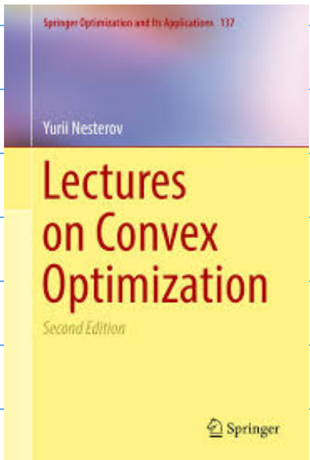
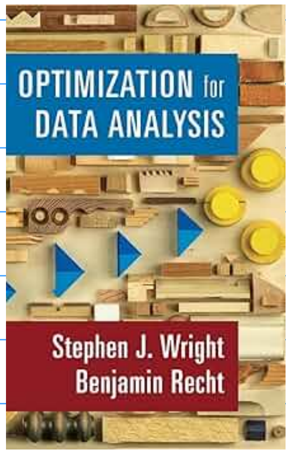
Complexity in continuous optimization (2/6)

February 6, 2024

Today: Accelerated gradient from convex to nonconvex

References:





"Convex Until Proven Guilty": Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions

ICML 2017

Yair Carmon John C. Duchi Oliver Hinder Aaron Sidford [1]

Mathematical Programming (2021) 185:315–355
https://doi.org/10.1007/s10107-019-01431-x

FULL LENGTH PAPER

Series A

Check for updates

Lower bounds for finding stationary points II: first-order methods

Yair Carmon[1] · John C. Duchi[2] · Oliver Hinder[3] · Aaron Sidford[3]

Where we stand:

$$\underset{x\in\mathbb{R}^n}{\text{minimize}}\ f(x) \quad , \quad f\ C_L^{1,1} \quad \left(\nabla f \text{ exists at every } x\in\mathbb{R}^n \text{ and } \|\nabla f(x)-\nabla f(y)\| \le L\|x-y\|\right)$$

Goal: Bound the number of iterations/gradient evaluations/function evaluations to find a point such that $\|\nabla f(x)\| \le \varepsilon$

One algorithm: Gradient descent

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k) \quad , \quad \alpha_k > 0 \quad (\text{e.g. } \alpha_k = \tfrac{1}{L})$$

Complexity: $\|\nabla f(x_k)\| \le \varepsilon$ after at most $O(\varepsilon^{-2})$ iterations

(Valid $\forall \varepsilon > 0$, implies $\liminf_{k\to\infty} \|\nabla f(x_k)\| = 0$)

Upper bound is $O(\varepsilon^{-2})$

Lower bound matches the upper bound: There exists a function f $C_L^{1,1}$ such that GD takes exactly ε^{-2} iterations to reach a point such that $\|\nabla f(x)\| \le \varepsilon$.

$\Rightarrow$ Sharp analysis (Lower and upper bound match)

$\Rightarrow$ How can we improve these results? Look at subclasses of $C_L^{1,1}$ functions

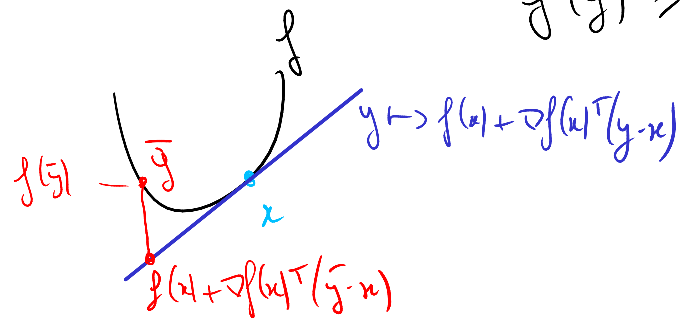
Today:
- convex
- $C_L^{1,1} + C^2$ (nonconvex)

1) Convex optimization

In this part, we suppose that f is $C^{1,1}_L$ and convex or strongly convex.

f $C^{1,1}_L$ is convex $\iff \forall (x,y) \in (\mathbb{R}^n)^2$,

$$f(y) \geq \underbrace{f(x) + \nabla f(x)^T (y-x)}_{\text{linear function of } y}$$

