# Chapter 3 Gradient descent

- We assume existance of minimizer $\mathbf{x}^\star$.
- The goal is to find approximation $\mathbf{x}$, s.t. $f(\mathbf{x}) - f\left(\mathbf{x}^\star\right) < \varepsilon$.
  - no need for $\|\mathbf{x} - \mathbf{x}^\star\|$ to be close

### Notation on Norm

- $\|\cdot\|_{a*}$ is the dual norm, $\|u\|_{a*} = \sup_{v\neq 0} \frac{u^T v}{\|v\|_a} = \sup_{\|v\|_a=1} u^T v$
- This implies generalized Cauchy-Schwarz inequality $|u^T v| \leq \|u\|_{a*}\|v\|_a, \forall u, v$.
- For Eculidiean norm $\|\cdot\| = \|\cdot\|_* = \|\cdot\|_2$
- the parameter $L$ depends on choice of norm
- $\|x\|_2 \leq \|x\|_1 \leq \sqrt{n}\|x\|_2$, $\frac{1}{\sqrt{n}}\|x\|_2 \leq \|x\|_\infty \leq \|x\|_2$.

## Smoothness and Lipschitz

### (Gradient) Lipschitz

- Definition (Lipschitz) The gradient of $f$ is *Lipschitz continuous* with parameter $L > 0$ w.r.t. norm $\|\cdot\|_a$, if $\|\nabla f(x) - \nabla f(y)\|_{a*} \leq L\|x - y\|_a$.
- Lemma A (generalized Cauchy–Schwarz inequality) $\nabla f$ is $L$-Lipschitz $\Rightarrow \forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$, $(\nabla f(x) - \nabla f(y))^T (x - y) \leq L\|x - y\|_a^2$
  - Proof
    - $(\nabla f(x) - \nabla f(y))^T (x - y) \leq \|\nabla f(x) - \nabla f(y)\|_{a*}\|x - y\|_a \leq L\|x - y\|_a^2$

### Smoothness

- Definition 3.2 (Smoothness) Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be a differentiable function, $X \subset \mathbf{dom}(f)$ convex and $L \in \mathbb{R}_+$. Function $f$ is called *smooth with parameter* $L$ over $X$ (over norm $\|\cdot\|_a$) if $f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_a^2$
  - PS: NO need to assume $f$ to be convex
- Lemma 3.3 (equivalance of smoothness in 2-norm) $f$ is $L$-smooth over 2-norm $\Leftrightarrow g(\mathbf{x}) = \frac{L}{2}\mathbf{x}^\top\mathbf{x} - f(\mathbf{x})$ is convex over $\mathbf{dom}(f)$.
  - Proof
    - $\nabla g(\mathbf{x}) = L\mathbf{x} - \nabla f(\mathbf{x})$
    - $g$ convex $\Leftrightarrow \forall \mathbf{x}, \mathbf{y}, g(\mathbf{y}) \geq g(\mathbf{x}) + \nabla g(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) \Leftrightarrow \frac{L}{2}\mathbf{y}^\top\mathbf{y} - f(\mathbf{y}) \geq \frac{L}{2}\mathbf{x}^\top\mathbf{x} - f(\mathbf{x}) + [L\mathbf{x} - \nabla f(\mathbf{x})]^\top (\mathbf{y} - \mathbf{x}) \Leftrightarrow f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|_2^2$
  - If $f$ twice differentiable, Hessian of $g$ is $LI - \nabla^2 f \succeq \mathbf{0}$, this means $\forall x, \lambda_{\max}\left(\nabla^2 f(x)\right) \leq L$.
- Lemma B (quadratic upper bound) If $\mathbf{dom}\, f = \mathbf{R}^n$ and $f$ is $L$-smooth over norm $a$, and has minimizer $\mathbf{x}^\star$, then $\forall z$, $\frac{1}{2L}\|\nabla f(z)\|_{a*}^2 \leq f(z) - f\left(x^\star\right) \leq \frac{L}{2}\|z - x^\star\|_a^2$
  - Proof
    - Right-hand inequality by smoothness setting $y = z, x = x^\star$, $f(\mathbf{z}) \leq f(\mathbf{x}^\star) + \underbrace{\nabla f(\mathbf{x}^\star)}^\top (\mathbf{z} - \mathbf{x}^\star) + \frac{L}{2}\|\mathbf{x}^\star - \mathbf{z}\|_a^2$ $=0$
    - Left-hand ineuqality, by smoothness $\inf_y f(y) \leq \inf_y \left(f(z) + \nabla f(z)^T (y - z) + \frac{L}{2}\|y - z\|^2\right)$
      - by seperation of direction and magnitude LHS $\leq= \inf_{\|v\|=1} \inf_t \left(f(z) + t\nabla f(z)^T v + \frac{Lt^2}{2}\right) = \inf_{\|v\|=1} \left(f(z) - \frac{1}{2L}\left(\nabla f(z)^T v\right)^2\right)$ by maximum of quadratic function.
      - By the definition of dual norm, we have $\inf_y f(y) = f(z^\star) \leq f(z) - \frac{1}{2L}\|\nabla f(z)\|_*^2$
- Lemma 3.4 (quadratic function) quadratic form $f(\mathbf{x}) = \mathbf{x}^\top Q\mathbf{x} + \mathbf{b}^\top\mathbf{x} + c$ is $2\|Q\|_{\text{spec}}$-smooth over 2-norm.
- Example of Non-Smooth func: $f(x) = x^4$ since $\forall L, y^4 \leq \frac{L}{2}y^2$ does not hold for all $y$ near $x = 0$.
- Lemma 3.6 (Operation that preserves smoothness)

