## AI709 Presentation:

# Convex and Non-convex Optimization under Generalized Smoothness

**Haochuan Li*, Jian Qian*, Yi Tian, Alexander Rakhlin, Ali Jadbababie**

**NeurIPS 2023 (Spotlight)**

**Speaker: Hanseul Cho**

# Classical Analyses of Optimization Algorithms

## Under Lipschitz smoothness

- Unconstrained optimization $\min_{x \in \mathbb{R}^d} f(x)$ with first-order algorithms

# Classical Analyses of Optimization Algorithms

## Under Lipschitz smoothness

- Unconstrained optimization $\min_{x \in \mathbb{R}^d} f(x)$ with first-order algorithms
- Classical textbook analyses [Nemirovskij and Yudin, 1983, Nesterov, 2018]
  - $f$ is **Lipschitz smooth** with constant $L$: $\|\nabla^2 f(x)\| \leq L$ a.e.*

*a.e. = almost everywhere with respect to the Lebesgue measure

# Classical Analyses of Optimization Algorithms

## Under Lipschitz smoothness

- Unconstrained optimization $\min_{x \in \mathbb{R}^d} f(x)$ with first-order algorithms
- Classical textbook analyses [Nemirovskij and Yudin, 1983, Nesterov, 2018]
  - $f$ is **Lipschitz smooth** with constant $L$: $\|\nabla^2 f(x)\| \leq L$ a.e.*
  - A consequence: $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|^2$
  - E.g., gradient descent: $f(x_{t+1}) \leq f(x_t) - \eta(1 - \eta L/2)\|\nabla f(x_t)\|^2 \leq f(x_t)$

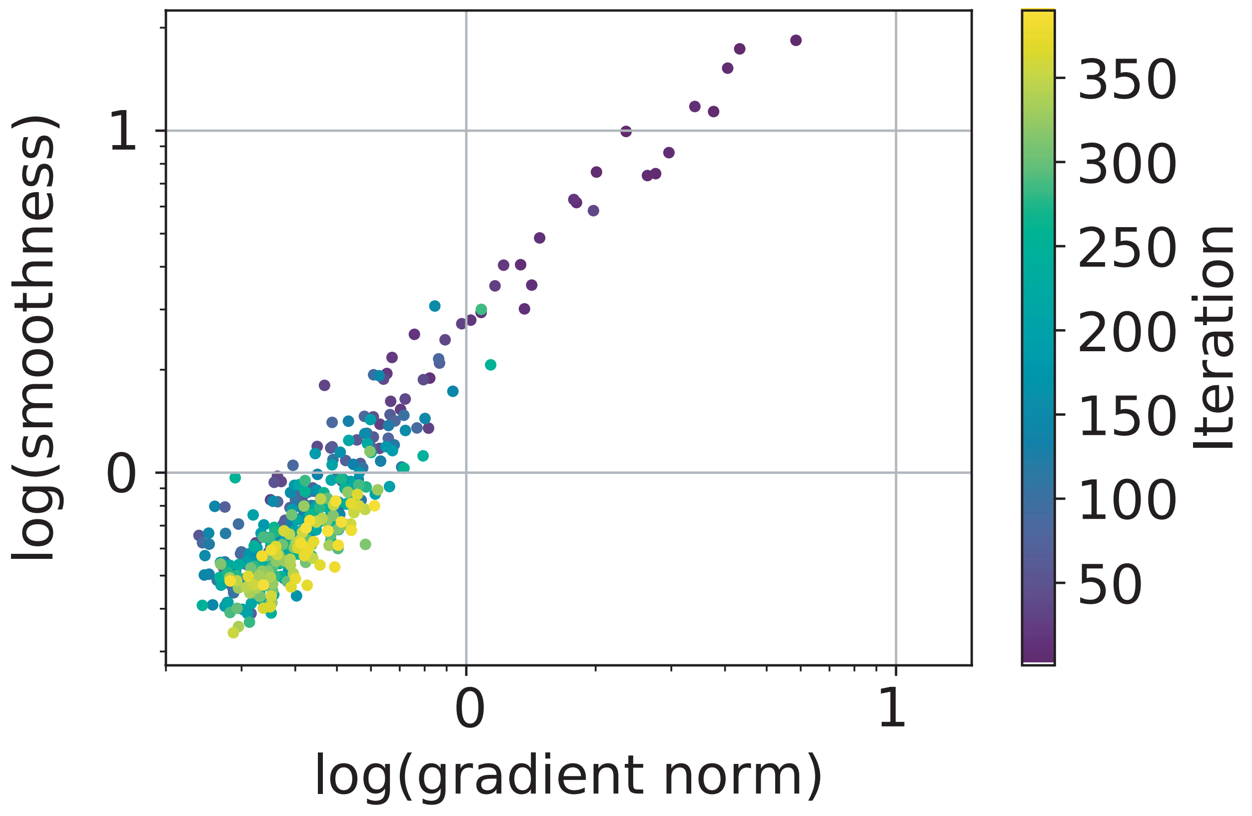
*a.e. = almost everywhere with respect to the Lebesgue measure

