Gradient Descent:

$$x_{k+1} = x_k - \frac{1}{L} \nabla f(x_k)$$

L — f is L-smooth & convex

Thm: If $f(x)$ is convex and L-smooth, then for Gradient Descent with step-size $1/L$, we have

$$f(x_{k+1}) - f^* \leq \frac{L \|x_0 - x^*\|^2}{2(k+1)} = \underbrace{O\left(\frac{1}{k}\right)}_{\varepsilon}$$

$O\left(\frac{1}{\varepsilon}\right)$ iterations

Proof:

$$f(x_{k+1}) \leq f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

[$\because f$ is L-smooth]

$$= \underbrace{f(x_k) + \nabla f(x_k)^T (x - x_k)}_{\leq f(x)} + \nabla f(x_k)^T (x_{k+1} - x) + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$$f(x_{k+1}) \leq f(x) + \nabla f(x_k)^T (x_{k+1} - x) + \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$x_{k+1} = x_k - \frac{1}{L}\nabla f(x_k)$ [$\because$ GD with step-size $1/L$]

$$\nabla f(x_k) = L(x_k - x_{k+1})$$

$$f(x_{k+1}) \leq f(x) + \frac{L}{2}\|x_k - x\|^2 - \frac{L}{2}\|x_{k+1} - x\|^2 \quad \text{——— (1)}$$

When $x = x_k$:

$$f(x_{k+1}) \leq f(x_k) - \frac{L}{2}\|x_{k+1} - x_k\|^2$$

$$f(x_{k+1}) < f(x_k) \quad \text{[decreasing seq}^n\text{]}$$

When $x = x^*$

$$f(x_{k+1}) \le f^* + \frac{L}{2}\|x_k - x^*\|^2 - \frac{L}{2}\|x_{k+1} - x^*\|^2$$

$$\left(f(x_{k+1}) - f^*\right) \le \frac{L}{2}\|x_k - x^*\|^2 - \frac{L}{2}\|x_{k+1} - x^*\|^2$$

$$\sum_{k=0}^{K}\left(f(x_{k+1}) - f^*\right) \le \frac{L}{2}\|x_0 - x^*\|^2 - \frac{L}{2}\|x_{K+1} - x^*\|^2$$

$$\le \frac{L}{2}\|x_0 - x^*\|^2$$

Since $f(x_k)$ is decreasing

$$\sum_{k=0}^{K}\left(f(x_{k+1}) - f^*\right) \ge (K+1)\left(f(x_{K+1}) - f^*\right)$$

$$(K+1)\left(f(x_{K+1}) - f^*\right) \le \frac{L}{2}\|x_0 - x^*\|^2$$

$$\boxed{\left(f(x_{K+1}) - f^*\right) \le \frac{L}{2(K+1)}\|x_0 - x^*\|^2} \quad (*)$$

$$\lim_{k \to \infty} \frac{|f(x_{k+1}) - f^*|}{|f(x_k) - f^*|^q} = \rho \rightarrow \text{rate of convergence}$$

$\rho = 1$ sublinear convergence

$\rho \in (0,1)$ linear convergence

$\rho = 0$ superlinear convergence

* For GD [f is L-smooth & convex]

$$\frac{f(x_{k+1}) - f^*}{f(x_k) - f^*} = \frac{k}{k+1}$$

$$\lim_{k\to\infty} = 1 \quad \text{(Sublinear convergence)}$$

Algorithm	L-smooth + convex
GD	$O\left(\frac{1}{\varepsilon}\right)$

* GD for strongly convex functions:

Assumptions: (i) f is L-smooth
(ii) f is convex.
(iii) f satisfies PL-inequality.

$$\frac{1}{2\mu}\|\nabla f(x)\|^2 \geq f(x) - f^*$$

f is μ-SC.

