Chapter 2

Training Algorithms

In mathematics, the art of proposing a question must be held of higher value than solving it.

– Georg Cantor

THIS chapter will provide background on the training algorithm for neural networks, including several topics in applications. Since the loss function is generally non-convex, the theoretical studies are more qualitative than quantitative.

2.1 Gradient Descent Method

The training of neural networks relies on optimization techniques. Gradient descent methods iteratively adjust the parameters to minimize the loss function.

Let the loss function be $\mathcal{L}(\theta)$. In the ideal continuum regime, the gradient flow (with infinitesimal learning rate) shows

$$\frac{d}{dt}\theta = -\nabla\mathcal{L}(\theta), \quad \theta(0) = \theta_0.$$

Therefore (if the ODE solves uniquely),

$$\frac{d}{dt}\mathcal{L}(\theta) = -\nabla\mathcal{L}(\theta) \cdot \frac{d}{dt}\theta = -\|\nabla\mathcal{L}(\theta)\|^2 \leq 0.$$

It shows that, as long as $\nabla\mathcal{L} \neq 0$, the gradient flow decreases the loss function. In the discrete setting, the gradient descent method becomes

$$\theta_{k+1} = \theta_k - \eta\nabla\mathcal{L}(\theta_k).$$

The parameter η denotes the learning rate (i.e., the time step size). It need not be constant. For simplicity, let us assume it is for now.

2.1.1 Convexity

Definition 2.1.1. *A function f is convex on domain Ω if for any $a \in [0, 1]$,*

$$af(x) + (1-a)f(y) \leq f(ax + (1-a)y), \quad \forall x, y \in \Omega.$$

If f is convex and differentiable, then (expand the RHS at $a \approx 0$), we have

$$f(x) \geq \nabla f(y) \cdot (x - y) + f(y).$$

It means the function's graph is above any tangent planes.

2.1.2 L-smoothness

Definition 2.1.2. *If $\mathcal{L}$ is differentiable and satisfies*

$$\langle\nabla\mathcal{L}(x) - \nabla\mathcal{L}(y), x - y\rangle \leq L\|x - y\|_2^2,$$

then $\nabla\mathcal{L}$ is L-Lipschitz continuous or $\mathcal{L}$ is L-smooth.

Theorem 2.1.3. *If $\mathcal{L} \in C^1$ and $\nabla \mathcal{L}$ is L-Lipschitz, then*

$$\mathcal{L}(\theta') \leq \mathcal{L}(\theta) + \langle \nabla \mathcal{L}(\theta), (\theta' - \theta) \rangle + \frac{L}{2} \|\theta' - \theta\|^2$$

Question 2.1.4. *Prove the above theorem.*

The next theorem is a useful argument.

Theorem 2.1.5 (Co-Coercivity). *If $\mathcal{L}$ is convex and $\nabla \mathcal{L}$ is L-Lipschitz, then*

$$\langle \nabla \mathcal{L}(x) - \nabla \mathcal{L}(y), x - y \rangle \geq \frac{1}{L} \|\nabla \mathcal{L}(x) - \nabla \mathcal{L}(y)\|^2.$$

Proof. Use Theorem 2.1.3,

$$\begin{aligned} \inf_{\theta'} \mathcal{L}(\theta') &\leq \inf_{\theta'} \mathcal{L}(\theta) + \langle \nabla \mathcal{L}(\theta), (\theta' - \theta) \rangle + \frac{L}{2} \|\theta' - \theta\|^2 \\ &= \mathcal{L}(\theta) - \frac{1}{2L} \|\nabla \mathcal{L}(\theta)\|^2. \end{aligned}$$

The above inequality only uses the L-Lipschitz property of the gradient. Let

$$g(\theta) := \mathcal{L}(\theta) - \langle \nabla \mathcal{L}(\theta'), \theta - \theta' \rangle \geq \mathcal{L}(\theta').$$

Now, ∇g is L-Lipschitz, we use the convexity (optimality holds at $\theta = \theta'$), then use the above inequality,

$$\begin{aligned} \mathcal{L}(\theta') = \inf_{\theta} g(\theta) &\leq g(\theta) - \frac{1}{2L} \|\nabla g(\theta)\|^2 \\ &= \mathcal{L}(\theta) - \langle \nabla \mathcal{L}(\theta'), \theta - \theta' \rangle - \frac{1}{2L} \|\nabla \mathcal{L}(\theta) - \nabla \mathcal{L}(\theta')\|^2. \end{aligned}$$

Switch θ and θ', then combine. □

Problem 2.1.6. *Let $\mathcal{L}$ be convex and L-smooth. Suppose θ^* is the global minimizer. Prove that*

1. $\frac{1}{L} \|\nabla \mathcal{L}(\theta_k)\|^2 \leq \langle \theta_k - \theta^*, \nabla \mathcal{L}(\theta_k) \rangle$.

Continue on next page ⇓

Continue from last page ⇑

2. $\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*) + \frac{1}{2L}\|\nabla\mathcal{L}(\theta_k)\|^2 \le \langle \theta_k - \theta^*, \nabla\mathcal{L}(\theta_k)\rangle$.
3. $\langle \theta_k - \theta^*, \nabla\mathcal{L}(\theta_k)\rangle \le \mathcal{L}(\theta_k) - \mathcal{L}(\theta^*) + \frac{L}{2}\|\theta_k - \theta^*\|^2$
4. $\mathcal{L}(\theta_k) - \mathcal{L}(\theta_{k-1}) + \frac{1}{2L}\|\nabla\mathcal{L}(\theta_k) - \nabla\mathcal{L}(\theta_{k-1})\|^2 \le \langle \theta_k - \theta_{k-1}, \nabla\mathcal{L}(\theta_k)\rangle$.
5. $\mathcal{L}(\theta_k) - \mathcal{L}(\theta_{k-1}) \le \langle \theta_k - \theta_{k-1}, \nabla\mathcal{L}(\theta_k)\rangle$.

Which statements use convexity, and which statements use L-smoothness?

Problem 2.1.7. *Find out the continuous version of Problem 2.1.6.*

Based on the Theorem 2.1.3, we find that

$$\begin{aligned}\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta_k) &\le \langle \nabla\mathcal{L}(\theta_k), \theta_{k+1} - \theta_k\rangle + \frac{L}{2}\|\theta_{k+1} - \theta_k\|_2^2 \\ &= -\eta\|\nabla\mathcal{L}(\theta_k)\|^2 + \frac{L\eta^2}{2}\|\nabla\mathcal{L}(\theta_k)\|^2 = \left(\frac{L}{2}\eta - 1\right)\eta\|\nabla\mathcal{L}(\theta_k)\|^2.\end{aligned}$$

Therefore, we need $\eta < \frac{2}{L}$ to make the loss decay. The decay on the right-hand side is maximized when $\eta = \frac{1}{L}$. In general, we only consider $\eta \le \frac{1}{L}$ for the following reason.

Theorem 2.1.8. *Let $\mathcal{L}$ be convex and $\nabla\mathcal{L}$ be L-Lipschitz. If θ^* is the global minimizer, then*

$$\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*) \le \frac{1}{2\eta}\left(\|\theta_k - \theta^*\|^2 - \|\theta_{k+1} - \theta^*\|^2\right) + \frac{\eta - \frac{1}{L}}{2}\|\nabla\mathcal{L}(\theta_k)\|^2.$$

Proof. Combine

$$\|\theta_{k+1} - \theta^*\|^2 = \|\theta_k - \theta^*\|^2 + \eta^2\|\nabla\mathcal{L}(\theta_k)\|^2 - 2\eta\langle\nabla\mathcal{L}(\theta_k), \theta_k - \theta^*\rangle$$

with co-coercivity (see last equation of Theorem 2.1.5),

$$\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*) + \frac{1}{2L}\|\nabla\mathcal{L}(\theta_k)\|^2 \le \langle\nabla\mathcal{L}(\theta_k), \theta_k - \theta^*\rangle.$$

Then,

$$\|\theta_{k+1} - \theta^*\|^2 \le \|\theta_k - \theta^*\|^2 + (\eta^2 - \frac{\eta}{L})\|\nabla\mathcal{L}(\theta_k)\|^2 - 2\eta(\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*)).$$

□

Remark 2.1.9. *The existence of a finite global minimizer does not follow from convexity and Lipschitz continuity. For instance, the function $f(x) = -\log x$ is convex on $(1, \infty)$, but the global minimizer does not exist.*

Corollary 2.1.10. *Let $\mathcal{L}$ be convex and $\nabla\mathcal{L}$ be L-Lipschitz. If θ^* is the global minimizer and $L\eta \leq 1$, then*

$$\|\theta_k - \theta^*\| \geq \|\theta_{k+1} - \theta^*\|.$$

We can further determine the convergence rate for $\mathcal{L}(\theta_k)$ (not for θ_k).