# Does Lipschitz Smoothness Reflect Reality?

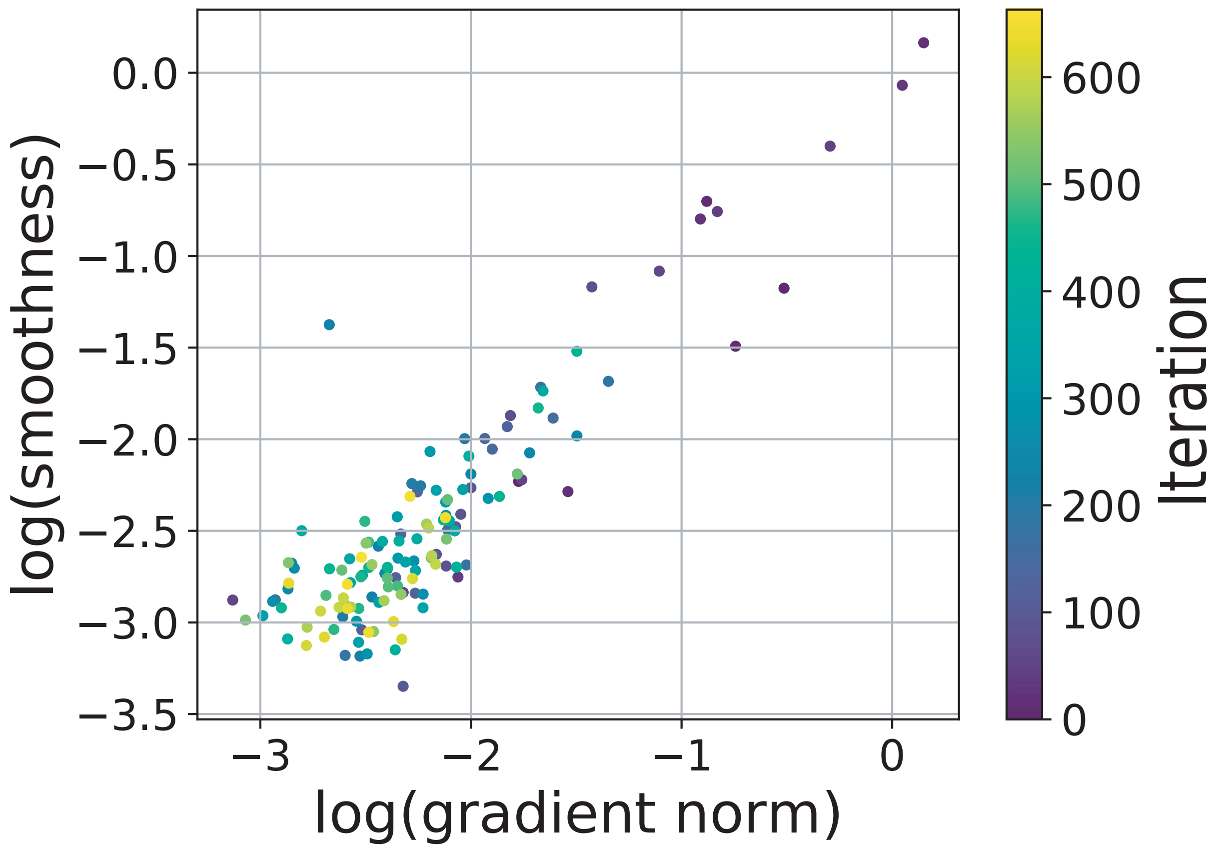
- Lipschitz smoothness is too strict!
  - Violated by polynomial ($\deg \geq 3$), rational, exponential, and logarithmic functions.