- (i) *Linear combination* $f := \sum_{i=1}^m \lambda_i f_i$, where $\lambda_i > 0$.
- (ii) *Affine Transform* $f(A\mathbf{x} + \mathbf{b}))$.
- Proof Straightforward

## Equivalence of Smoothness and Gradient Lipschitz under Convexity

- Definition (co-coercivity) The property of *co-coercivity* of $\nabla f$ with parameter $L$ over norm $a$ is $\forall x, y$, $(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{L} \|\nabla f(x) - \nabla f(y)\|_*^2$
- Lemma 3.5 (Equivalence of Smoothness and Lipschitz under convexity) If (i) $\mathbf{dom}(f) = \mathbb{R}^d$ (ii) $f$ convex and differentiable. Then $f$ is smooth with parameter $L$ under norm $a \Leftrightarrow \nabla f$ is Lipschitz continuous with parameter $L$ under norm $a$.
  - *Proof Route* Lipschitzness -(*Lemma A*)-> Smoothness -(*Lemma B*)-> Co-coercivity -> Lipschitzness
  - Proof
    - Lipschitzness -(*Lemma A*)-> Smoothness
      - Define $g(t \in [0,1]) = f(x + t(y - x))$, (well-defined because of convexity of $\mathbf{dom}(f)$)
      - $g'(t) - g'(0) = (\nabla f(x + t(y - x)) - \nabla f(x))^T (y - x)$
      - Since $f$ $L$-Lipschitz, by Lemma A, we have $\forall \mathbf{x}, \mathbf{y} \in \mathbf{dom}(f)$, $(\nabla f(x) - \nabla f(y))^T (x - y) \leq L \|x - y\|_a^2$
        - so that $g'(t) - g'(0) \leq tL\|x - y\|_a^2$
      - $f(y) = g(1) = g(0) + \int_0^1 g'(t)dt \leq g(0) + g'(0) + \frac{L}{2}\|x - y\|^2 = f(x) + \nabla f(x)^T (y - x) + \frac{L}{2}\|x - y\|^2$
      - We have smoothness
    - Smoothness -(*Lemma B*)-> Co-coercivity
      - Define two function $f_x(z) = f(z) - \nabla f(x)^T z, f_y(z) = f(z) - \nabla f(y)^T z$
      - Since we only add first-order term, $f_x z, f_y(z)$ are still convex.
      - Since $\nabla f_x(z_1) - \nabla f_x(z_2) = \nabla f(z_1) - \nabla f(z_2)$, $f_x z, f_y(z)$ are still $L$-smooth.
      - Also $\nabla f_x(z = x) = 0$, by convexity, it reaches it minimum, using Smoothness and left-hand inequality of Lemma B, and taking $f = f_x, z = y$, we get $f(y) - f(x) - \nabla f(x)^T (y - x) \geq \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_{a*}^2$
      - Similarly $f(x) - f(y) - \nabla f(y)^T (x - y) \geq \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|_{a*}^2$
      - Adding these together and we get $(\nabla f(x) - \nabla f(y))^T (x - y) \geq \frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_{a*}^2$, co-coercivity
    - Co-coercivity -> Lipschitzness
      - $\frac{1}{L}\|\nabla f(x) - \nabla f(y)\|_{a*}^2 \leq (\nabla f(x) - \nabla f(y))^T (x - y) \leq \|\nabla f(x) - \nabla f(y)\|_{a*} \cdot \|x - y\|_a$
      - $\Rightarrow \|\nabla f(x) - \nabla f(y)\|_{a*} \leq L\|x - y\|_a$