Theorem 2.1.11. *Let $\mathcal{L}$ be convex and $\nabla\mathcal{L}$ be L-Lipschitz. If θ^* is the global minimizer and $L\eta \leq 1$, then*

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) \leq \frac{1}{2\eta(T+1)}\|\theta_0 - \theta^*\|^2.$$

Proof. The standard method averages $\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*)$, that is,

$$\begin{aligned}
\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) &\leq \frac{1}{T+1}\sum_{i=0}^{T} \mathcal{L}(\theta_i) - \mathcal{L}(\theta^*) \\
&= \frac{1}{T+1}\sum_{i=0}^{T} \frac{1}{2\eta}\left(\|\theta_i - \theta^*\|^2 - \|\theta_{i+1} - \theta^*\|^2\right) \\
&\leq \frac{1}{T+1}\cdot\frac{1}{2\eta}\|\theta_0 - \theta^*\|^2.
\end{aligned}$$

□

Remark 2.1.12. *Consider the loss function $\mathcal{L}(\theta) = \frac{1}{2p(2p-1)}\theta^{2p}$ on $[-1, 1]$, its gradient is 1-Lipschitz, thus we take $\eta = 1$. The gradient descent becomes*

$$\theta_{k+1} = \theta_k - \frac{\theta_k^{2p-1}}{2p-1}.$$

We show that if $\theta_1 = 1$, then

$$\theta_k \asymp k^{-\frac{1}{2p-2}}.$$

The loss function $\mathcal{L}(\theta_k) \asymp k^{-\frac{2p}{2p-2}}$. It shows that $\mathcal{O}(k^{-1})$ is sort of optimal as $p \to \infty$.

Remark 2.1.13. *If $\mathcal{L}$ is just convex and $\nabla\mathcal{L}$ is not continuous, and $\nabla\mathcal{L}$ is just a sub-gradient, the convergence rate becomes $\mathcal{O}(k^{-1/2})$.*

Question 2.1.14. *Show the gradient flow (suppose the solution is unique) for a convex $\mathcal{L}(\theta)$ satisfying*

$$\mathcal{L}(\theta(t)) - \mathcal{L}(\theta^*) \leq \frac{\|\theta_0 - \theta^*\|^2}{2t}.$$

The GD (gradient descent) method is a special case of first-order methods, that is, only gradient information is used:

$$\theta_k \in \theta_0 + \text{span}\{\nabla\mathcal{L}(\theta_0), \cdots, \nabla\mathcal{L}(\theta_{k-1})\}.$$

It is known (see Nesterov et al. (2018)) that first-order methods generally can attain at most second-order convergence without further assumptions (other than convexity, L-smoothness, and existence of a finite global minimizer).

Theorem 2.1.15 (Nesterov). *For any $1 \leq k \leq \frac{1}{2}(d-1)$ and any $\theta_0 \in \mathbb{R}^d$, there exists a convex loss function $\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$ with L-Lipschitz gradient $\nabla\mathcal{L}$ that any first-order method leads to*

$$\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*) \geq \frac{3L\|\theta_0 - \theta^*\|^2}{32},$$

and

$$\|\theta_k - \theta^*\|^2 \geq \frac{1}{8}\|\theta_0 - \theta^*\|^2.$$

A proof can be found in (Nesterov et al., 2018, Chapter 2). The second-order convergence will be introduced in the momentum methods.

2.1.3 Strong Convexity

Definition 2.1.16. *$\mathcal{L}$ is μ-strongly convex is $\mathcal{L}(\theta) - \frac{\mu}{2}\|\theta\|^2$ is convex.*

The μ-strongly convexity and L-Lipschitz is equivalent to say

$$\mathcal{J}(\theta) := \mathcal{L}(\theta) - \mu\frac{\|\theta\|^2}{2}$$

is $(L-\mu)$-Lipschitz and convex.

Problem 2.1.17. *Show that if $\mathcal{L}$ is μ-strongly convex, then*

$$\mathcal{L}(\theta') \geq \mathcal{L}(\theta) + \nabla\mathcal{L}(\theta)\cdot(\theta'-\theta) + \frac{\mu}{2}\|\theta'-\theta\|^2.$$

If θ^* is a global minimizer, then

$$\mathcal{L}(\theta) - \mathcal{L}(\theta^*) \leq \frac{1}{2\mu}\|\nabla\mathcal{L}(\theta)\|^2.$$

Question 2.1.18. *Prove that the gradient flow now satisfies*

$$\|\theta(t)-\theta^*\|^2 \leq \exp(-\mu t)\|\theta_0-\theta^*\|^2.$$

Theorem 2.1.19 (Co-Coercivity). *The μ-strongly convexity and L-Lipschitz imply the following inequality.*

$$\langle\nabla\mathcal{L}(\theta)-\nabla\mathcal{L}(\theta'),\theta-\theta'\rangle \geq \frac{L\mu}{L+\mu}\|\theta-\theta'\|^2 + \frac{1}{L+\mu}\|\nabla\mathcal{L}(\theta)-\nabla\mathcal{L}(\theta')\|^2.$$

Proof. Since $\mathcal{L}(\theta) - \frac{\mu}{2}\|\theta\|^2$ is convex and its gradient is $(L-\mu)$-Lipschitz, the co-coercivity implies

$$\langle\nabla\mathcal{L}(\theta)-\nabla\mathcal{L}(\theta'),\theta-\theta'\rangle - \mu\|\theta-\theta'\|^2 \geq \frac{\|\nabla\mathcal{L}(\theta)-\nabla\mathcal{L}(\theta')-\mu(\theta-\theta')\|^2}{(L-\mu)}.$$

Rearrange the terms, and obtain

$$(L+\mu)\langle\nabla\mathcal{L}(\theta)-\nabla\mathcal{L}(\theta'),\theta-\theta'\rangle \geq \|\nabla\mathcal{L}(\theta)-\nabla\mathcal{L}(\theta')\|^2 + \mu L\|\theta-\theta'\|^2.$$

□

Theorem 2.1.20. *Let $\mathcal{L}$ be μ-strongly convex and $\nabla\mathcal{L}$ be L-Lipschitz. If $\eta \leq \frac{2}{L+\mu}$, then*

$$\|\theta_{k+1}-\theta^*\|^2 \leq \left(1-\eta\frac{2\mu L}{\mu+L}\right)\|\theta_k-\theta^*\|^2.$$

Proof. Use the co-coercivity,

$$\begin{aligned}\|\theta_{k+1}-\theta^*\|^2 &= \|\theta_k-\eta\nabla\mathcal{L}(\theta_k)-\theta^*\|^2\\ &= \|\theta_k-\theta^*\|^2-2\eta\langle\theta_k-\theta^*,\nabla\mathcal{L}(\theta_k)\rangle+\eta^2\|\nabla\mathcal{L}(\theta_k)\|^2\\ &\le \left(1-\frac{2\eta L\mu}{L+\mu}\right)\|\theta_k-\theta^*\|^2+\left(\eta^2-\frac{2\eta}{L+\mu}\right)\|\nabla\mathcal{L}(\theta_k)\|^2.\end{aligned}$$

□

We can see from the above theorem that $\mu > 0$ is essential for exponential decay.

Corollary 2.1.21. *Let $\mathcal{L}$ be μ-strongly convex and $\nabla\mathcal{L}$ be L-Lipschitz. If $\eta \le \frac{2}{L+\mu}$, then*

$$\|\theta_k-\theta^*\|^2 \le \left(1-\eta\frac{2\mu L}{\mu+L}\right)^k\|\theta_0-\theta^*\|^2.$$

When $\eta = \frac{2}{L+\mu}$, the convergence becomes $\|\theta_k-\theta^*\| \le \rho^k\|\theta_0-\theta^*\|$, where

$$\rho=\sqrt{1-\frac{4\mu L}{(\mu+L)^2}}=\frac{(L-\mu)}{(L+\mu)}=\frac{\kappa-1}{\kappa+1},\quad \kappa=\frac{L}{\mu}.$$

Example 2.1.22. *If the loss function is $\mathcal{L}(\theta)=\frac{1}{2}\theta^TA\theta$, where A is a positive definite matrix. Then $\mu=\min(\lambda(A))$ and $L=\max(\lambda(A))$. The parameter κ represents the condition number. In this case, we can see $\eta=\left(\frac{1}{2}(\mu+L)\right)^{-1}$ is the optimal value for*

$$\min_{\eta\in\mathbb{R}}\max\left(|1-\eta\mu|,|1-\eta L|\right).$$

2.1.4 Non-convex Assumption

Standard loss functions (e.g., MSE, cross-entropy) for neural networks are generally nonconvex. Therefore, it is not guaranteed that optimization will converge to the global minimum.

Definition 2.1.23. *We consider a sequence of learning rates η_k that satisfies*

$$\sum_{k=1}^{\infty}\eta_k=\infty,\quad \sum_{k=1}^{\infty}\eta_k^2<\infty.$$

The constant learning rate in the previous section does not belong to this class. The learning rates with slow decay $\eta_k = \frac{C_0}{k+C_1}$ belong to this class. It is proved in Bertsekas and Tsitsiklis (2000) that the gradient sequence $\nabla\mathcal{L}(\theta_k)$ vanishes as $k \to \infty$ for the gradient descent method with step size η_k generically. The original proof is more general. We provide a much simpler version.

Theorem 2.1.24. *Let $\mathcal{L}(\theta)$ be differentiable and L-smooth, then with the aforementioned learning rates, either $\mathcal{L}(\theta_k) \to -\infty$, or $\lim_{k\to\infty} \|\nabla\mathcal{L}(\theta_k)\| = 0$.*

Rough Idea. The L-smoothness implies that

$$\mathcal{L}(\theta_{k+1}) \leq \mathcal{L}(\theta_k) + \nabla\mathcal{L}(\theta_k)\cdot(\theta_{k+1}-\theta_k) + \frac{L^2}{2}\|\theta_{k+1}-\theta_k\|^2.$$

Bring $\theta_{k+1} = \theta_k - \eta_k\nabla\mathcal{L}(\theta_k)$ into the above formula, then

$$\mathcal{L}(\theta_{k+1}) \leq \mathcal{L}(\theta_k) - \left(\eta_k - \frac{L^2}{2}\eta_k^2\right)\|\nabla\mathcal{L}(\theta_k)\|^2.$$

Choose a small $\varepsilon > 0$. Let us suppose $\eta_k < \varepsilon \ll \frac{2}{L}$ for all $k > k_0$ (the 2nd condition allows us to do so), we can see $\mathcal{L}(\theta_k)$ decays monotonically if $k > k_0$. If

$$\sum_{k>k_0} \eta_k\|\nabla\mathcal{L}(\theta_k)\|^2 = \infty,$$

then $\mathcal{L}(\theta_k) \to -\infty$. Otherwise, $\lim_{k\to\infty}\mathcal{L}(\theta_k)$ exists and

$$\sum_{k>k_0} \eta_k\|\nabla\mathcal{L}(\theta_k)\|^2 < \infty.$$

The above bound implies that $\liminf_{k\to\infty}\|\nabla\mathcal{L}(\theta_k)\| = 0$. Otherwise contradicts with the divergence of $\sum_{k>k_0}\eta_k = \infty$.