# Does Lipschitz Smoothness Reflect Reality?

- Lipschitz smoothness is too strict!
  - Violated by polynomial ($\deg \geq 3$), rational, exponential, and logarithmic functions.
- Observation in deep learning
  - Zhang et al. [2020] observe that local smoothness ($\|\nabla^2 f(x)\|$) varies a lot in terms of the gradient norm ($\|\nabla f(x)\|$) in deep architectures.




ResNet (Computer Vision)




AWS-LSTM (Language Model)

# Overview of Li et al. [2023]

- They generalize the standard Lipschitz smoothness to the $\ell$-smoothness condition: it assumes that **the Hessian norm is bounded by a non-decreasing function of the gradient norm**.

$$\|\nabla^2 f(x)\| \leq \ell(\|\nabla f(x)\|) \quad (\ell\text{: non-decreasing, continuous function}).$$

# Overview of Li et al. [2023]

- They generalize the standard Lipschitz smoothness to the $\ell$-smoothness condition: it assumes that **the Hessian norm is bounded by a non-decreasing function of the gradient norm**.

$$\|\nabla^2 f(x)\| \leq \ell(\|\nabla f(x)\|) \quad (\ell\text{: non-decreasing, continuous function}).$$

- They prove the convergence of constant-step-size first-order algorithms in the convex and non-convex settings, recovering the classical rates of:
  - Gradient descent (GD);
  - Stochastic gradient descent (SGD);
  - Nesterov's accelerated gradient method (NAG).

# Generalized Smoothness (1,2)

- **Definition 1** ($\ell$-smoothness). A real-valued differentiable function $f$ is $\ell$-smooth for a non-decreasing continuous function $\ell : [0, +\infty) \to (0, +\infty)$ if

$$\|\nabla^2 f(x)\| \leq \ell(\|\nabla f(x)\|) \text{ a.e.}$$

# Generalized Smoothness (1,2)

- **Definition 1** ($\ell$-smoothness). A real-valued differentiable function $f$ is $\ell$-smooth for a non-decreasing continuous function $\ell : [0, +\infty) \to (0, +\infty)$ if

$$\|\nabla^2 f(x)\| \le \ell(\|\nabla f(x)\|) \text{ a.e.}$$

- **Definition 2** ($(r, \ell)$-smoothness). A real-valued differentiable function $f$ is $(r, \ell)$-smooth for continuous functions $r, \ell : [0, +\infty) \to (0, +\infty)$ where $\ell$ is non-decreasing and $r$ is non-increasing if, for any $x \in \mathbb{R}^d$ and $x_1, x_2 \in \mathfrak{B}(x, r(\|\nabla f(x)\|))$,

$$\|\nabla f(x_1) - \nabla f(x_2)\| \le \ell(\|\nabla f(x)\|) \cdot \|x_1 - x_2\|.$$

*$\mathfrak{B}(x, R)$ = a closed Euclidean ball with radius $R$ centered at $x$

# Generalized Smoothness (1,2)

- **Definition 1** ($\ell$-smoothness). A real-valued differentiable function $f$ is $\ell$-smooth for a non-decreasing continuous function $\ell : [0, +\infty) \to (0, +\infty)$ if

$$\|\nabla^2 f(x)\| \leq \ell(\|\nabla f(x)\|) \text{ a.e.}$$

- **Definition 2** ($(r, \ell)$-smoothness). A real-valued differentiable function $f$ is $(r, \ell)$-smooth for continuous functions $r, \ell : [0, +\infty) \to (0, +\infty)$ where $\ell$ is non-decreasing and $r$ is non-increasing if, for any $x \in \mathbb{R}^d$ and $x_1, x_2 \in \mathfrak{B}(x, r(\|\nabla f(x)\|))$,

$$\|\nabla f(x_1) - \nabla f(x_2)\| \leq \ell(\|\nabla f(x)\|) \cdot \|x_1 - x_2\|.$$