* Acceleration under strong convexity:

Thm: If f is L-smooth and μ-SC, then for GD with step-size $1/L$, we have

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^{k+1} \frac{L}{2}\|x_0 - x^*\|^2$$

$$\lim_{k\to\infty} \frac{f(x_{k+1}) - f^*}{f(x_k) - f^*} = \left(1 - \frac{\mu}{L}\right) < 1 \quad \text{[Linear rate of convergence]}$$

Proof: $\Rightarrow$ $f(x_{k+1}) \leq f(x_k) + \underline{\nabla f(x_k)^T}(x_{k+1} - x_k) + \frac{L}{2}\|x_{k+1} - x_k\|^2$

[$\because f$ is L-smooth]

$x_{k+1} - x_k = -\frac{1}{L}\nabla f(x_k)$ [GD]

$$f(x_{k+1}) \leq f(x_k) - \frac{1}{L}\|\nabla f(x_k)\|^2 + \frac{1}{2L}\|\nabla f(x_k)\|^2$$

$$= f(x_k) - \frac{1}{2L}\underbrace{\|\nabla f(x_k)\|^2}$$

$$= f(x_k) - \frac{\mu}{L} \cdot \frac{1}{2\mu}\|\nabla f(x_k)\|^2$$

$$f(x_{k+1}) \leq f(x_k) - \frac{\mu}{L}\left(f(x_k) - f^*\right) \quad \text{[PL-inequality]}$$

$$f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)\left(f(x_k) - f^*\right)$$

$$\boxed{f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^{k+1}\left(f(x_0) - f^*\right)} \text{ --- } 1$$

$f(y) \leq f(x) + \nabla f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2$ [$\because f$ is L-smooth]

$y = x_0$

$x = x^*$

$$\left(f(x_0) - f^*\right) \leq \frac{L}{2}\|x_0 - x^*\|^2$$

$$\boxed{f(x_{k+1}) - f^* \leq \left(1 - \frac{\mu}{L}\right)^{k+1}\frac{L}{2}\|x_0 - x^*\|^2}$$

$$O\Big(\underbrace{\left(1 - \frac{\mu}{L}\right)^k}_{\varepsilon}\Big)$$

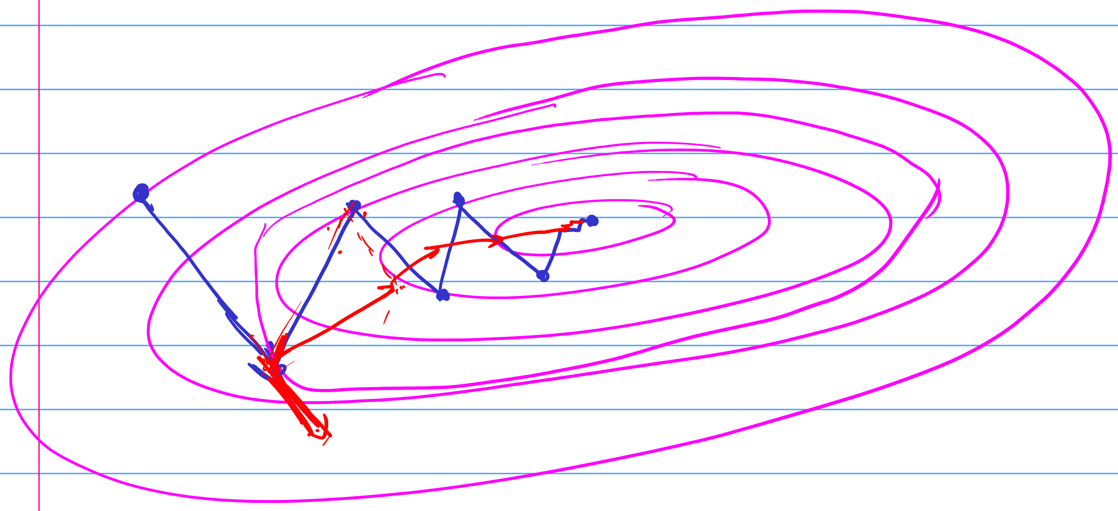
$$\left(1 - \frac{\mu}{L}\right)^k = \varepsilon$$

$$k \underbrace{\log\left(1 - \frac{\mu}{L}\right)} = \log \varepsilon$$

$$k\left(-\frac{\mu}{L}\right) = \log \varepsilon \Rightarrow \boxed{k = \frac{L}{\mu}\log\left(\frac{1}{\varepsilon}\right)}$$

# of iterations $O\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$

Algorithm	convex f	SC	
→ GD	$O\left(\frac{1}{\varepsilon}\right)$	$O\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$	} No of iterations
	Sublinear	Linear	



Q: Can we accelerate the convergence even further?