It is relatively simple to show $\liminf_{k\to\infty}\|\nabla\mathcal{L}(\theta_k)\| = 0$ (why?). Next, we show $\limsup_{k\to\infty}\|\nabla\mathcal{L}(\theta_k)\| = 0$. Otherwise, assume that

$$\limsup_{k\to\infty}\|\nabla\mathcal{L}(\theta_k)\| = 3\delta > 0.$$

Then, there are infinitely many (sufficiently large) k that $\|\nabla\mathcal{L}(\theta_k)\| \in (\delta, 2\delta)$ for a certain $\delta > 0$ (since the steps are small enough). For any such k, let the

consecutive neighboring indices $\mathcal{N}(k) := \{l \mid \|\nabla\mathcal{L}(\theta_l)\| \in (\delta, 2\delta)\}$. Thus,

$$\sum_{l\in\mathcal{N}(k)} \eta_l \|\nabla\mathcal{L}(\theta_l)\|^2 > \left(\sum_{l\in\mathcal{N}(k)} \eta_l \|\nabla\mathcal{L}(\theta_l)\|\right) \delta \asymp \frac{\delta^2}{L}.$$

where

$$\begin{aligned}\sum_{l\in\mathcal{N}(k)} \eta_l \|\nabla\mathcal{L}(\theta_l)\| &= \sum_{l\in\mathcal{N}(k)} \|\theta_{l+1} - \theta_l\| \\ &\geq \frac{1}{L} \sum_{l\in\mathcal{N}(k)} \big|\|\nabla\mathcal{L}(\theta_{l+1})\| - \|\nabla\mathcal{L}(\theta_l)\|\big| \asymp \frac{\delta}{L}.\end{aligned}$$

Hence, we obtain a contradiction due to the existence of an infinite such k, and $\lim_{k\to\infty} \|\nabla\mathcal{L}(\theta_k)\| = 0$. □

Remark 2.1.25. *It essentially shows that if $f \in \mathrm{Lip}(\mathbb{R})$ that*

$$|f(x) - f(y)| \leq L\|x - y\|,$$

and

$$\int_{\mathbb{R}} |f(x)|dx < \infty,$$

Then $\lim_{|x|\to\infty} f(x) = 0$. Here f can be viewed as $\nabla\mathcal{L}(\theta)$ and $d\theta \sim \theta_{k+1} - \theta_k$.

Question 2.1.26. *Does the above theorem 2.1.24 hold with constant learning rates $\eta_k \equiv \eta < \frac{2}{L}$?*

Question 2.1.27. *What if the learning rates only satisfy a weaker condition $\sum_k \eta_k = \infty$ and $\lim_{k\to\infty} \eta_k = 0$?*

Later, we will see that ℓ^2-finiteness of the learning rates is useful for stochastic gradient descent.

2.2 Momentum Methods

In this section, we return to the "optimal" convergence rate for first-order methods.

2.2.1 Polyak's method

Polyak's method, or the heavy-ball method, was proposed in Polyak (1964) to overcome the oscillations of the iterates in the usual gradient descent method. The Polyak iteration includes an additional momentum term, which represents a *second-order correction*.

$$\theta_{k+1} = \theta_k - \eta \nabla \mathcal{L}(\theta_k) + \beta(\theta_k - \theta_{k-1}).$$

Note that we need two initial conditions θ_0 and θ_{-1} to compute the iterations. Let θ_k be the position vector of a particle at $t = t_k$, then the above iteration becomes

$$(1-\beta)\frac{\theta_{k+1}-\theta_k}{\eta} + \beta\eta\left(\frac{(\theta_{k+1}-\theta_k)-(\theta_k-\theta_{k-1})}{\eta^2}\right) = -\nabla\mathcal{L}(\theta_k),$$

It can be *locally* viewed as an approximation to the ODE

$$\frac{d}{dt}\theta + \frac{\beta\eta}{1-\beta}\frac{d^2}{dt^2}\theta = -\frac{1}{1-\beta}\nabla\mathcal{L}(\theta).$$

Next, we show that Polyak's method has a better *local* convergence than the standard gradient descent method.

Remark 2.2.1. *A high-order ODE can be transformed into a first-order dynamical system. This is a typical treatment. In the above 2nd-order ODE, we can write*

$$\frac{d}{dt}\begin{pmatrix}\theta\\ \dot\theta\end{pmatrix} = \begin{pmatrix}0 & 1\\ 0 & -\frac{1-\beta}{\beta\eta}\end{pmatrix}\begin{pmatrix}\theta\\ \dot\theta\end{pmatrix} + \begin{pmatrix}0\\ -\frac{1}{\beta\eta}\nabla\mathcal{L}(\theta)\end{pmatrix}.$$

Remark 2.2.2. *From a physics perspective, the ODE can be viewed as Newton's 2nd law with friction. The variable θ represents the position and $\dot\theta$ represents the momentum. The loss function becomes a potential energy. However, it will be hard to explain the standard GD in a similar way using dimension analysis. We will see this in a later section.*

Theorem 2.2.3. *Set $p_k = \frac{\beta}{1-\beta}(\theta_k - \theta_{k-1})$. Let $\mathcal{L}$ be convex and L-smooth. If θ^* is the global minimizer, then*

$$\begin{aligned}\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*) \le (1-\beta)&\frac{\|\theta_k + p_k - p^*\|^2 - \|\theta_k + p_{k+1} - p^*\|^2}{2\eta\lambda}\\ &+ \frac{(\eta - (1-\beta)\frac{2-\lambda}{L})}{2(1-\beta)\lambda}\|\nabla\mathcal{L}(\theta_k)\|^2\\ &- \frac{\beta}{(1-\beta)\lambda}\left(\mathcal{L}(\theta_k) - \mathcal{L}(\theta_{k-1})\right).\end{aligned}$$

Proof. Due to the similarity of Polyak's method with the standard gradient descent ($\beta = 0$), the proof shares a similar flavor as Theorem 2.1.8. Then

$$\theta_{k+1} + p_{k+1} = \theta_k + p_k - \frac{\eta}{1-\beta}\nabla\mathcal{L}(\theta_k).$$

We have (recall Problem 2.1.6)

$$\begin{aligned}
\|\theta_{k+1} + p_{k+1} - \theta^*\|^2 &= \|\theta_k + p_k - \theta^*\|^2 - \frac{2\eta}{1-\beta}\langle \theta_k + p_k - \theta^*, \nabla\mathcal{L}(\theta_k)\rangle \\
&\quad + \left(\frac{\eta}{1-\beta}\right)^2 \|\nabla\mathcal{L}(\theta_k)\|^2 \\
&= \|\theta_k + p_k - \theta^*\|^2 + \left(\frac{\eta}{1-\beta}\right)^2 \|\nabla\mathcal{L}(\theta_k)\|^2 \\
&\quad - \frac{2\eta}{1-\beta}\langle \theta_k - \theta^*, \nabla\mathcal{L}(\theta_k)\rangle \\
&\quad - \frac{2\eta\beta}{(1-\beta)^2}\langle \theta_k - \theta_{k-1}, \nabla\mathcal{L}(\theta_k)\rangle \\
&\le \|\theta_k + p_k - \theta^*\|^2 + \left(\frac{\eta}{1-\beta}\right)^2 \|\nabla\mathcal{L}(\theta_k)\|^2 \\
&\quad - \frac{2\eta}{1-\beta}\frac{1-\lambda}{L}\|\nabla\mathcal{L}(\theta_k)\|^2 \\
&\quad - \frac{2\eta}{1-\beta}\lambda\left(\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*) + \frac{1}{2L}\|\nabla\mathcal{L}(\theta_k)\|^2\right) \\
&\quad - \frac{2\eta\beta}{(1-\beta)^2}(\mathcal{L}(\theta_k) - \mathcal{L}(\theta_{k-1}))
\end{aligned}$$

where $\lambda \in [0,1]$ is an arbitrary parameter. The conclusion is obtained by rearranging the terms. □

Therefore, if $\eta \le \frac{2-\lambda}{L}(1-\beta)$, we have

$$\begin{aligned}
\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*) &\le (1-\beta)\frac{\|\theta_k + p_k - \theta^*\|^2 - \|\theta_{k+1} + p_{k+1} - \theta^*\|^2}{2\eta\lambda} \\
&\quad - \frac{\beta}{(1-\beta)\lambda}\left(\mathcal{L}(\theta_k) - \mathcal{L}(\theta_{k-1})\right).
\end{aligned}$$

Take the sum over k, we find that

$$\begin{aligned}
\sum_{k=0}^{T} \mathcal{L}(\theta_k) - \mathcal{L}(\theta^*) &\le (1-\beta)\frac{\|\theta_0 + p_0 - \theta^*\|^2}{2\eta\lambda} + \frac{\beta}{(1-\beta)\lambda}(\mathcal{L}(\theta_{-1}) - \mathcal{L}(\theta_T)) \\
&\le (1-\beta)\frac{\|\theta_0 + p_0 - \theta^*\|^2}{2\eta(1-\beta)\lambda} + \frac{\beta}{(1-\beta)\lambda}(\mathcal{L}(\theta_{-1}) - \mathcal{L}(\theta^*)).
\end{aligned}$$

Using the convexity of $\mathcal{L}$, we can replace the LHS by the "moving average".