- **Proposition 3.2.**

$$(r, \ell)\text{-smooth} \Rightarrow \ell\text{-smooth} \Rightarrow \left(\frac{a}{\ell(\cdot + a)}, \ell(\cdot + a)\right)\text{-smooth} \quad (\forall a > 0)$$

*$\mathfrak{B}(x, R)$ = a closed Euclidean ball with radius $R$ centered at $x$

# Generalized Smoothness (3)

## Important subset of $\ell$-smoothness

- **Definition 3** ($(\rho, L_0, L_\rho)$-smoothness). A real-valued differentiable function $f$ is $(\rho, L_0, L_\rho)$-smooth for constants $\rho, L_0, L_\rho \geq 0$ if it is $\ell$-smooth with $\ell(u) = L_0 + L_\rho u^\rho$.
  - $\rho = 0$ or $L_\rho = 0$: standard Lipschitz smoothness.
  - $\rho = 1$: $(L_0, L_1)$-smoothness [Zhang et al., 2020].

| $\rho$ | 0 | 1 | 1 | 1+ | 1.5 | 2 | $\frac{p-2}{p-1}$ |
|---|---|---|---|---|---|---|---|
| Functions | Quadratic | Polynomial | $a^x$ | $a^{(b^x)}$ | Rational | Logarithmic | $x^p$ |

Table. Examples of univariate $(\rho, L_0, L_\rho)$-smooth functions. The parameters $a, b, p$ are real numbers such that $a, b > 1$ and $p \in (-\infty, 1) \cup [2, \infty)$. 1+ means any real number slightly larger than 1.

# Properties of Generalized Smoothness (1)

- **Lemma 3.3**. If $f$ is $(r, \ell)$-smooth, for any $x \in \mathbb{R}^d$ satisfying $\|\nabla f(x)\| \leq G$ and any $x_1, x_2 \in \mathfrak{B}(x, r(G))$, $f$ satisfies $\|\nabla f(x_1) - \nabla f(x_2)\| \leq L\|x_1 - x_2\|$ and $f(x_1) \leq f(x_2) + \langle \nabla f(x_2), x_1 - x_2 \rangle + \frac{L}{2}\|x_1 - x_2\|^2$.

# Properties of Generalized Smoothness (1)

- **Lemma 3.3**. If $f$ is $(r, \ell)$-smooth, for any $x \in \mathbb{R}^d$ satisfying $\|\nabla f(x)\| \le G$ and any $x_1, x_2 \in \mathfrak{B}(x, r(G))$, $f$ satisfies $\|\nabla f(x_1) - \nabla f(x_2)\| \le L\|x_1 - x_2\|$ and $f(x_1) \le f(x_2) + \langle \nabla f(x_2), x_1 - x_2 \rangle + \frac{L}{2}\|x_1 - x_2\|^2$.

- *Proof Sketch.* Since $\ell$ is non-decreasing and $r$ is non-increasing, we have $\ell(\|\nabla f(x)\|) \le \ell(G)$ and $r(G) \le r(\|\nabla f(x)\|)$. Thus, the first inequality holds by definition. The second inequality follows from the first one (proof: use integrals.)

- ***Remark.*** If we properly bound the gradient norm along the optimization trajectory, then we can recover the classical analysis established upon Lipschitz smoothness!

# Properties of Generalized Smoothness (2)

- If $\ell$ is sub-quadratic ($\lim_{u\to\infty} \ell(u)/u^2 = 0$), bounded function values imply bounded gradient norms.
- Let $f^* = \inf_{x\in\mathbb{R}^d} f(x)$.
- **Corollary 3.6.** Suppose $f$ is $\ell$-smooth where $\ell$ is sub-quadratic. If $f(x) - f^* \leq F$, then we have $\|\nabla f(x)\| \leq G := \sup\{u \geq 0 \,|\, u^2 \leq 2\ell(2u)\cdot F\} < \infty$.

