_t, d_t\rangle\right) - \frac{L}{2}\|g_t\|_2^2 + \frac{L}{2}\,2\langle g_t, \theta_t+\cancel{d_t}-\theta^*\rangle$$

$$= -\lambda_t\frac{L}{2}\|g_t\|_2^2 + \lambda_t\frac{L}{2}\cdot 2\langle g_t, d_t\rangle + \frac{L}{2}\|g_t\|_2^2 - \frac{L}{2}\cdot 2\langle g_t, d_t\rangle$$
$$- \frac{L}{2}\|g_t\|_2^2 + \frac{L}{2}\cdot 2\langle g_t, d_t\rangle + \frac{L}{2}\cdot 2\langle g_t, \theta_t-\theta^*\rangle$$

$$= -\frac{L}{2}\left(\lambda_t\|g_t\|_2^2 - 2\lambda_t\langle g_t, d_t\rangle - 2\langle g_t, \theta_t-\theta^*\rangle\right)$$

$$= -\frac{L}{2}\left(\lambda_t\|g_t\|_2^2 - 2\langle g_t, \theta_t+\lambda_t d_t-\theta^*\rangle\right)$$

$$\leqslant -\frac{L}{2}\left(\lambda_t \|g_t\|_2^2 - 2\langle g_t, \theta_t + \lambda_t d_t - \theta^*\rangle\right)$$

$$= -\frac{L}{2\lambda_t}\left(\|\lambda_t g_t\|_2^2 - 2\langle \lambda_t g_t, \theta_t + \lambda_t d_t - \theta^*\rangle\right)$$

$$= -\frac{L}{2\lambda_t}\left(\|-\lambda_t g_t + \theta_t + \lambda_t d_t - \theta^*\|^2 - \|\theta_t + \lambda_t d_t - \theta^*\|^2\right)$$

$$= -\frac{L}{2\lambda_t}\left(\|-\lambda_t g_t - \theta^* + \theta_t + \lambda_t d_t\|_2^2 - \|\theta_t + \lambda_t d_t - \theta^*\|_2^2\right)$$

$$= -\frac{L}{2\lambda_t}\left(\|\theta_{t+1} + \lambda_{t+1} d_{t+1} - \theta^*\|_2^2 - \|\theta_t + \lambda_t d_t - \theta^*\|_2^2\right)$$

Since the choice of the momentum intensity is precisely ensuring that

$$\begin{aligned}
\theta_t - \lambda_t g_t + \lambda_t d_t &= \theta_{t+1} + (\lambda_t - 1)(g_t + d_t) \\
&= \theta_{t+1} + (\lambda_t - 1)(\theta_{t+1} - \theta_t) \\
&= \theta_{t+1} + \lambda_{t+1} \cdot \underbrace{\frac{\lambda_t - 1}{\lambda_{t+1}}(\theta_{t+1} - \theta_t)}_{d_{t+1}}
\end{aligned}$$

It follows from the choice of λ_t that:

$$\lambda_t^2 \delta_{t+1} - \lambda_{t-1}^2 \delta_t = \lambda_t^2 \delta_{t+1} - (\lambda_t^2 - \lambda_t)\delta_t$$

$$\leqslant -\frac{L}{2}\left(\|\theta_{t+1} + \lambda_{t+1} d_{t+1} - \theta^*\|_2^2 - \|\theta_t + \lambda_t d_t - \theta^*\|_2^2\right)$$

and hence since $\lambda_{-1} = 0$ and $\lambda_t \geq \frac{t+1}{2}$

$$\left(\frac{T}{2}\right)^2 \delta_T \leqslant \lambda_{T-1}^2 \delta_T \leqslant \frac{L}{2} \| \theta_0 + \lambda_0 d_0 - \theta^* \|_2^2 = \frac{L \| \theta_0 - \theta^* \|_2^2}{2} .$$