## Smoothness $\neq$ Lipschitz in $f$ itself

- Counter Examples
  - Lipschitz $\not\Rightarrow$ Smoothness $f(x) = \begin{cases} |x|^{3/2}, & |x| \leq 1 \\ \frac{3}{2}|x| - \frac{1}{2}, & |x| \geq 1 \end{cases}$, $f$ is $3/2$-Lipschitz but $\infty$-smooth.
  - Smoothness $\not\Rightarrow$ Lipschitz $f(x) = x^2, x \in (-\infty, \infty)$, $f$ is $2$-smooth, but $\infty$-Lipschitz.

## Strong convexity

- Definition and Lemma C (Strong Convexity) A differentiable function $f$ is strongly convex with parameter $\mu$ over norm $\|\cdot\|_a$ if $\mathbf{dom}(f)$ is convex and (for following are equivalent)
  - (i) $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\mu}{2}\|y - x\|_a^2, \forall x, y$ (This is also used in text book)
  - (ii) $f(\theta x + (1 - \theta)y) \leq \theta f(x) + (1 - \theta) f(y) - \frac{\mu}{2}\theta(1 - \theta)\|x - y\|_a^2, \theta \in [0, 1]$
  - (iii) $\forall x, y \in \mathbf{dom}(f), g(t) := f(x + t(y - x)) - \frac{m}{2}t^2\|x - y\|_a^2$ is convex over $t$
  - (iiii) $(\nabla f(x) - \nabla f(y))^T (x - y) \geq \mu\|x - y\|_a^2$ (also known as *strong monotonicity* or *coercivity* of $\nabla f$)
  - Proof of Equivalence
    - (i) -> (ii)
      - set $y = \theta x + (1 - \theta)y, x = x$ and $y = \theta x + (1 - \theta)y, x = y$, we get
        - (a) $f(\theta x + (1 - \theta)y) \geq f(x) + (1 - \theta)\nabla f(x)^T (y - x) + (1 - \theta)^2 \frac{\mu}{2}\|y - x\|_a^2$ and
        - (b) $f(\theta x + (1 - \theta)y) \geq f(y) - \theta \nabla f(y)^T (y - x) + \theta^2 \frac{\mu}{2}\|y - x\|_a^2$
      - $\theta \times$(a)$+(1 - \theta) \times$(b) we get (ii)
    - (ii) -> (iii)
      - Consider function $h(t) := f(x + t(y - x))$, by (ii) we have $h(\theta t_1 + (1 - \theta)t_2) \leq \theta h(t_1) + (1 - \theta)h(t_2) - \frac{\mu}{2}\theta(1 - \theta)(t_1 - t_2)^2\|y - x\|_a^2$
      - Since $\theta(1 - \theta)(t_1 - t_2)^2 = \theta t_1^2 + (1 - \theta)t_2^2 - (\theta t_1 + (1 - \theta)t_2)^2$, we have equivalently