# Properties of Generalized Smoothness (2)

- If $\ell$ is sub-quadratic ($\lim_{u\to\infty} \ell(u)/u^2 = 0$), bounded function values imply bounded gradient norms.
- Let $f^* = \inf_{x\in\mathbb{R}^d} f(x)$.
- **Corollary 3.6.** Suppose $f$ is $\ell$-smooth where $\ell$ is sub-quadratic. If $f(x) - f^* \le F$, then we have $\|\nabla f(x)\| \le G := \sup\{u \ge 0 \,|\, u^2 \le 2\ell(2u) \cdot F\} < \infty$.
- *Proof Sketch*. This is a corollary of Lemma 3.5: If $f$ is $\ell$-smooth, then we can show that $\|\nabla f(x)\|^2 \le 2 \cdot \ell(2\|\nabla f(x)\|) \cdot (f(x) - f^*)$.
- ***Remark.*** In order to bound the gradients along the trajectory, it suffices to bound the function values, which is usually easier!

# Gradient Descent — Convex Setting

- **Lemma 4.1**. For any $x \in \mathbb{R}^d$ satisfying $\|\nabla f(x)\| \le G$, define $x^+ := x - \eta \nabla f(x)$. If $f$ is convex and $(r, \ell)$-smooth, and $\eta \le \min\left\{\frac{2}{\ell(G)}, \frac{r(G)}{2G}\right\}$, we have $\|\nabla f(x^+)\| \le \|\nabla f(x)\| \le G$.

# Gradient Descent — Convex Setting

- **Lemma 4.1**. For any $x \in \mathbb{R}^d$ satisfying $\|\nabla f(x)\| \leq G$, define $x^+ := x - \eta \nabla f(x)$. If $f$ is convex and $(r, \ell)$-smooth, and $\eta \leq \min\left\{\frac{2}{\ell(G)}, \frac{r(G)}{2G}\right\}$, we have $\|\nabla f(x^+)\| \leq \|\nabla f(x)\| \leq G$.

- *Proof Sketch.* Recall that $\ell(\|\nabla f(x)\|) \leq \ell(G)$, $r(G) \leq r(\|\nabla f(x)\|)$. Also, we can prove that convexity and $(r, \ell)$-smoothness imply the local co-coercivity: $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \frac{1}{\ell(\|\nabla f(x)\|)}\|y - x\|^2$ for all $x$ and $y \in \mathfrak{B}(x, r(\|\nabla f(x)\|)/2)$. Note that $\|x^+ - x\| = \|\eta \nabla f(x)\| \leq \eta G \leq r(G)/2$. Then by applying the local co-coercivity,

$$
\begin{aligned}
\|\nabla f(x^+)\|^2 - \|\nabla f(x)\|^2 &= 2\langle \nabla f(x^+) - \nabla f(x), \nabla f(x) \rangle + \|\nabla f(x^+) - \nabla f(x)\|^2 \\
&= -\frac{2}{\eta}\langle \nabla f(x^+) - \nabla f(x), x^+ - x \rangle + \|\nabla f(x^+) - \nabla f(x)\|^2 \\
&\leq -\left(\frac{2}{\eta \cdot \ell(\|\nabla f(x)\|)} - 1\right)\|\nabla f(x^+) - \nabla f(x)\|^2 \leq 0.
\end{aligned}
$$

# Gradient Descent — Convex Setting

- **Theorem 4.2–3.** Suppose $f$ is convex and $(r, \ell)$-smooth. Denote $G = \|\nabla f(x_0)\|$. Choose the step size $\eta \leq \min\left\{\frac{1}{\ell(G)}, \frac{r(G)}{2G}\right\}$. Then the gradient descent iterates ($x_{t+1} = x_t - \eta\nabla(x_t)$) satisfy $\|\nabla f(x_t)\| \leq G$ for all $t \geq 0$ and

$$f(x_T) - f^* \leq \frac{\|x_0 - x^*\|^2}{2\eta T}. \quad \text{(Thm 4.2)}$$

  If $f$ is $\mu$-strongly convex, then

$$f(x_T) - f^* \leq \frac{\mu(1-\eta\mu)^T}{2(1-(1-\eta\mu)^T)}\|x_0 - x^*\|^2. \quad \text{(Thm 4.3)}$$

- *Proof Sketch.* Apply Lemma 4.1 and the usual potential function analysis [Bansal and Gupta, 2019].