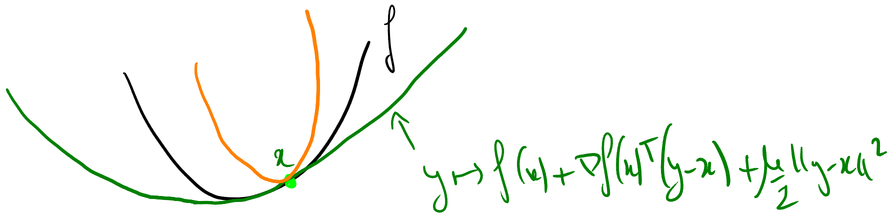


Property: If f is C^1 convex, then $\nabla f(\bar{x}) = 0 \iff \bar{x} \in \underset{x}{\operatorname{argmin}} f(x)$

($\bar{x}$ global minimum of f)

- f $C^{1,1}_L$ is μ-strongly convex ($\mu > 0$)

$\iff \forall (x,y) \in (\mathbb{R}^n)^2$, $$f(y) \geq \underbrace{f(x) + \nabla f(x)^T (y-x) + \frac{\mu}{2} \|y-x\|^2}_{\text{quadratic function of } y}$$



(NB: f $C^{1,1}_L \implies \forall (x,y) \in (\mathbb{R}^n)^2$, $f(y) \leq \underbrace{f(x) + \nabla f(x)^T (y-x) + \frac{L}{2} \|y-x\|^2}$)

Property: If f $C^{1,1}_L$ and μ-strongly convex, then it has a unique minimum which is the unique solution of $\nabla f(x) = 0_{\mathbb{R}^n}$

f is μ-strongly convex $\iff$ $x \mapsto \underbrace{f(x) - \frac{\mu}{2}\|x\|^2}_{\bar{f}}$ is convex.

$$\underset{x,y}{\text{minimize}} \quad \bar{f}(y) - \bar{f}(x) - \nabla \bar{f}(x)^T (y-x)$$

$$f\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right) = x_1^2$$

$\hookrightarrow$ Gradient descent on convex problems

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

$\rightarrow$ f $C_L^{1,1}$ and convex, and has at least 1 minimum

Then, GD reaches a point x_K such that

$f(x_K) - \min_{x \in \mathbb{R}^d} f(x) \leq \varepsilon$ in at most $O(\varepsilon^{-1})$ iterations

$\iff$ $\|\nabla f(x_K)\| \leq \varepsilon$ in at most $O(\varepsilon^{-1})$ iterations)

Typical criterion for complexity in the convex setting (replaces $\|\nabla f(x)\| \leq \varepsilon$ used in the nonconvex setting)

$\Rightarrow$ Indicates how far the current function value is from the minimum value

Proof: • Since f is $C_L^{1,1}$, $\forall k \in \mathbb{N}$,

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

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

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

Let $x^* \in \underset{x}{\operatorname{argmin}} f(x)$ and $f^* = f(x^*) = \min_x f(x)$

By convexity, $f(x^*) \geq f(x_k) + \nabla f(x_k)^T (x^* - x_k)$

$$\Leftrightarrow \quad f(x_k) \leq f(x^*) + \nabla f(x_k)^T (x_k - x^*)$$

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

$\forall k \in \mathbb{N}$
$$f(x_{k+1}) - f(x^*) \leq \nabla f(x_k)^T (x_k - x^*) - \frac{1}{2L} \|\nabla f(x_k)\|^2$$
$$= \frac{L}{2} \left(\|x_k - x^*\|^2 - \|\underbrace{x_{k+1}}_{x_k - \frac{1}{L}\nabla f(x_k)} - x^*\|^2 \right)$$

Suppose that $f(x_k) - f(x^*) > \varepsilon \quad \forall k = 0, \ldots, K$

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

$$K\varepsilon < \sum_{k=0}^{K-1} \underbrace{\left(f(x_{k+1}) - f(x^*) \right)}_{> \varepsilon} \leq \boxed{\frac{L}{2} \|x_0 - x^*\|^2}$$

$$\Rightarrow \quad K < \frac{L}{2} \|x_0 - x^*\|^2 \varepsilon^{-1} = O(\varepsilon^{-1})$$