Corollary 2.2.4. *If $\eta \leq \frac{2-\lambda}{L}(1-\beta)$, then*

$$\mathcal{L}(\overline{\theta}_T) - \mathcal{L}(\theta^*) \leq \frac{1}{T+1}\left((1-\beta)\frac{\|\theta_0 + p_0 - \theta^*\|^2}{2\eta\lambda} + \frac{\beta(\mathcal{L}(\theta_{-1}) - \mathcal{L}(\theta^*))}{(1-\beta)\lambda}\right),$$

where $\overline{\theta}_T = \frac{1}{T+1}\sum_{k=0}^{T}\theta_k$ is the Cesáro average.

Remark 2.2.5. *If $\theta_{-1} = \theta_0$ as the initial points, then $p_0 = 0$ and by L-smoothness,*

$$\mathcal{L}(\theta_0) - \mathcal{L}(\theta^*) \leq \frac{L}{2}\|\theta_0 - \theta^*\|^2.$$

The above RHS can be reduced to

$$\frac{1}{2\lambda}\cdot\frac{1}{T+1}\left(\frac{1-\beta}{\eta} + \frac{\beta L}{(1-\beta)}\right)\|\theta_0 - \theta^*\|^2.$$

The optimal β^* satisfies $\beta^* = 1 - \sqrt{\eta L}$.

Theorem 2.2.6. *Take $\lambda = 1$. If $\eta \leq \frac{1}{L}$ and $\beta = 1 - \sqrt{\eta L}$, then*

$$\left[\min_{0\leq k\leq T}\mathcal{L}(\theta_k)\right] - \mathcal{L}(\theta^*) \leq \frac{1}{2(T+1)}\left(\frac{2\sqrt{\eta L} - \eta L}{\eta}\right)\|\theta_0 - \theta^*\|^2.$$

We can compare the bound for $\beta = 0$, which reduces to $\frac{1}{2(T+1)}\frac{1}{\eta}\|\theta_0 - \theta^*\|^2$ for the standard gradient descent method. The Polyak's bound is always better since

$$2\sqrt{\eta L} - \eta L - 1 \leq -(1 - \sqrt{\eta L})^2 \leq 0.$$

These two bounds coincide when $\eta = \frac{1}{L}$.

Problem 2.2.7. *Does Polyak's iteration outperform the standard GD in the sense of the last iteration $\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*)$? What is the main difficulty or difference from the standard GD?*

If we allow variable learning rates and momentum parameters. The convergence can be quantified in the desired way.

Corollary 2.2.8. *Let*

$$\beta_k = \frac{k}{k+1}, \quad \eta_k = \frac{\eta_0}{k+1}.$$

Then we can show that Polyak's method has linear convergence in the standard sense.

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) \leq \frac{\|\theta_0 - \theta^*\|^2}{2\eta_0(T+1)}.$$

The proof is no different from Theorem 2.2.3 but can be significantly simplified.

Problem 2.2.9. *Prove the above corollary.*

Problem 2.2.10. *Show that the ODE's solution $\theta(t)$ to*

$$(1-\beta)\frac{d}{dt}\theta + \beta\eta\frac{d^2}{dt^2}\theta = -\nabla\mathcal{L}(\theta)$$

with initial condition $\theta(0) = \theta_0$ and $\theta'(0) = 0$ satisfies

$$\int_0^T (\mathcal{L}(\theta(t)) - \mathcal{L}(\theta^*))dt \leq \left((1-\beta)\|\theta_0 - \theta^*\|^2 + \frac{\beta\eta}{1-\beta}(\mathcal{L}(\theta_0) - \mathcal{L}(\theta^*))\right).$$

Next, we consider $\mathcal{L}$ as a μ-strongly and L-smooth convex loss function. The following proof is based on Ghadimi et al. (2015). The following simple lemma describes when a sequence decays exponentially.

Lemma 2.2.11 (Exponential Decaying Sequence). *Let $\{A_k\}_{k\geq 0}$ and $\{B_k\}_{k\geq 0}$ be two non-negative sequences. Suppose there are constants $a_1, a_2, b \in \mathbb{R}_+$ and $c \in \mathbb{R}$ that*

$$A_{k+1} + bB_{k+1} \leq a_1 A_k + a_2 A_{k-1} + cB_k,$$

where $A_0 = A_{-1}$, $a_1 + a_2 < 1$, $c < b$. Then

$$A_k \leq \rho^k \left((\rho - a_1 + 1)A_0 + bB_0\right),$$

Continue on next page ⇓

Continue from last page ⇑

where

$$\rho = \max\left(\frac{c}{b}, \frac{a_1 + \sqrt{a_1^2 + 4a_2}}{2}\right).$$

Proof. Let $r \geq 0$, we construct the relation

$$\begin{aligned} A_{k+1} + rA_k + bB_{k+1} &\leq (a_1 + r)A_k + a_2 A_{k-1} + cB_k \\ &\leq \underbrace{(a_1 + r)}_{=\rho}[A_k + rA_{k-1} + bB_k]. \end{aligned}$$

This requires

$$(a_1 + r)r \geq a_2, \quad (a_1 + r)b \geq c.$$

That is,

$$r \geq \max\left(\frac{c}{b} - a_1, \frac{-a_1 + \sqrt{a_1^2 + 4a_2}}{2}\right).$$

Thus,

$$\begin{aligned} A_k \leq A_k + rA_{k-1} + bB_k &\leq \rho^k(A_0 + rA_{-1} + bB_0) \\ &= \rho^k((1 + \rho - a_1)A_0 + bB_0). \end{aligned}$$

□

The general idea of the proof of convergence in the μ-strongly convex case is similar to that in the previous convex case. The (Lyapunov) functional will be a combination of the loss function and the quadratic distance.

Theorem 2.2.12. *Suppose $\mathcal{L}$ is μ-strongly convex and L-smooth. There are constants $\rho \in (0, 1)$ and $C < \infty$ that*

$$\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*) \leq C\rho^k(\mathcal{L}(\theta_0) - \mathcal{L}(\theta^*)).$$

Proof. We consider the following quantities

$$A_{k+1} = \mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta^*), \quad B_{k+1} = \|\theta_{k+1} - \theta_k\|^2.$$

We verify that

$$\begin{aligned} \mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta_k) &\leq \nabla\mathcal{L}(\theta_k) \cdot (\theta_{k+1} - \theta_k) + \frac{L}{2}\|\theta_{k+1} - \theta_k\|^2 \\ &= -\eta(1 - \frac{L\eta}{2})\|\nabla\mathcal{L}(\theta_k)\|^2 + \frac{L\beta^2}{2}\|\theta_k - \theta_{k-1}\|^2 \\ &\quad + \beta(1 - \eta L)\nabla\mathcal{L}(\theta_k) \cdot (\theta_k - \theta_{k-1}), \end{aligned}$$

and

$$\|\theta_{k+1}-\theta_k\|^2 = \eta^2\|\nabla\mathcal{L}(\theta_k)\|^2 + \beta^2\|\theta_k-\theta_{k-1}\|^2 - 2\eta\beta\nabla\mathcal{L}(\theta_k)\cdot(\theta_k-\theta_{k-1}).$$

The following inequalities are straightforward (from the previous exercises):

$$\begin{aligned}\|\nabla\mathcal{L}(\theta_k)\|^2 &\geq 2\mu(\mathcal{L}(\theta_k)-\mathcal{L}(\theta^*)),\\ \nabla\mathcal{L}(\theta_k)\cdot(\theta_k-\theta_{k-1}) &\geq \mathcal{L}(\theta_k)-\mathcal{L}(\theta_{k-1})+\frac{\mu}{2}\|\theta_k-\theta_{k-1}\|^2.\end{aligned}$$

Suppose we find $b\in\mathbb{R}_+$ that

$$\frac{(1-\eta L)}{2\eta} < b < \frac{2-\eta L}{2\eta}.$$

Then (why?)

$$\begin{aligned}&A_{k+1}+bB_{k+1}-A_k-bB_k\\ &= \mathcal{L}(\theta_{k+1})-\mathcal{L}(\theta_k)+b\left(\|\theta_{k+1}-\theta_k\|^2-\|\theta_k-\theta_{k-1}\|^2\right)\\ &\leq \left(b\eta^2-\eta(1-\frac{L\eta}{2})\right)\|\nabla\mathcal{L}(\theta_k)\|^2+\frac{(L+2b)\beta^2-2b}{2}\|\theta_k-\theta_{k-1}\|^2\\ &\quad+\left(\beta(1-\eta L)-2b\eta\beta\right)\nabla\mathcal{L}(\theta_k)\cdot(\theta_k-\theta_{k-1})\\ &\leq 2\mu\left(b\eta^2-\eta(1-\frac{L\eta}{2})\right)(\mathcal{L}(\theta_k)-\mathcal{L}(\theta^*))+\frac{(L+2b)\beta^2-2b}{2}\|\theta_k-\theta_{k-1}\|^2\\ &\quad+\left(\beta(1-\eta L)-2b\eta\beta\right)\left[\mathcal{L}(\theta_k)-\mathcal{L}(\theta_{k-1})+\frac{\mu}{2}\|\theta_k-\theta_{k-1}\|^2\right]\\ &= (a_1-1)A_k+a_2A_{k-1}+(c-b)\|\theta_k-\theta_{k-1}\|^2.\end{aligned}$$

where the constants

$$\begin{aligned}a_1 &= 1+2\mu\eta\left(b\eta+\frac{L\eta}{2}-1\right)+\beta\left(1-\eta L-2b\eta\right),\\ a_2 &= -\beta(1-\eta L-2b\eta),\\ c &= \frac{L+2b}{2}\beta^2-\frac{\mu}{2}a_2.\end{aligned}$$