# Gradient Descent — Convex Setting

- **Theorem 4.2–3.** Suppose $f$ is convex and $(r, \ell)$-smooth. Denote $G = \|\nabla f(x_0)\|$. Choose the step size $\eta \leq \min\left\{\frac{1}{\ell(G)}, \frac{r(G)}{2G}\right\}$. Then the gradient descent iterates ($x_{t+1} = x_t - \eta \nabla(x_t)$) satisfy $\|\nabla f(x_t)\| \leq G$ for all $t \geq 0$ and

$$f(x_T) - f^* \leq \frac{\|x_0 - x^*\|^2}{2\eta T}. \quad \text{(Thm 4.2)}$$

If $f$ is $\mu$-strongly convex, then

$$f(x_T) - f^* \leq \frac{\mu(1-\eta\mu)^T}{2(1-(1-\eta\mu)^T)}\|x_0 - x^*\|^2. \quad \text{(Thm 4.3)}$$

- *Proof Sketch.* Apply Lemma 4.1 and the usual potential function analysis [Bansal and Gupta, 2019].
- ***Remark.*** Theorems above recover the classical convergence rates:
  - Theorem 4.2 gives $O(1/\epsilon)$ gradient complexity for convex $(r, \ell)$-smooth functions to achieve $f(x_T) - f^* \leq \epsilon$.
  - Theorem 4.3 gives $O((\eta\mu)^{-1}\log(1/\epsilon))$ gradient complexity for $\mu$-strongly convex $(r, \ell)$-smooth functions.

# Gradient Descent — Non-convex Setting

## With sub-quadratic $\ell$

- **Lemma 5.1.** Suppose $f$ is $\ell$-smooth where $\ell$ is sub-quadratic. For any given $F \geq 0$, let $G := \sup\{u \geq 0 \,|\, u^2 \leq 2\ell(2u) \cdot F\}$ and $L = \ell(2G)$. For any $x \in \mathbb{R}^d$ satisfying $f(x) - f^* \leq F$, define $x^+ := x - \eta \nabla f(x)$. If $\eta \leq \frac{1}{L}$, we have $f(x^+) \leq f(x)$.

# Gradient Descent — Non-convex Setting

## With sub-quadratic $\ell$

- **Lemma 5.1.** Suppose $f$ is $\ell$-smooth where $\ell$ is sub-quadratic. For any given $F \geq 0$, let $G := \sup\{u \geq 0 \,|\, u^2 \leq 2\ell(2u) \cdot F\}$ and $L = \ell(2G)$. For any $x \in \mathbb{R}^d$ satisfying $f(x) - f^* \leq F$, define $x^+ := x - \eta \nabla f(x)$. If $\eta \leq \frac{1}{L}$, we have $f(x^+) \leq f(x)$.

- *Proof Sketch.* By Corollary 3.6, we know $\|\nabla f(x)\| \leq G$. By Proposition 3.2, we know $\ell$-smoothness implies $(\frac{G}{\ell(\cdot + G)}, \ell(\cdot + G))$-smoothness. Thus, by Lemma 3.3, $f$ is locally Lipschitz $L$-smooth on a closed Euclidean ball with a radius $G/L$. Note that $\|x^+ - x\| = \|\eta \nabla f(x)\| \leq \eta G \leq G/L$. Then applying the usual descent lemma,

$$
\begin{aligned}
f(x^+) - f(x) &\leq \langle \nabla f(x), x^+ - x\rangle + \frac{L}{2}\|x^+ - x\|^2 \\
&= -\eta\left(1 - \frac{\eta L}{2}\right)\|\nabla f(x)\|^2 \leq 0.
\end{aligned}
$$