For μ-strongly convex case, can show

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

(Behind this: $\|\nabla f(x)\|^2 \geq 2\mu \, (f(x) - f(x^*))$, mathematical consequence of strong convexity inequality)

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

$$\leq f(x_k) - \frac{\mu}{L} (f(x_k) - f(x^*))$$

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

$\rightarrow$ The proof in the strongly convex setting is (in some way) easier than the proof in the convex setting

Complexity of GD : $O\left(\frac{L}{\mu} \ln(\varepsilon^{-1})\right)$

Same GD algorithm	$f \in C_L^{1,1}$	
f nonconvex	$\|\nabla f(x)\| \leq \varepsilon$	$O(\varepsilon^{-2})$
f convex	$f(x) - \min_x f(x) \leq \varepsilon$	$O(\varepsilon^{-1})$
f μ-strongly convex	"	$O\left(\frac{L}{\mu} \ln(\varepsilon^{-1})\right)$

constant in terms of ε but important for the complexity

L/μ : "condition number"

Pb: In the convex/strongly convex settings, the upper bounds for GD do not match the lower bounds

Nemirovski & Yudin 1983

Existence result: there exists an algorithm that uses only one gradient per iteration

→ If f $C^{1,1}_L$ convex, there exists an algorithm with complexity $O(\varepsilon^{-1/2})$

→ If f $C^{1,1}_L$ μ-strongly convex, $\exists$ algorithm with complexity $O\left(\sqrt{\frac{L}{\mu}}\ln(\varepsilon^{-1})\right)$

↑ $\mu \le L$ $\quad \sqrt{\frac{L}{\mu}} \le \frac{L}{\mu}$

→ Algorithm was unknown until Yurii Nesterov discovered it in 1983

Accelerated gradient / Nesterov's method ($x_0 \in \mathbb{R}^n$)

$$x_{k+1} = x_k - \alpha_k \nabla f\big(x_k + \underbrace{\beta_k (x_k - x_{k-1})}_{\text{Momentum term}}\big) + \beta_k (x_k - x_{k-1})$$

$\alpha_k > 0$, $\beta_k > 0$ $\quad (\beta_0 = 0,\ x_{-1} = x_0)$

Equivalent formulation:

$$\begin{cases} y_k = x_k + \beta_k (x_k - x_{k-1}) \\ x_{k+1} = y_k - \nabla f(y_k) \end{cases}$$

$\beta_0 = 0$
$x_{-1} = x_0$
$y_0 = x_0$

Note: Momentum methods in ML (SGD with momentum, Adam)

$$x_{k+1} = x_k - \alpha_k \nabla f_{i_k}(x_k) + \beta_k (x_k - x_{k-1})$$

$\nabla f_{i_k}(x_k)$: stochastic gradient $\simeq$ sample

Analyzing Nesterov's method

$$\begin{cases} y_k = x_k + \beta_k (x_k - x_{k-1}) \\ x_{k+1} = y_k - \alpha_k \nabla f(y_k) \end{cases} \qquad y_0 = x_0 \ , \ \beta_0 = 0 \ , \ x_{-1} = x_0$$

a) f is μ-strongly convex

$$\rightarrow \alpha_k = \frac{1}{L} \ , \ \underset{k \geq 1}{\beta_k} = \frac{\sqrt{\kappa} - 1}{\sqrt{\kappa} + 1} \quad \text{where} \quad \kappa = \frac{L}{\mu}$$

<u>Key difference with GD analysis</u> $\qquad f^* = \min f(x)$

- In GD, we look at $f(x_k) - f^*$ and we show

$$f(x_k) - f^* \leq \left(1 - \frac{\mu}{L}\right)^k \left(f(x_0) - f^*\right)$$
$$= \left(1 - \frac{1}{\kappa}\right)^k \left(f(x_0) - f^*\right)$$

"Lyapunov function for GD"

- In AG (Accelerated Gradient)'s analysis, we use

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

with $\rho^2 = \left(1 - \frac{1}{\sqrt{\kappa}}\right) = 1 - \sqrt{\frac{\mu}{L}}$

$\Rightarrow$ the analysis shows $V_k \leq \rho^k V_0$

$$V_0 = f(x_0) - f^* + \frac{L}{2}\|(1-\rho^2)(x_0 - x^*)\|^2$$
$$= f(x_0) - f^* + \frac{\mu}{2}\|x_0 - x^*\|^2$$