Therefore,

$$A_{k+1}+bB_{k+1}\leq a_1A_k+a_2A_{k-1}+cB_k.$$

and we verify the conditions.

$$a_1,a_2,b>0,\quad a_1+a_2<1,\quad c<b.$$

The parameter β should satisfy

$$0 \leq \beta < \frac{\mu((L+2b)\eta - 1) + \sqrt{\mu^2((L+2b)\eta-1)^2 + 4b(L+2b)}}{2(L+2b)},$$

and the ratio

$$\rho = \max\left(\frac{c}{b}, \frac{a_1 + \sqrt{a_1^2 + 4a_2}}{2}\right).$$

Therefore,

$$\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta^*) \leq \rho^{k+1}(1 + \rho - a_1)(\mathcal{L}(\theta_0) - \mathcal{L}(\theta^*)).$$

□

Remark 2.2.13. *Using the relation that $\mathcal{L}(\theta) - \mathcal{L}(\theta^*) \sim \|\theta - \theta^*\|^2$, this theorem implies the exponential decay of $\|\theta_k - \theta^*\|$. The ODE version of the above theorem can be constructed using a similar idea.*

2.2.2 Nesterov's method

Nesterov's method Nesterov (1998) improves on Polyak's method. Instead of the $\mathcal{O}(T^{-1})$ convergence rate for L-smooth convex loss function, Nesterov's method can attain the "optimal" $\mathcal{O}(T^{-2})$ convergence rate using **variable** parameters, which is known as Nesterov Accelerated Gradient (NAG) method. The update scheme is slightly different (and more complicated) from Polyak's momentum method. We first review the case with constant parameters.

$$\begin{aligned} \theta_{k+1} &= \gamma_k - \eta \nabla \mathcal{L}(\gamma_k) \\ \gamma_k &= \theta_k + \beta \underbrace{(\theta_k - \theta_{k-1})}_{\text{momentum}}. \end{aligned} \tag{NAG}$$

Note that there is a difference from Polyak's method

$$\begin{aligned} \theta_{k+1} &= \gamma_k - \eta \nabla \mathcal{L}(\theta_k) \\ \gamma_k &= \theta_k + \beta \underbrace{(\theta_k - \theta_{k-1})}_{\text{momentum}}. \end{aligned} \tag{Heavy-Ball}$$

The iteration γ_k can be viewed as a "lookahead" correction of θ_k in the sense that if γ_k improves the convergence over θ_k, then why not use a more accurate version in the update of gradient descent?

Remark 2.2.14. *If we let $v_k = \theta_k - \theta_{k-1}$ be the momentum variable, then the Heavy-*

Ball algorithm becomes

$$v_{k+1} = \beta v_k - \eta \nabla \mathcal{L}(\theta_k), \qquad \theta_{k+1} = \theta_k + v_{k+1}.$$

The NAG becomes

$$v_{k+1} = \beta v_k - \eta \nabla \mathcal{L}(\underbrace{\theta_k + \beta v_k}_{\text{lookahead point}}), \qquad \theta_{k+1} = \theta_k + v_{k+1}.$$

However, the constant parameter setting in NAG still yields linear convergence for convex functions.

Theorem 2.2.15. *Let $\mathcal{L}$ be convex and L-smooth. Assume $\beta \in [0, 1)$, and $\theta_0 = \theta_{-1}$ Then, NAG iteration satisfies*

$$\mathcal{L}(\overline{\gamma}_T) - \mathcal{L}(\theta^*) \leq \frac{1}{T+1}\left(\frac{\beta}{1-\beta}(\mathcal{L}(\gamma_{-1}) - \mathcal{L}(\theta^*)) + \frac{1-\beta}{2\eta}\|\gamma_0 - \theta^*\|^2\right),$$

where $\overline{\gamma}_T = \frac{1}{T+1}\sum_{k=0}^{T}\gamma_k$ is the Cesáro average.

Proof. Let $p_k = \frac{\beta}{1-\beta}(\gamma_k - \gamma_{k-1} + \eta\nabla\mathcal{L}(\gamma_{k-1})) = \frac{\beta}{1-\beta}(\gamma_k - \theta_k) = \frac{\beta^2}{1-\beta}(\theta_k - \theta_{k-1})$. Then

$$\begin{aligned}\gamma_{k+1} + p_{k+1} &= \frac{1}{1-\beta}\gamma_{k+1} + \frac{\beta}{1-\beta}(\eta\nabla\mathcal{L}(\gamma_k) - \gamma_k) \\ &= \gamma_k + p_k - \frac{\eta}{1-\beta}\nabla\mathcal{L}(\gamma_k).\end{aligned}$$

This is essentially the same as the proof for Polyak's method. Then, let $\gamma^* = \theta^*$,

recall Problem 2.1.6,

$$\begin{aligned}
\|\gamma_{k+1} + p_{k+1} - \gamma^*\|^2 &= \|\gamma_k + p_k - \gamma^*\|^2 - \frac{2\eta}{1-\beta}\langle \gamma_k + p_k - \gamma^*, \nabla\mathcal{L}(\gamma_k)\rangle \\
&\quad + \frac{\eta^2}{(1-\beta)^2}\|\nabla\mathcal{L}(\gamma_k)\|^2 \\
&= \|\gamma_k + p_k - \gamma^*\|^2 + \frac{\eta^2}{(1-\beta)^2}\|\nabla\mathcal{L}(\gamma_k)\|^2 \\
&\quad - \frac{2\eta}{1-\beta}\langle \gamma_k - \gamma^*, \nabla\mathcal{L}(\gamma_k)\rangle \\
&\quad - \frac{2\beta\eta}{(1-\beta)^2}\langle \gamma_k - \gamma_{k-1}, \nabla\mathcal{L}(\gamma_k)\rangle \\
&\quad - \frac{2\beta\eta^2}{(1-\beta)^2}\langle \nabla\mathcal{L}(\gamma_{k-1}), \nabla\mathcal{L}(\gamma_k)\rangle \\
&\leq \|\gamma_k + p_k - \gamma^*\|^2 + \frac{\eta^2}{(1-\beta)^2}\|\nabla\mathcal{L}(\gamma_k)\|^2 \\
&\quad - \frac{2\eta}{1-\beta}(\mathcal{L}(\gamma_k) - \mathcal{L}(\gamma^*) + \frac{1}{2L}\|\nabla\mathcal{L}(\gamma_k)\|^2) \\
&\quad - \frac{2\beta\eta}{(1-\beta)^2}(\mathcal{L}(\gamma_k) - \mathcal{L}(\gamma_{k-1}) + \frac{1}{2L}\|\nabla\mathcal{L}(\gamma_k) - \nabla\mathcal{L}(\gamma_{k-1})\|^2) \\
&\quad - \frac{2\beta\eta^2}{(1-\beta)^2}\langle \nabla\mathcal{L}(\gamma_{k-1}), \nabla\mathcal{L}(\gamma_k)\rangle
\end{aligned}$$

We rearrange the terms and obtain

$$\begin{aligned}
&\frac{2\eta}{(1-\beta)^2}(\mathcal{L}(\gamma_k) - \mathcal{L}(\gamma^*)) + \|\gamma_{k+1} + p_{k+1} - \gamma^*\|^2 \\
&\leq \frac{2\beta\eta}{(1-\beta)^2}(\mathcal{L}(\gamma_{k-1}) - \mathcal{L}(\gamma^*)) + \|\gamma_k + p_k - \gamma^*\|^2 \\
&\quad + \underbrace{\left(\frac{\eta^2}{(1-\beta)^2} - \frac{\eta}{(1-\beta)L} - \frac{\beta\eta}{(1-\beta)^2}\frac{1}{L}\right)}_{=\frac{\eta(\eta-1/L)}{(1-\beta)^2}}\|\nabla\mathcal{L}(\gamma_k)\|^2 - \frac{\beta\eta}{L(1-\beta)^2}\|\nabla\mathcal{L}(\gamma_{k-1})\|^2 \\
&\quad + \underbrace{\left(\frac{2\beta\eta}{L(1-\beta)^2} - \frac{2\beta\eta^2}{(1-\beta)^2}\right)}_{=\frac{2\beta\eta}{(1-\beta)^2}(\frac{1}{L}-\eta)}\langle \nabla\mathcal{L}(\gamma_{k-1}), \nabla\mathcal{L}(\gamma_k)\rangle
\end{aligned}$$

If $0 < \eta \leq \frac{1}{L}$, the quadratic form of the last 3 terms has a determinant

$$\begin{aligned}\Delta &= \left(\frac{2\beta\eta}{(1-\beta)^2}(\frac{1}{L}-\eta)\right)^2 + 4\frac{\eta(\eta-1/L)}{(1-\beta)^2}\frac{\beta\eta}{L(1-\beta)^2} \\ &= \frac{4\eta(\eta-1/L)}{(1-\beta)^2}\frac{\beta\eta}{L(1-\beta)^2}\left(\beta(L\eta-1)-1\right) \leq 0.\end{aligned}$$