# Gradient Descent — Non-convex Setting

## With sub-quadratic $\ell$

- **Theorem 5.2.** Suppose $f$ is $\ell$-smooth where $\ell$ is sub-quadratic. Let $G := \sup\{u \geq 0 \,|\, u^2 \leq 2\ell(2u) \cdot (f(x_0) - f^*)\}$ and $L = \ell(2G)$. Choose the step size $\eta \leq \frac{1}{L}$. Then the gradient descent iterates ($x_{t+1} = x_t - \eta \nabla(x_t)$) satisfy $\|\nabla f(x_t)\| \leq G$ for all $t \geq 0$ and

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \leq \frac{2(f(x_0) - f^*)}{\eta T}.$$

- *Proof Sketch.* Applying Lemma 5.1 and Corollary 3.6, we obtain $f(x_t) \leq f(x_0)$ and thus $\|\nabla f(x_t)\| \leq G$. Following the proof of Lemma 5.1, we obtain $f(x_{t+1}) - f(x_t) \leq -\eta\left(1 - \frac{\eta L}{2}\right)\|\nabla f(x_t)\|^2$. Taking a summation over $t = 0, \cdots, T-1$ and rearranging terms, we complete the proof.

- ***Remark.*** Theorem above recovers the classical convergence rates:
  - Theorem 5.2 gives $O(1/\epsilon^2)$ gradient complexity for $(r, \ell)$-smooth functions to achieve an $\epsilon$-stationary point, which is optimal as it matches the lower bound in Carmon et al. [2020].

# Gradient Descent — Non-convex Setting

## What about non-sub-quadratic $\ell$? ($\rho \geq 2$)

- The gradient complexity is at least exponentially large in the problem parameter.
- **Theorem 5.4.** Given $L_0, L_2, F_0, G_0 > 0$ such that $L_2 F_0 \geq 10$, for any $\eta \geq 0$, there exists a $(\mathbf{2}, L_0, L_2)$-smooth univariate function $f$, which is bounded below, and an initial point $x_0$ satisfying $|f'(x_0)| \leq G_0$ and $f(x_0) - f^* \leq F_0$, such that GD with step size $\eta$ either cannot reach a 1-stationary point or takes at least $\exp(L_2 F_0/8)/6$ steps to reach a 1-stationary point.
- *Proof Sketch*. If $\eta > \frac{L_0}{2}$, taking $f(x) = \frac{L_0}{2}x^2$, GD will diverge. Otherwise, we carefully take a piecewise logarithmic/quadratic function (which is $(2, L_0, L_2)$-smooth, independent to the step-size) so that either GD gets stuck or takes exponentially many steps to reach a 1-stationary point.

# Nesterov's Accelerated Gradient Method

**Convex & Sub-quadratic $\ell$ ➡ Optimal $O(1/\sqrt{\epsilon})$ gradient complexity**

**Algorithm 1:** Nesterov's Accelerated Gradient Method (NAG)

**input** A convex and $\ell$-smooth function $f$, stepsize $\eta$, initial point $x_0$

**Initialize** $z_0 = x_0$, $B_0 = 0$, and $A_0 = 1/\eta$.
**for** $t = 0, \ldots$ **do**
  $B_{t+1} = B_t + \frac{1}{2}\left(1 + \sqrt{4B_t + 1}\right)$
  $A_{t+1} = B_{t+1} + 1/\eta$
  $y_t = x_t + (1 - A_t/A_{t+1})(z_t - x_t)$
  $x_{t+1} = y_t - \eta \nabla f(y_t)$
  $z_{t+1} = z_t - \eta(A_{t+1} - A_t)\nabla f(y_t)$
**end for**