        - $h(\theta t_1 + (1-\theta)t_2 \frac{\mu}{2}\|y-x\|_a^2) - (\theta t_1 + (1-\theta)t_2)^2 \leq \theta(h(t_1) - t_1^2\frac{\mu}{2}\|y-x\|_a^2) + (1-\theta)(h(t_2) - t_2^2\frac{\mu}{2}\|y-x\|_a^2)$
    - or equivalently, define $g(t) := h(t) - t^2\frac{\mu}{2}\|y-x\|_a^2$, $g$ is a convex function (along this domain of finite line between $x, y$)
        - if $f$ differentiable, then $g$ also
- PS: (iii) -> (ii) straightforward by setting $t = \theta$, and compare with $t = 0, 1$.
- (iii) -> (i)
    - The first order condition of $g$ convex gives $g(1) \geq g(0) + g'(0)$
        - $\Rightarrow f(y) - \frac{\mu}{2}\|y-x\|_a^2 \geq f(x) + \nabla f(x)^\top (y-x) - \mu\|y-x\|_a^2$, this is (i)
- (i) -> (iiii) switch between $x, y$ and add them together.
- (iiii) -> (iii)
    - let $x = x, y = x + t(y-x)$ in (iiii), and we have $(\nabla f(x + t(y-x)) - \nabla f(x))^T (y-x) \geq \mu t \|x-y\|^2$
    - now we compute $g'(t)$ in (iii), we have $g'(t) = \nabla f(x + t(y-x))^\top (y-x) - \mu t \|y-x\|_a^2$
    - assume $t_1 \geq t_2$, then $g'(t_1) - g'(t_2) = \nabla (f(x + t_1(y-x)) - f(x + t_2(y-x)))^\top (y-x) - \mu(t_1 - t_2)\|y-x\|_a^2$
      $\geq \mu(t1 - t2)\|x-y\|_a^2 - \mu(t_1 - t_2)\|y-x\|_a^2 = 0$
    - $g'(t)$ nondecreasing, convex -> (iii).

- **Lemma D (boundedness and global minimum)** If $f$ is differentiable and $\mu$-strongly convex, then $f$ have bounded sublevel.
    - **Proof**
        - Since $f$ $\mu$-strongly convex, given any $x, y$, value along this line will go to $+\infty$ by quadratic function.
        - suppose $f^{\leq \alpha}$ non-empty, since $f$ convex -> continuous, then $\exists y$ s.t. $f(y) = \alpha$.
        - Then $\forall x, f(x) \geq f(y) + \nabla f(y)^\top (y-x) + \frac{\mu}{2}\|y-x\|_a^2 \geq f(y) - \|\nabla f(y)\|_{a*}\|y-x\| + \frac{\mu}{2}\|y-x\|_a^2$,
        - If $\|y-x\|_a \geq 2\|\nabla f(y)\|_{a*}/\mu$, then $f(x) \geq f(y) = \alpha$, then $f^{\leq \alpha}$ bounded.
    - This means $\inf_{x \in \mathbf{dom}(f)} f > -\infty$, if further one of $f^{\leq \alpha}$ is closed, then $f$ have a global minimum.
    - The global minimum is also unique (trivial).

- **Lemma E (quadratic lower bound)** If $f$ is closed (means has closed sublevel sets), and $\mathbf{x}^\star$ is its unique minimizer, then $\forall z \in \mathbf{dom}(f), \frac{m}{2}\|z - x^\star\|^2 \leq f(z) - f(x^\star) \leq \frac{1}{2m}\|\nabla f(z)\|_*^2$.
    - **Proof** similar to smoothness case.