(Yes)

↳ by using momentum

60s ↳ Polyak's Heavy-Ball Method (HB)

80s ↳ Nesterov's Accelerated Gradient Method (AGD)

* Heavy-Ball Method:

$$y_k = \beta(x_k - x_{k-1})$$

$$x_{k+1} = x_k - \alpha \nabla f(x_k) + y_k$$

"Momentum"

* Convergence rate for HB Method:

(i) f is μ-SC (and L-smooth):

# of iterations $O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$

(ii) If f is simply convex (and L-smooth),

# of iterations $O\left(\frac{1}{\varepsilon}\right)$

Algorithms	f is convex	f is SC
GD	$O(1/\varepsilon)$	$O\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$
HB	$O(1/\varepsilon)$	$O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$

HB method improves acceleration for SC case.

* AGD:

Adaptive

$$\boxed{\begin{aligned} y_k &= x_k + \frac{\theta_k(1-\theta_{k-1})}{\theta_{k-1}}(x_k - x_{k-1}) \\ x_{k+1} &= y_k - \frac{1}{L}\nabla f(y_k) \end{aligned}}$$

$$\theta_k = \frac{\sqrt{\theta_{k-1}^4 + 4\theta_{k-1}^2} - \theta_{k-1}^2}{2} \; ; \quad x_0 = x_{-1}, \; \theta_0 = 1$$

* For SC:

$$\boxed{\begin{aligned} y_k &= x_k + \left(\frac{\sqrt{L}-\sqrt{\mu}}{\sqrt{L}+\sqrt{\mu}}\right)(x_k - x_{k-1}) \\ x_{k+1} &= y_k - \frac{1}{L}\nabla f(y_k) \end{aligned}}$$

$$f(x_{k+1}) - f^* \leq O\left(\left(1-\sqrt{\frac{\mu}{L}}\right)^k\right)$$

Number of iterations. $\longrightarrow$ $O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$

(ii) Convex case: $O\left(\sqrt{\frac{1}{\varepsilon}}\right)$ iterations

Algorithms	Convex f	SC - f
GD	$O\left(\frac{1}{\varepsilon}\right)$	$O\left(\frac{L}{\mu}\log\frac{1}{\varepsilon}\right)$
HB	$O\left(\frac{1}{\varepsilon}\right)$	$O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$
✓ AGD	$O\left(\sqrt{\frac{1}{\varepsilon}}\right)$	$O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$
Optimal	$O\left(\sqrt{\frac{1}{\varepsilon}}\right)$	$O\left(\sqrt{\frac{L}{\mu}}\log\frac{1}{\varepsilon}\right)$

* **Thm:** If f is L-smooth, then AGD requires $O\left(\sqrt{\frac{L}{\varepsilon}}\right)$ iterations

i.e.,

$$f(x_{k+1}) - f^* \leq O\left(\frac{1}{k^2}\right)$$

→ f is convex.

GD: $x_{k+1} = x_k - \eta \nabla f(x_k)$

$$\lim_{\eta \to 0} \frac{x_{k+1} - x_k}{\eta} = -\nabla f(x_k)$$

$$\boxed{\dot{x} = -\nabla f(x)} \longrightarrow \text{Gradient flow}$$

$$f(x) = \frac{1}{2}x^2 \qquad \nabla f(x) = x$$

$$\boxed{\dot{x} = -x} \longrightarrow \text{Exponential convergence guarantees}$$

$$\dot{x} = -10x \longrightarrow x(t) = e^{-10t}x_0$$

$$\dot{x} = -x \longrightarrow x(t) = e^{-t}x_0$$

$$\dot{x} = -\frac{\nabla f(x)}{\|\nabla f(x)\|} \qquad \dot{x} = -\operatorname{sgn}(x)$$