Therefore, we take the summation over $k = 0, \cdots, T$,

$$\begin{aligned}&\sum_{k=0}^{T}(\mathcal{L}(\gamma_k)-\mathcal{L}(\gamma^*)) + \sum_{k=0}^{T}\frac{\beta}{(1-\beta)}(\mathcal{L}(\gamma_k)-\mathcal{L}(\gamma^*)) \\ &\quad + \sum_{k=0}^{T}\frac{1-\beta}{2\eta}\|\gamma_{k+1}+p_{k+1}-\gamma^*\|^2 \\ &\leq \sum_{k=0}^{T}\frac{\beta}{(1-\beta)}(\mathcal{L}(\gamma_{k-1})-\mathcal{L}(\gamma^*)) + \sum_{k=0}^{T}\frac{1-\beta}{2\eta}\|\gamma_k+p_k-\gamma^*\|^2.\end{aligned}$$

That is,

$$\sum_{k=0}^{T}(\mathcal{L}(\gamma_k)-\mathcal{L}(\gamma^*)) \leq \frac{\beta}{(1-\beta)}\left(\mathcal{L}(\gamma_{-1})-\mathcal{L}(\gamma^*)\right) + \frac{1-\beta}{2\eta}\|\gamma_0+p_0-\gamma^*\|^2.$$

The LHS is bounded below by $(T+1)\mathcal{L}(\overline{\gamma}_T)$ for the Cesáro average. □

Remark 2.2.16. *The conclusion is essentially the same as in Polyak's case. The exponential convergence in the μ-strongly convex setting is similar.*

Regarding μ-strongly convex functions, the constant parameters $\eta = \frac{1}{L}$ and $\beta = \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}$, $\kappa = \frac{\mu}{L}$ are typical,

$$\begin{aligned}\theta_{k+1} &= \gamma_k - \frac{1}{L}\nabla\mathcal{L}(\gamma_k) \\ \gamma_k &= \theta_k + \frac{1-\sqrt{\kappa}}{1+\sqrt{\kappa}}\underbrace{(\theta_k-\theta_{k-1})}_{\text{momentum}},\end{aligned} \tag{NAG-sc}$$

and the convergence is exponentially fast

$$\mathcal{L}(\theta_T)-\mathcal{L}(\theta^*) \leq \rho^T\left(\mathcal{L}(\theta_0)-\mathcal{L}(\theta^*)+\frac{\mu}{2}\|\theta_0-\theta^*\|^2\right), \quad \rho := (1-\sqrt{\kappa}).$$

This setting is often known as NAG-sc. The proof is based on the construction of a Lyapunov function.

Theorem 2.2.17. *If $\mathcal{L}$ is μ-strongly convex and L-smooth, then NAG-sc converges exponentially,*

$$\mathcal{L}(\theta_T) - \mathcal{L}(\theta^*) \leq \rho^T \left(\mathcal{L}(\theta_0) - \mathcal{L}(\theta^*) + \frac{\mu}{2}\|\theta_0 - \theta^*\|^2 \right).$$

Proof. Let $\mathcal{A}_k = \mathcal{L}(\theta_k) - \mathcal{L}(\theta^*)$, and $\mathcal{B}_k = \|\theta_k - \rho\theta_{k-1} - (1-\rho)\theta^*\|^2$. Then

$$\begin{aligned}
\mathcal{A}_{k+1} &= \mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta^*) \\
&\leq \mathcal{L}(\gamma_k) - \mathcal{L}(\theta^*) + \langle \nabla\mathcal{L}(\gamma_k), \theta_{k+1} - \gamma_k \rangle + \frac{L}{2}\|\theta_{k+1} - \gamma_k\|^2 \\
&= \mathcal{L}(\gamma_k) - \mathcal{L}(\theta^*) - \frac{L}{2}\|\theta_{k+1} - \gamma_k\|^2
\end{aligned}$$

and use the following to rewrite the first term

$$\begin{aligned}
\mathcal{L}(\gamma_k) - \mathcal{L}(\theta_k) &\leq -\langle \nabla\mathcal{L}(\gamma_k), \theta_k - \gamma_k \rangle - \frac{\mu}{2}\|\gamma_k - \theta_k\|^2, \\
\mathcal{L}(\gamma_k) - \mathcal{L}(\theta^*) &\leq -\langle \nabla\mathcal{L}(\gamma_k), \theta^* - \gamma_k \rangle - \frac{\mu}{2}\|\gamma_k - \theta^*\|^2.
\end{aligned}$$

We obtain that

$$\begin{aligned}
\mathcal{A}_{k+1} + \frac{L}{2}\mathcal{B}_{k+1} \leq\ & \rho \left(\mathcal{A}_k - \langle \nabla\mathcal{L}(\gamma_k), \theta_k - \gamma_k \rangle - \frac{\mu}{2}\|\gamma_k - \theta_k\|^2 \right) \\
& + (1-\rho) \left(-\langle \nabla\mathcal{L}(\gamma_k), \theta^* - \gamma_k \rangle - \frac{\mu}{2}\|\gamma_k - \theta^*\|^2 \right) \\
& + \frac{L}{2}\|\theta_{k+1} - \rho\theta_k\|^2 - \frac{L}{2}\|\theta_{k+1} - \gamma_k\|^2 \\
=\ & \rho(\mathcal{A}_k + \frac{L}{2}\mathcal{B}_k) \\
& + \rho L\langle \theta_{k+1} - \gamma_k, \theta_k - \gamma_k \rangle - \rho\frac{\mu}{2}\|\gamma_k - \theta_k\|^2 \\
& + (1-\rho)L\langle \theta_{k+1} - \gamma_k, \theta^* - \gamma_k \rangle - (1-\rho)\frac{\mu}{2}\|\gamma_k - \theta^*\|^2 \\
& + \frac{L}{2}\|\theta_{k+1} - \rho\theta_k - (1-\rho)\theta^*\|^2 - \frac{L}{2}\|\theta_{k+1} - \gamma_k\|^2 \\
& - \rho\frac{L}{2}\|\theta_k - \rho\theta_{k-1} - (1-\rho)\theta^*\|^2
\end{aligned}$$

Replace $\theta_k - \rho\theta_{k-1} = (1-\rho)\theta_k + \frac{\rho}{\beta}(\gamma_k - \theta_k)$, we can compute the remainder part using the prescribed parameters as

$$-\frac{1}{2}(L-\mu)\sqrt{\frac{\mu}{L}}\|\theta_k - \gamma_k\|^2 \leq 0.$$

The verification step can be done symbolically through Mathematica or SymPy.

```
>> rho = 1 - Sqrt[mu/L];
>> beta = (1 - Sqrt[mu/L])/(1 + Sqrt[mu/L]);
>> Simplify[
rho*L*((theta_k1) - (gamma_k))*((theta_k) -(gamma_k))
- rho*(mu/2)*((gamma_k) - (theta_k))^2
+ (1 - rho)*L*((theta_k1) - (gamma_k))*((theta_star) - (gamma_k))
- (1 - rho)*(mu/2)*((gamma_k) - (theta_star))^2
+ (L/2)*((theta_k1) - rho*(theta_k)
- (1 - rho)*(theta_star))^2 - (L/2)*((theta_k1) - (gamma_k))^2
- rho*(L/2)*((1 - rho)*(theta_k) + rho * (1/(beta))*((gamma_k)
   - (theta_k)) - (1 - rho)*(theta_star))^2,
   Assumptions -> L > 0 && mu > 0]
>>-(1/2) (L-mu) Sqrt[mu/L] (gamma_k-theta_k)^2
```

□

Now, let us consider the NAG-c (Nesterov Accelerated Gradient) method for general **convex** functions with a variable step size.

$$\begin{aligned} \theta_{k+1} &= \gamma_k - \eta \nabla \mathcal{L}(\gamma_k), \\ \gamma_k &= \theta_k + \beta_k(\theta_k - \theta_{k-1}). \end{aligned} \tag{NAG-c}$$

The parameters satisfies $0 < \eta < \frac{1}{L}$ and $\beta_k = \frac{k}{k+3}$. The following theorem implies that NAG-c achieves the optimal convergence rate $\mathcal{O}(T^{-2})$ for convex and L -smooth functions.