- **Lemma 3.11 (equivalance of strong convexity in 2-norm)** $f$ $\mu$-strong convex over 2-norm iff $k(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ is convex
    - **Proof** similar to smoothness case.
    - If $f$ twice differentiable, Hessian of $h$ is $\nabla^2 f - \mu I \succeq 0$, this means $\forall x, \lambda_{\min}\left(\nabla^2 f(x)\right) \geq \mu$.
        - Strictly Convex.

## Smooth and Strong convexity

- **Lemma F (Extension of co-coercivity)** Suppose $f$ is $\mu$-strongly convex and $L$-smooth for 2-norm, and $\mathbf{dom}(f) = \mathbb{R}^n$
    - then function $h(x) = f(x) - \frac{\mu}{2}\|x\|_2^2$ is $L - \mu$-smooth
    - co-coercivity of $\nabla h$ gives $\forall x, y, (\nabla f(x) - \nabla f(y))^T (x-y) \geq \frac{\mu L}{\mu + L}\|x-y\|_2^2 + \frac{1}{\mu+L}\|\nabla f(x) - \nabla f(y)\|_2^2$

## Speed metric

- $\lim_{k\to\infty} \frac{f(w^{k+1}) - f(w^*)}{(f(w^k) - f(w^*))^p} = q$
    - $p = 1, q \in (0, 1)$, *linear* rate. E.g. $\Delta f_t = O(e^{-\alpha t}), \alpha > 0$.
    - $p = 1, q = 1$, *sublinear* rate. E.g. $\Delta f_t = O(t^{-\beta}), \beta > 0$.
    - $p = 1, q = 0$, *super linear* rate. E.g. $\Delta f_t = O(e^{-\alpha t^2}), \alpha > 0$.
    - $p > 1, q > 0$, *convergence of order $p$*. E.g. $\Delta f_t = O(e^{-\alpha p^t}), \alpha > 0$.
        - when $p = 2$, *quadratic convergence*.

## Vanilla GD

- Update: $\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \mathbf{g}_t$, $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$
- Consider quantity $\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{2\gamma}\left(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right)$
  $= \frac{\gamma}{2}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\left(\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right)$
- sum over $t = 0, \ldots, T-1$, we have
  $\sum_{t=0}^{T-1} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\left(\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \|\mathbf{x}_T - \mathbf{x}^\star\|^2\right) \leq= \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$

- By convexity $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \le \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star)$, we have $\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \le \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$
  - This gives upper bound on average error $\mathbb{E}_t[f(\mathbf{x}_t) - f(\mathbf{x}^\star)]$
  - note that last iterate is not necessarily the best one. Some algorithms guarantee last iterate convergence.
  - Bad result since we cannot control $\|\mathbf{g}_t\|$

## Improvement Case 1: Lipschitz $f$: $\mathcal{O}\left(1/\varepsilon^2\right)$ Steps

- By Theorem 2.9, if we assume convex function $f$ is $B$-Lipschitz, then $\|Df(x)\| \le B$, the spectual norm in Thm 2.9 is Euclidian norm when $Df \in \mathbb{R}^{1\times d}$
  - then $\|\mathbf{g}_t\| \le B$ -> Theorem 3.1
- Theorem 3.1 Suppose $f \in C^1(x)$ convex, with global minimum $\mathbf{x}^\star$. If (1) $\|\mathbf{x}_0 - \mathbf{x}^\star\| \le R$, (2) $\forall \mathbf{x}, \|\nabla f(\mathbf{x})\| \le B$, when with step size $\gamma := \frac{R}{B\sqrt{T}}$, we have average error bounded by $\frac{1}{T}\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \le \frac{RB}{\sqrt{T}}$
  - Proof Straighforward from vanilla case, $\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \le \frac{\gamma}{2}B^2T + \frac{1}{2\gamma}R^2$
  - Time Complexity To achieve $\min_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \le \varepsilon$, we need $T \ge \frac{R^2B^2}{\varepsilon^2}$ iterations, $\mathcal{O}\left(1/\varepsilon^2\right)$.
  - but no dependence on dimension $d$.