- **Theorem 4.4.** Suppose $f$ is convex and $\ell$-smooth where $\ell$ is sub-quadratic. Let $G$ be a constant satisfying $G \geq \max\left\{8\sqrt{\ell(2G)((f(x_0) - f^*) + \|x_0 - x^*\|^2)}, \|\nabla f(x_0)\|\right\}$. Denote $L = \ell(2G)$ and choose $\eta \leq \min\{\frac{1}{16L^2}, \frac{1}{2L}\}$. The iterates generated by NAG satisfy

$$f(x_T) - f^* \leq \frac{4(f(x_0) - f^*) + r\|x_0 - x^*\|^2}{\eta T^2 + 4}.$$

# Stochastic Gradient Descent

**Non-convex & Sub-quadratic $\ell$ ➡ Optimal $O(1/\epsilon^4)$ gradient complexity (w.h.p.)**

- Assumption: Stochastic gradient $g_t$ is unbiased and has bounded variance ($\sigma^2$).
- **Theorem 5.3.** Suppose $\ell$-smooth where $\ell$ is sub-quadratic. For any $\delta \in (0,1)$, denote $F = 8(f(x_0) - f^* + \sigma)/\delta$ and $G = \sup\{u \geq 0 \,|\, u^2 \leq 2\ell(2u) \cdot F\}$. Denote $L = \ell(2G)$ and choose $\eta \leq \min\{\frac{1}{2L}, \frac{1}{4G\sqrt{T}}\}$ and $T \geq \frac{F}{\eta\epsilon^2}$ for any $\epsilon > 0$. Then with probability at least $1 - \delta$, the iterates generated by SGD satisfy $\|\nabla f(x_t)\| \leq G$ for all $t < T$ and

$$\frac{1}{T}\sum_{t=0}^{T-1} \|\nabla f(x_t)\|^2 \leq \epsilon^2$$

# Summary

Table 1: Summary of the results. $\epsilon$ denotes the sub-optimality gap of the function value in convex settings, and the gradient norm in non-convex settings. "$*$" denotes optimal rates.

| Method | Convexity | $\ell$-smoothness | Gradient complexity |
|---|---|---|---|
| GD | Strongly convex | No requirement | $\mathcal{O}(\log(1/\epsilon))$ (Theorem 4.3) |
| | Convex | | $\mathcal{O}(1/\epsilon)$ (Theorem 4.2 ) |
| | Non-convex | Sub-quadratic $\ell$ | $\mathcal{O}(1/\epsilon^2)^*$ (Theorem 5.2) |
| | | Quadratic $\ell$ | $\Omega$(exp. in cond #) (Theorem 5.4 ) |
| NAG | Convex | Sub-quadratic $\ell$ | $\mathcal{O}(1/\sqrt{\epsilon})^*$ (Theorem 4.4 ) |
| SGD | Non-convex | Sub-quadratic $\ell$ | $\mathcal{O}(1/\epsilon^4)^*$ (Theorem 5.3) |

# Discussions

- All the results are in the form of:
  - Generalized smoothness (assumption)
  - + Bounded gradients along the trajectory (not an assumption)

  ➡ Standard Lipschitz smoothness! Similar analyses to the classical ones!
- Generalized smoothness might give a better geometry than the standard Lipschitz smoothness.
  - If generalized smoothness can give a tighter upper bound on the Hessian norm than the Lipschitz smoothness along the trajectory, shouldn't we have gotten a better convergence rate, rather than obtaining the identical rate as the classical one?
- In the non-convex setting (sub-quadratic $\ell$), (S)GD is still rate-optimal. In practice, vanilla (S)GD performs worse than methods with momentum or adaptive methods. This means either…
  - Although the rate is optimal, the hidden constants are too large, which hurts the performance in reality.
  - Or, generalized smoothness might not be enough.

# References


- Nikhil Bansal, and Anupam Gupta. *Potential-function proofs for first-order methods*. Theory of Computing. 15: 1-32. 2019.
- Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. *Lower bounds for finding stationary points I.* Mathematical Programming, 184(1–2): 71–120, Nov 2020. ISSN 0025-5610.
- Arkadij Semenovič Nemirovskij and David Borisovich Yudin. *Problem complexity and method efficiency in optimization.* Wiley-Interscience, 1983.
- **Haochuan Li, Jian Qian, Yi Tian, Alexander Rakhlin, and Ali Jadbabaie. Convex and Non-convex Optimization under Generalized Smoothness. NeurIPS 2023.**
- Yurii Nesterov. *Lectures on Convex Optimization.* Springer Publishing Company, Incorporated, 2nd edition, 2018. ISBN 3319915770.
- Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. ICLR 2020.