Theorem 2.2.18. *Let $\mathcal{L}$ be convex and L-smooth. The NAG-c algorithm satisfies*

$$\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*) \leq \frac{9}{2\eta(k+2)^2}\|\theta_0 - \theta^*\|^2.$$

Proof. For any x, denote $x_+ = x - \eta\nabla\mathcal{L}(x)$, then

$$\begin{aligned} \mathcal{L}(x_+) - \mathcal{L}(x) &\leq \langle \nabla\mathcal{L}(x), x_+ - x\rangle + \frac{L}{2}\|x_+ - x\|^2 \\ &\leq \left(-\frac{1}{\eta} + \frac{L}{2}\right)\|x_+ - x\|^2 \leq -\frac{1}{2\eta}\|x_+ - x\|^2. \end{aligned}$$

Then

$$\begin{aligned}\mathcal{L}(x^+) - \mathcal{L}(z) &\le \mathcal{L}(x) - \mathcal{L}(z) - \frac{\eta}{2}\|\nabla\mathcal{L}(x)\|^2 \\ &\le \langle \nabla\mathcal{L}(x), x - z\rangle - \frac{\eta}{2}\|\nabla\mathcal{L}(x)\|^2 \\ &= -\frac{\eta}{2}\|\nabla\mathcal{L}(x) - \frac{1}{\eta}(x-z)\|^2 + \frac{1}{2\eta}\|x-z\|^2 \\ &= -\frac{1}{2\eta}\|x_+ - z\|^2 + \frac{1}{2\eta}\|x-z\|^2.\end{aligned} \tag{$*$}$$

Apply $x = \gamma_k$ and $z = \theta^*$ and $z = \theta_k$ in ($*$), then by affinity (verify!)

$$\begin{aligned}&\mathcal{L}(\theta_{k+1}) - ((1-\beta_k)\mathcal{L}(\theta^*) + \beta_k\mathcal{L}(\theta_k)) \\ &\le \frac{1}{2\eta}\left(\|\gamma_k - ((1-\beta_k)\theta^* + \beta_k\theta_k)\|^2 - \|\theta_{k+1} - ((1-\beta_k)\theta^* + \beta_k\theta_k)\|^2\right) \\ &= \frac{(1-\beta_k)^2}{2\eta}\left(\left\|\theta_k + \frac{\beta_k}{1-\beta_k}(\theta_k - \theta_{k-1}) - \theta^*\right\|^2 - \left\|\theta_{k+1} + \frac{\beta_k}{1-\beta_k}(\theta_{k+1} - \theta_k) - \theta^*\right\|^2\right)\end{aligned}$$

That is,

$$\begin{aligned}&\frac{\eta}{(1-\beta_k)^2}(\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta^*)) + \frac{1}{2}\left\|\theta_{k+1} + \frac{\beta_k}{1-\beta_k}(\theta_{k+1} - \theta_k) - \theta^*\right\|^2 \\ &\le \frac{\beta_k\eta}{(1-\beta_k)^2}(\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*)) + \frac{1}{2}\left\|\theta_k + \frac{\beta_k}{1-\beta_k}(\theta_k - \theta_{k-1}) - \theta^*\right\|^2 \\ &\le \frac{\eta}{(1-\beta_{k-1})^2}(\mathcal{L}(\theta_k) - \mathcal{L}(\theta^*)) + \frac{1}{2}\left\|\theta_k + \frac{\beta_k}{1-\beta_k}(\theta_k - \theta_{k-1}) - \theta^*\right\|^2\end{aligned}$$

Since

$$\frac{\beta_k}{(1-\beta_k)^2} = \frac{k(k+3)}{9} < \frac{(k+2)^2}{9} = \frac{1}{(1-\beta_{k-1})^2}.$$

Therefore, we have constructed a monotone sequence, and

$$\begin{aligned}&\frac{\eta}{(1-\beta_k)^2}(\mathcal{L}(\theta_{k+1}) - \mathcal{L}(\theta^*)) + \frac{1}{2}\left\|\theta_{k+1} + \frac{\beta_k}{1-\beta_k}(\theta_{k+1} - \theta_k) - \theta^*\right\|^2 \\ &\le \frac{\eta}{(1-\beta_0)^2}(\mathcal{L}(\theta_{-1}) - \mathcal{L}(\theta^*))) + \frac{1}{2}\|\theta_0 - \theta^*\|^2 \\ &\le \frac{1}{2}\|\theta_0 - \theta^*\|^2.\end{aligned}$$

□

Remark 2.2.19. *There are many ways to define suitable sequences β_k as long as*

$$\frac{\beta_k}{(1-\beta_k)^2} \le \frac{1}{(1-\beta_{k-1})^2}.$$

Thus, we can choose $\beta_k = \frac{k}{k+2}$ *and the convergence constant becomes* $\frac{2}{\eta(k+1)^2}$.

The use of NAG-c for strongly convex functions has been extensively studied over the past decade. Early results Su et al. (2016) show that the naive way of ODE formulation of NAG-c cannot converge exponentially fast, while the numerical experiments show exponential convergence. Later, the gap can be closed by adding higher-order correction terms in the ODE formulation.

2.2.3 Lyapunov Functions

In the continuum limit, the previous sections can be viewed from ODE perspectives. Consider the dynamical system

$$\frac{d\theta}{dt} = \phi(\theta(t)).$$

Let $V(\theta(t))$ be a **Lyapunov function** satisfying

$$V(\theta(t)) \geq V^\star,$$

where $V^\star$ is a lower bound. The time derivative of V along trajectories is

$$\frac{dV}{dt} = \left\langle \nabla V(\theta(t)), \frac{d\theta}{dt} \right\rangle = \langle \nabla V(\theta(t)), \phi(\theta(t)) \rangle \leq 0.$$

Theorem 2.2.20. *Assume* $V(\theta(t))$ *is bounded from below by* V^* *and*

$$-\frac{dV}{dt} \geq W(\theta(t)) \geq 0.$$

Then:

1. $\liminf_{t\to\infty} W(\theta(t)) = 0$,
2. *For any* $T > 0$,

$$\min_{0\leq t\leq T} W(\theta(t)) \leq \frac{1}{T}\big(V(\theta(0)) - V(\theta(T))\big) \leq \frac{1}{T}\big(V(\theta(0)) - V^*\big).$$

Proof. Since $W(\theta(t)) \leq -\frac{dV}{dt}$, integrating over $[0, T]$ gives

$$\int_0^T W(\theta(t))\, dt \leq V(\theta(0)) - V(\theta(T)) \leq V(\theta(0)) - V^\star.$$

Hence,

$$T \min_{0 \le t \le T} W(\theta(t)) \le V(\theta(0)) - V^\star,$$

which implies the result. Letting $T \to \infty$ yields $\liminf_{t \to \infty} W(\theta(t)) = 0$. □

The following theorem is a direct conclusion from convexity.

Theorem 2.2.21. *If $W(\theta)$ is convex, then*

$$W\big(\bar{\theta}_T\big) \le \frac{1}{T}\big(V(\theta(0)) - V^\star\big),$$

where

$$\bar{\theta}_T := \frac{1}{T}\int_0^T \theta(t)\, dt.$$

Theorem 2.2.22. *If there exists $\gamma > 0$ such that*

$$W(\theta(t)) \ge \gamma\big(V(\theta(t)) - V^\star\big),$$

then

$$V(\theta(t)) - V^\star \le \big(V(\theta(0)) - V^\star\big)e^{-\gamma t}.$$

Proof. We have

$$\frac{dV}{dt} \le -W(\theta(t)) \le -\gamma\big(V(\theta(t)) - V^\star\big).$$

The result follows from Grönwall's inequality. □

Next, we consider the specific example of gradient flow

$$\frac{d\theta}{dt} = -\nabla\mathcal{L}(\theta(t)).$$

Theorem 2.2.23. *If $\mathcal{L} \in C^1$ and $\mathcal{L}(\theta) \ge \mathcal{L}^\star$, then*

$$\min_{0 \le t \le T} \|\nabla\mathcal{L}(\theta(t))\|^2 \le \frac{1}{T}\big(\mathcal{L}(\theta(0)) - \mathcal{L}^\star\big).$$

Proof. Take

$$W(\theta(t)) = \|\nabla \mathcal{L}(\theta(t))\|^2, \qquad V(\theta(t)) = \mathcal{L}(\theta(t)).$$

Then

$$\frac{dV}{dt} = \langle \nabla \mathcal{L}(\theta(t)), \dot{\theta}(t) \rangle = -\|\nabla \mathcal{L}(\theta(t))\|^2,$$

Thus take

$$W(\theta(t)) := \|\nabla \mathcal{L}(\theta(t))\|^2.$$

and apply Theorem 2.2.20. □

Theorem 2.2.24. *If*

$$\|\nabla \mathcal{L}(\theta(t))\|^2 \geq 2\mu\big(\mathcal{L}(\theta(t)) - \mathcal{L}^\star\big),$$

then

$$\mathcal{L}(\theta(t)) - \mathcal{L}^\star \leq e^{-2\mu t}\big(\mathcal{L}(\theta(0)) - \mathcal{L}^\star\big).$$

Proof. Apply Theorem 2.2.22 (exponential decay under $W(\theta) \geq \gamma(V(\theta) - V^\star)$) with $\gamma = 2\mu$. □

Theorem 2.2.25. *If $\mathcal{L}$ is convex and $\mathcal{L}^\star = \mathcal{L}(\theta^\star)$, then*

$$\mathcal{L}(\bar{\theta}_T) - \mathcal{L}^\star \leq \frac{1}{T}\|\theta(0) - \theta^\star\|^2,$$

where

$$\bar{\theta}_T = \frac{1}{T}\int_0^T \theta(t)\, dt.$$

Proof. Let $V(\theta) = \|\theta - \theta^\star\|^2$. Then

$$\frac{dV}{dt} = 2\langle \theta(t) - \theta^\star, \dot{\theta}(t) \rangle = 2\langle \theta(t) - \theta^\star, -\nabla \mathcal{L}(\theta(t)) \rangle.$$

By convexity,

$$\mathcal{L}(\theta(t)) - \mathcal{L}^\star \leq \langle \nabla \mathcal{L}(\theta(t)), \theta(t) - \theta^\star \rangle.$$

Hence

$$\frac{dV}{dt} \leq -2\big(\mathcal{L}(\theta(t)) - \mathcal{L}^\star\big).$$

Thus

$$-\frac{dV}{dt} \geq W(\theta(t)), \qquad W(\theta(t)) := 2(\mathcal{L}(\theta(t)) - \mathcal{L}^\star).$$