## Improvement Case 2: Smooth Convex $f$: $\mathcal{O}\left(1/\varepsilon\right)$ Steps

- Lemma 3.7 (Sufficient Descent) Suppose $f: \mathbb{R}^d \to \mathbb{R} \in C^1$ is $L$-smooth. Setting $\gamma := L^{-1}$, then gradient descent gives $f(\mathbf{x}_{t+1}) \le f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$. (Proof Straightforward)
- Lemma 3.8 (Last Iteration Bound for smooth convex funciton) Suppose $f: \mathbb{R}^d \to \mathbb{R} \in C^1$ is $L$-smooth with global minimum $\mathbf{x}^\star$. With step size $\gamma := L^{-1}$ gradient descent gives $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$.
  - Proof
    - Sum *sufficient descent* over $t = 0 : T-1$, we get $\frac{1}{2L}\sum_{t=0}^{T-1}\|\nabla f(\mathbf{x}_t)\|^2 \le \sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) = f(\mathbf{x}_0) - f(\mathbf{x}_T)$
    - Take this into vanilla GD bound, we get $\sum_{t=1}^{T}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \le \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$.
    - By sufficient descent, monotocity of $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ is guaranteed, so $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \le \frac{1}{T}\sum_{t=1}^{T}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \le \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$
  - Time Complexity $T \ge \frac{R^2L}{2\varepsilon} = \mathcal{O}(1/\varepsilon)$

## Improvement Case 3: Acceleration for smooth $f$: $\mathcal{O}(1/\sqrt{\varepsilon})$ Steps

- Algorithm
  - $\mathbf{y}_{t+1} := \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t)$
  - $\mathbf{z}_{t+1} := \mathbf{z}_t - \frac{t+1}{2L}\nabla f(\mathbf{x}_t)$
  - $\mathbf{x}_{t+1} := \frac{t+1}{t+3}\mathbf{y}_{t+1} + \frac{2}{t+3}\mathbf{z}_{t+1}$
- Theorem (Nesterov Acceleration) Suppose $f \in C^1(\mathbb{R}^d \to \mathbb{R})$ is convex and $L$-smooth with global minimum $\mathbf{x}^\star$. Then Nesterov's Accelerated gradient descent algorithm gives $f(\mathbf{y}_T) - f(\mathbf{x}^\star) \le \frac{2L\|\mathbf{z}_0 - \mathbf{x}^\star\|_2^2}{T(T+1)}$.
  - Proof (*potential function argument*)
    - Define potential function $\Phi(t) := t(t+1)(f(\mathbf{y}_t) - f(\mathbf{x}^\star)) + 2L\|\mathbf{z}_t - \mathbf{x}^\star\|_2^2$, we aim to show that $\Phi(T) \le \Phi(0)$, then we can prove this theorem.
    - Define $\Delta := \frac{\Phi(t+1) - \Phi(t)}{t+1}$, and we need to show that it's less than zero
      - $\Delta = \frac{1}{t+1}\left[(t+2)(t+1)(f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star)) - t(t+1)(f(\mathbf{y}_t) - f(\mathbf{x}^\star)) + 2L(\|\mathbf{z}_{t+1} - \mathbf{x}^\star\|_2^2 - \|\mathbf{z}_t - \mathbf{x}^\star\|_2^2)\right]$
      - $= t(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)) + 2(f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star)) + \frac{2L}{t+1}\left(\|\mathbf{z}_{t+1} - \mathbf{x}^\star\|_2^2 - \|\mathbf{z}_t - \mathbf{x}^\star\|_2^2\right)$
      - Since $\mathbf{g}_t = \frac{2L}{t+1}(\mathbf{z}_t - \mathbf{z}_{t+1})$, we consider $\mathbf{g}_t^\top(\mathbf{z}_t - \mathbf{x}^\star) = \frac{2L}{t+1}(\mathbf{z}_t - \mathbf{z}_{t+1})^\top(\mathbf{z}_t - \mathbf{x}^\star)$
      - by $v^\top w = \frac{\|v\|_2^2 + \|w\|_2^2 - \|v-w\|_2^2}{2}$, we have $\mathbf{g}_t^\top(\mathbf{z}_t - \mathbf{x}^\star) = \frac{t+1}{4L}\|\mathbf{g}_t\|_2^2 + \frac{L}{t+1}\left(\|\mathbf{z}_t - \mathbf{x}^\star\|_2^2 - \|\mathbf{z}_{t+1} - \mathbf{x}^\star\|_2^2\right)$
      - plug this into $\Delta$ we get $\Delta = t(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)) + 2(f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star)) - 2\mathbf{g}_t^\top(\mathbf{z}_t - \mathbf{x}^\star) + \frac{t+1}{2L}\|\mathbf{g}_t\|_2^2$
      - By sufficient descent of $y_{t+1}$, $f(y_{t+1}) \le f(x_t) - \frac{1}{2L}\|\mathbf{g}_t\|_2^2$, we replace $f(y_{t+1})$ w.r.t $f(x_t)$ and ommit term $-\frac{1}{2L}\|\mathbf{g}_t\|_2^2$
        - $\Delta \le t(f(\mathbf{x}_t) - f(\mathbf{y}_t)) + 2(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) - 2\mathbf{g}_t^\top(\mathbf{z}_t - \mathbf{x}^\star)$
        -
      - By convexity, $\Delta \le t\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{y}_t) + 2\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) - 2\mathbf{g}_t^\top(\mathbf{z}_t - \mathbf{x}^\star) = \mathbf{g}_t^\top[(t+2)\mathbf{x}_t - t\mathbf{y}_t - 2\mathbf{z}_t] = 0$ by definition of $\mathbf{x}_t$ .