$\Rightarrow$ Gives the $O\left(\sqrt{\frac{L}{\mu}}\ln(\varepsilon^{-1})\right)$ bound

vs $O\left(\frac{L}{\mu}\ln(\varepsilon^{-1})\right)$ for gradient descent

b) Convex case (Not strongly convex)

Nesterov's implementation of accelerated gradient

$$\begin{cases} y_k = x_k + \beta_k(x_k - x_{k-1}) \\ x_{k+1} = y_k - \frac{1}{L}\nabla f(y_k) \end{cases}$$

$\beta_0 = 0$

$\beta_k = \rho_k \rho_{k-1}^2 \quad k \geq 1$

$\rho_0 = \rho_{-1} = 0$

Most non-intuitive part of the algorithm: $\{\rho_k\}$ and $\{\beta_k\}$ is defined independently of the problem (f) and x_0

ρ_{k+1} is the positive root of $1 + \rho_{k+1}(\rho_k^2 - 1) - \rho_{k+1}^2 = 0$

$$\rho_{k+1} = \frac{1 + \sqrt{1 + 4\rho_k^2}}{2}$$

$\rightarrow$ Analysis relies on the Lyapunov function

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

$\rightarrow$ Show $W_k \leq \rho_{k-1}^2 \cdots \rho_1^2 W_1 = (1 - \rho_{k-1}^2)^2 W_1$

and $W_1 \leq \boxed{\frac{L}{2}\|x_0 - x^*\|^2}$

$$\rightarrow \quad 1-\rho_k^2 \leq \frac{2}{k+2} \quad \forall k \quad (\text{by definition of } \rho_k)$$

Overall, prove
$$f(x_k) - f^* \leq V_k \leq \frac{2L}{(k+1)^2} \|x_0 - x^*\|^2$$

$\Rightarrow$ $f(x_k) - f^* \leq \varepsilon$ after at most $O(\varepsilon^{-1/2})$ iterations.

$\rightarrow$ Nesterov's method attains the upper bound $O(\varepsilon^{-1/2})$

$\rightarrow$ Nesterov also showed a lower bound for the convex setting

$$f(x) = \frac{1}{2} x^T A x - e_1^T x \qquad \text{convex quadratic}$$
$$= \frac{1}{2}(x_1 - 1)^2 + \frac{1}{2} \sum_{i=1}^{n-1} (x_i - x_{i+1})^2$$

$$A = \begin{bmatrix} 2 & -1 & & 0 \\ -1 & \ddots & \ddots & \\ & \ddots & \ddots & -1 \\ 0 & & -1 & 2 \end{bmatrix}, \quad e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \quad \text{unique solution } x^* = \begin{bmatrix} 1 - 1/_{n+1} \\ \vdots \\ 1 - n/_{n+1} \end{bmatrix}$$

Any algorithm starting from $x_0 = 0$ and using only 1 gradient per iteration will satisfy

$$f(x_k) - f^* \geq \frac{3}{8(k+1)^2} \|x_0 - x^*\|^2 \qquad \forall k \leq \frac{n}{2} - 1$$

Nesterov's example: dimension "$\varepsilon^{-1/2}$" but quadratic convex
$\overset{\shortparallel}{m}$

Cartis, Gould Toint examples: dimension 1 or 2, but very nonlinear nonquadratic

2) Acceleration in the nonconvex setting

→ If f is just $C_L^{1,1}$ and nonconvex, cannot do better than GD!

→ If f is $C_L^{1,1}$, nonconvex and $C^{2,2}$, then you can do better than GD, and you can do that with a variant of Nesterov's method (Carmon et al '17)

Idea: "Convex until proven guilty"

- Run AG as if the function were strongly convex
- Two cases:
  - Either the method works as in the strongly convex setting
  - Or it doesn't, and take a negative curvature step instead of doing 1 iteration of AG

Negative curvature

f nonconvex C^2: $\quad \bar{x} \in \operatorname{argmin}_x f(x) \Rightarrow \begin{cases} \nabla f(\bar{x}) = 0 \\ \nabla^2 f(\bar{x}) \succeq 0 \end{cases}$

↑

$\forall v \in \mathbb{R}^n, \quad v^T \nabla^2 f(\bar{x}) v \geq 0$

When f is convex C^2, $\nabla^2 f(x) \succeq 0$ true $\forall x$!