Applying Theorem 2.2.21 yields

$$\mathcal{L}(\bar{\theta}_T) - \mathcal{L}^\star \leq \frac{1}{2T}\big(V(\theta(0)) - 0\big) = \frac{1}{2T}\|\theta(0) - \theta^\star\|^2.$$

□

Theorem 2.2.26. *If $\mathcal{L}$ is μ-strongly convex, then*

$$\mathcal{L}(\theta(t)) - \mathcal{L}^\star \leq e^{-\gamma t}\big(\mathcal{L}(\theta(0)) - \mathcal{L}^\star\big), \qquad \gamma = \frac{\mu}{2}.$$

Proof. Strong convexity implies

$$\mathcal{L}(\theta(t)) - \mathcal{L}^\star \geq \frac{\mu}{2}\|\theta(t) - \theta^\star\|^2.$$

Using the previous Lyapunov argument with

$$W(\theta(t)) = \mathcal{L}(\theta(t)) - \mathcal{L}^\star,$$

and applying Theorem 2.2.22 gives exponential convergence with $\gamma = \mu/2$. □

Next, we consider another example of the second-order system for Polyak's momentum method

$$\ddot{\theta} + a\dot{\theta} + b\nabla\mathcal{L}(\theta) = 0.$$

This can be written as the first-order system

$$\begin{cases} \dot{\theta} = -y, \\ \dot{y} = -ay - b\nabla\mathcal{L}(\theta). \end{cases}$$

Theorem 2.2.27. *Let $\mathcal{L} \in C^2$ and assume $\|\nabla^2\mathcal{L}(\theta)\| \leq L$ (i.e., $\mathcal{L}$ is L-smooth) and $\mathcal{L}(\theta(t)) \geq \mathcal{L}^\star$. Suppose $y(0) = 0$, then*

$$\min_{0\leq t\leq T} \|\nabla\mathcal{L}(\theta(t))\|^2 \leq \frac{1}{T}\frac{a^2 + 2bL}{ab}\big(\mathcal{L}(\theta(0)) - \mathcal{L}^*\big).$$

Proof. Let the Lyapunov function be defined as:

$$V(\theta(t)) = \mathcal{L}(\theta(t)) - \mathcal{L}^* + \frac{a}{a^2+2bL}\langle \nabla\mathcal{L}(\theta(t)), y\rangle + \frac{L}{a^2+2bL}\|y\|^2.$$

Then, using the property of L-smoothness where $\mathcal{L}(\theta) - \mathcal{L}^* \geq \frac{1}{2L}\|\nabla\mathcal{L}(\theta)\|^2$:

$$\begin{aligned} V(\theta(t)) &\geq \frac{1}{2L}\|\nabla\mathcal{L}\|^2 + \frac{a}{a^2+2bL}\langle \nabla\mathcal{L}(\theta(t)), y\rangle + \frac{L}{a^2+2bL}\|y\|^2 \\ &\geq c_1\left(\|\nabla\mathcal{L}(\theta(t))\|^2 + \|y\|^2\right). \end{aligned}$$

The time derivative is given by:

$$\begin{aligned} -\frac{dV}{dt} = \langle \nabla\mathcal{L}(\theta(t)), y\rangle &- \frac{a}{a^2+2bL}\langle \underbrace{\nabla^2\mathcal{L}(\theta(t))}_{\leq L} y, y\rangle \\ &- \frac{a}{a^2+2bL}\langle \nabla\mathcal{L}(\theta(t)), -ay - b\nabla\mathcal{L}\rangle \\ &- \frac{2L}{a^2+2bL}\langle y, -ay - b\nabla\mathcal{L}\rangle. \end{aligned}$$

Bounding the terms:

$$\begin{aligned} -\frac{dV}{dt} &\geq -\langle \nabla\mathcal{L}(\theta(t)), y\rangle - \frac{aL}{a^2+2bL}\|y\|^2 + \frac{a^2}{a^2+2bL}\langle \nabla\mathcal{L}(\theta(t)), y\rangle \\ &\quad + \frac{ab}{a^2+2bL}\|\nabla\mathcal{L}(\theta(t))\|^2 + \frac{2La}{a^2+2bL}\|y\|^2 + \frac{2bL\langle y, \nabla\mathcal{L}(\theta(t))\rangle}{a^2+2bL} \\ &= \frac{La}{a^2+2bL}\|y\|^2 + \frac{ab}{a^2+2bL}\|\nabla\mathcal{L}(\theta(t))\|^2 =: W(\theta(t)). \end{aligned}$$

From this, it follows that:

$$\min_{0\leq t\leq T} W(\theta(t)) \leq \frac{1}{T}\left(V(\theta(0)) - V^*\right)$$

□

Theorem 2.2.28. *Suppose $\mathcal{L}$ is convex, then*

$$\mathcal{L}(\overline{\theta}(T)) - \mathcal{L}^* \leq \frac{1}{T}\left(\mathcal{L}(\theta(0)) - \mathcal{L}^* + \frac{a^2}{2b}\left\|\theta(0) - \theta^* + \frac{1}{a}y(0)\right\|^2\right)$$

Proof. Let $V(\theta(t)) = \mathcal{L}(\theta(t)) - \mathcal{L}^* + \frac{a^2}{2b}\left\|\theta(t) - \theta^* + \frac{1}{a}y\right\|^2$. And we can check

$$\begin{aligned}
\frac{dV(\theta(t))}{dt} &= \langle \nabla\mathcal{L}(\theta(t)), \dot{\theta}\rangle + \frac{a^2}{b}\langle \theta(t) - \theta^* + \frac{1}{a}y, \dot{\theta} + \frac{1}{a}\dot{y}\rangle \\
&= \langle \nabla\mathcal{L}(\theta(t)), y\rangle + \frac{a^2}{b}\langle \theta(t) - \theta^* + \frac{1}{a}y, y + \frac{1}{a}(-ay - b\nabla\mathcal{L}(\theta(t)))\rangle \\
&= \langle \nabla\mathcal{L}(\theta(t)), y\rangle + \frac{a^2}{b}\langle \theta(t) - \theta^* + \frac{1}{a}y, -\frac{b}{a}\nabla\mathcal{L}(\theta(t))\rangle \\
&= \langle \nabla\mathcal{L}(\theta(t)), y\rangle - a\langle \theta(t) - \theta^*, \nabla\mathcal{L}(\theta(t))\rangle - \langle y, \nabla\mathcal{L}(\theta(t))\rangle \\
&= -a\langle \nabla\mathcal{L}(\theta(t)), \theta(t) - \theta^*\rangle
\end{aligned}$$

Thus, $W(\theta(t)) = a(\mathcal{L}(\theta(t)) - \mathcal{L}^*)$ and we use Theorem 2.2.21. □

Finally, we consider the Nesterov acceleration case that

$$\ddot{\theta} + \frac{a}{t}\dot{\theta} + b\nabla\mathcal{L}(\theta) = 0$$

Theorem 2.2.29. *Suppose $a \geq 3$, then*

$$\mathcal{L}(\theta(T)) - \mathcal{L}^* \leq \frac{(a-1)^2}{2b}\frac{\|(\theta(0) - \theta^*)\|^2}{T^2}$$

Proof. Let $V(\theta(t)) = t^2(\mathcal{L}(\theta(t)) - \mathcal{L}^*) + \frac{1}{2b}\|(a-1)(\theta(t) - \theta^*) + ty\|^2$. Then,

$$\begin{aligned}
\frac{dV(\theta(t))}{dt} &= 2t(\mathcal{L}(\theta(t)) - \mathcal{L}^*) + t^2\langle \nabla\mathcal{L}(\theta(t)), y\rangle \\
&\quad + \frac{1}{b}\langle (a-1)(\theta(t) - \theta^*) + ty, -bt\nabla\mathcal{L}(\theta(t))\rangle \\
&= 2t(\mathcal{L}(\theta(t)) - \mathcal{L}^*) - t\langle (a-1)(\theta(t) - \theta^*), \nabla\mathcal{L}(\theta(t))\rangle \\
&\leq t(3-a)(\mathcal{L}(\theta(t)) - \mathcal{L}^*).
\end{aligned}$$

Therefore, as long as $a \geq 3$, we can choose $W \equiv 0$ and use Theorem 2.2.20,

$$V(\theta(T)) \leq V(\theta(0)).$$

□

Remark 2.2.30. *This convergence bound matches the previous NAG-c's constant $\frac{2}{\eta(k+2)^2}$ if we take $a = 3$.*

2.2.4 Acceleration vs. Robustness

TODO

2.3 Stochastic Variants

Stochastic variants of training algorithms are essential in practice due to hardware limitations.

Extended Reading

Bertsekas, D. P. and Tsitsiklis, J. N. (2000). Gradient convergence in gradient methods with errors. *SIAM Journal on Optimization*, 10(3):627–642.

Ghadimi, E., Feyzmahdavian, H. R., and Johansson, M. (2015). Global convergence of the heavy-ball method for convex optimization. In *2015 European control conference (ECC)*, pages 310–315. IEEE.

Nesterov, Y. (1998). Introductory lectures on convex programming volume i: Basic course. *Lecture notes*, 3(4):5.

Nesterov, Y. et al. (2018). *Lectures on convex optimization*, volume 137. Springer.

Polyak, B. T. (1964). Some methods of speeding up the convergence of iteration methods. *Ussr computational mathematics and mathematical physics*, 4(5):1–17.

Su, W., Boyd, S., and Candes, E. J. (2016). A differential equation for modeling nesterov's accelerated gradient method: Theory and insights. *Journal of Machine Learning Research*, 17(153):1–43.