## Improvement Case 4: Smooth and Strongly convex $f$: $\mathcal{O}(\log(1/\varepsilon))$ Steps, linear rate

- Theorem 3.14 (Linear rate for smooth and strongly convex function) Suppose $f \in C^1(\mathbb{R}^d \to \mathbb{R})$ is $L$-smooth and $\mu$-strongly convex with global minimum $\mathbf{x}^\star$. Choosing step size of $\gamma := L^{-1}$, then gradient descent with arbitrary $\mathbf{x}_0$,
  - (i) Geometric decrease for $\|\mathbf{x}_t - \mathbf{x}^\star\|^2$, $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)\|\mathbf{x}_t - \mathbf{x}^\star\|^2$
  - (ii) Exponential decrease for absoulute error $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$
  - Proof
    - By storng convexity, $\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) = \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^\star) \geq f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2$
      - again by the fact $\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{\gamma}(\mathbf{x}_t - \mathbf{x}_{t+1})^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{2\gamma}(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2)$
      - so that $f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{1}{2\gamma}(\gamma^2\|\mathbf{g}_t\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2) - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2$
      - Equivalently, $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq (1 - \mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2 + \gamma^2\|\mathbf{g}_t\|^2 - 2\gamma(f(\mathbf{x}_t) - f(\mathbf{x}^\star))$
    - By sufficient decrease, $f(\mathbf{x}^\star) - f(\mathbf{x}_t) \leq f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t) \leq -\frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2$
      - so that $\gamma^2\|\mathbf{g}_t\|^2 - 2\gamma(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq 0$
    - Therefore, $\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq (1 - \mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2 = (1 - \frac{\mu}{L})\|\mathbf{x}_t - \mathbf{x}^\star\|^2$
    - Iteratively, $\|\mathbf{x}_T - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$
    - By Strong convexity, $f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \nabla f(\mathbf{x}^\star)^\top (\mathbf{x}_T - \mathbf{x}^\star) + \frac{L}{2}\|\mathbf{x}_T - \mathbf{x}^\star\|^2 = \frac{L}{2}\|\mathbf{x}_T - \mathbf{x}^\star\|^2$.
  - Time Complexiy $T \geq \frac{L}{\mu}\ln\left(\frac{R^2 L}{2\varepsilon}\right) = \mathcal{O}(\log(1/\varepsilon))$