For nonconvex f, if $\nabla^2 f(x) \succeq 0$ does not hold,

the function decreases from x along any direction v such that $v^T \nabla^2 f(x) v < 0$

(and $v^T \nabla f(x) \leq 0$)

Negative curvature direction

Negative curvature step: $x + \alpha v$, $\alpha > 0$

Algorithm $(x_0, L, \varepsilon, \varepsilon_H)$

$\varepsilon > 0$ for complexity: used in the algorithm

L Lipschitz constant for the gradient

For $k = 0, 1, \ldots$

→ Define $\hat{f}_k(z) = f(z) + \varepsilon_H \|z - x_k\|^2$ ≃ Lyapunov function

→ Run AG on $\hat{f}_k$ to find a point z_k such that $\|\nabla \hat{f}_k(z_k)\| \leq \varepsilon/10$

assuming that the function $\hat{f}_k$ is ε_H-strongly convex

true if f is convex, false otherwise

. At every iteration of AG, check whether the strong convexity inequality holds between every pair of points that it computed, and stop if the inequality is violated.

$\hat{f}(u) \geq \hat{f}_k(v) + \nabla \hat{f}_k(v)^T (u-v) + \frac{\varepsilon_H}{2}\|u-v\|^2$

- If AG stops with $\|\nabla \hat{f}_k(z_k)\| \leq \frac{\varepsilon}{10}$, define $x_{k+1} = z_k$.

The function is nonconvex "enough" around x_k for AG to fail →

- Otherwise, get a pair (u_k, v_k) that violates the strong convexity inequality and use $u_k - v_k$ as a negative curvature direction.

$$x_{k+1} = x_k + \eta(u_k - v_k)$$

Analysis: At every iteration k,

- $f(x_k) - f(x_{k+1}) \geq O\left(\frac{\|\nabla f(x_k)\|^2}{\varepsilon_H}\right) \geq O\left(\frac{\varepsilon^2}{\varepsilon_H}\right)$ if AG succeeds (if $\|\nabla f(x_k)\| \geq \varepsilon$)

- $f(x_k) - f(x_{k+1}) \geq O(\varepsilon_H^3)$ if AG fails

⇒ The method computes x_k such that $\|\nabla f(x_k)\| \leq \varepsilon$ in at most $O\left(\max\left(\varepsilon^{-2}\varepsilon_H, \varepsilon_H^{-3}\right)\right)$ iterations

⇒ With $\varepsilon_H = \varepsilon^{1/2}$, get $O(\varepsilon^{-3/2})$ outer iterations

Better than $O(\varepsilon^{-2})$ for GD!

→ Every call to AG terminates (with success/failure) after at most $O\left(\sqrt{\frac{L+2\varepsilon_H}{\varepsilon_H}}\ln\left(\frac{1}{\varepsilon}\right)\right) = O\left(\sqrt{\frac{1}{\varepsilon_H}}\ln\left(\frac{1}{\varepsilon}\right)\right)$

⇒ Total number of iterations:

$$O\left(\frac{1}{\sqrt{\varepsilon_H}}\ln\left(\frac{1}{\varepsilon}\right)\right) \times O\left(\max\left(\varepsilon^{-2}\varepsilon_H, \varepsilon_H^{-3}\right)\right)$$

$\varepsilon_H = \varepsilon^{1/2}$ ↪ $O\left(\varepsilon^{-7/4}\ln(1/\varepsilon)\right) = \tilde{O}\left(\varepsilon^{-7/4}\right)$

⇒ Still better (for small ε) than $O(\varepsilon^{-2})$

$\varepsilon^{-7/4}$: Best known upper bound for algorithms that use only gradient information to optimize a $C^{1,1} \cap C^{2,2}$ function

Lower bound (for this class of methods/functions): $O\left(\varepsilon^{-12/7}\right)$

⇒ Attained for an example with dimension $O(\varepsilon^{-12/7})$

Takeaways

- AG better than GD for convex / strongly convex) + variants
- Analysis: Lyapunov functions + convexity inequalities
- Extension to nonconvex (+ negative curvature)