Optimization for Data Science
ETH Zürich, FS 2023 261-5110-00L

Lecture 3: Gradient Descent

Bernd Gärtner
Niao He

https://www.ti.inf.ethz.ch/ew/courses/ODS23/index.html

March 6, 2023

Idea

$f : \mathbb{R}^d \to \mathbb{R}$ convex, differentiable, has a global minimum $\mathbf{x}^\star$.

Goal: For given $\varepsilon > 0$, find $\mathbf{x} \in \mathbb{R}^d$ such that
this is a very small value for the practical computation.

$$f(\mathbf{x}) - f(\mathbf{x}^\star) \leq \varepsilon.$$

Note that there can be several global minima $\mathbf{x}_1^\star \neq \mathbf{x}_2^\star$ with $f(\mathbf{x}_1^\star) = f(\mathbf{x}_2^\star)$.

the function is not convex in this case.

Iterative Algorithm: choose $\mathbf{x}_0 \in \mathbb{R}^d$.

$$\mathbf{x}_{t+1} := \mathbf{x}_t + \mathbf{v}_t$$

for **times** $t = 0, 1, \ldots,$ and **steps** $\mathbf{v}_t \in \mathbb{R}^d$.
this is the step size for x

Rate of convergence: how quickly does the sequence $(f(\mathbf{x}_t) - f(\mathbf{x}^\star))_{t \in \mathbb{N}}$ converge to 0? We distinguish sublinear, linear, superlinear convergence (Section 3.1.1).

The Algorithm

$$\mathbf{x}_{t+1} := \mathbf{x}_t + \mathbf{v}_t$$

How to choose $\mathbf{v}_t$ to get $f(\mathbf{x}_{t+1}) < f(\mathbf{x}_t)$?

Differentiability of f at $\mathbf{x}_t$: for $\|\mathbf{v}_t\|$ tending to 0,

$$f(\mathbf{x}_t + \mathbf{v}_t) = f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{v}_t + \underbrace{r(\mathbf{v}_t)}_{o(\|\mathbf{v}_t\|)} \approx f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top \mathbf{v}_t.$$

Among all steps of the same length, make the one minimizing $\nabla f(\mathbf{x}_t)^\top \mathbf{v}_t$!

$\Rightarrow$ Let $\mathbf{v}_t$ point into the direction of the negative gradient $-\nabla f(\mathbf{x}_t)$!

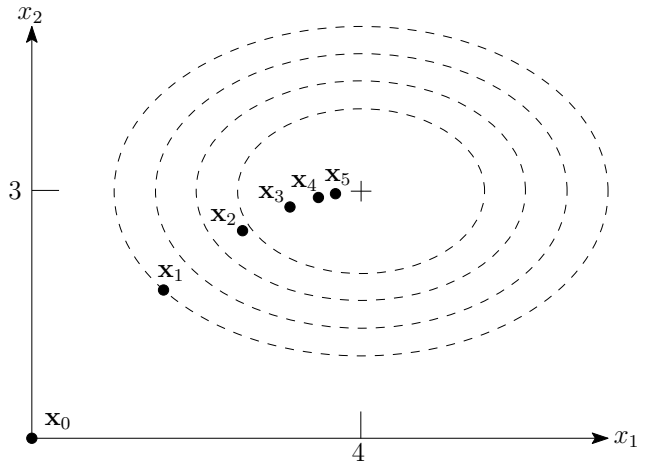
Gradient descent: choose $\mathbf{x}_0 \in \mathbb{R}^d$.

$$\mathbf{x}_{t+1} := \mathbf{x}_t - \gamma \nabla f(\mathbf{x}_t),$$

this is the form of gradient descent.

for **times** $t = 0, 1, \ldots,$ and **stepsize** $\gamma > 0$.

Example



$$f(x_1, x_2) = 2(x_1 - 4)^2 + 3(x_2 - 3)^2, \mathbf{x}_0 = (0, 0), \gamma = 0.1$$

Vanilla analysis

How to bound $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$? First-order characterization of convexity!

$$f(\mathbf{x}^\star) \geq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top (\mathbf{x}^\star - \mathbf{x}_t) \quad \Leftrightarrow \quad f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \nabla f(\mathbf{x}_t)^\top (\mathbf{x}_t - \mathbf{x}^\star).$$

- Abbreviate $\mathbf{g}_t := \nabla f(\mathbf{x}_t)$ (gradient descent: $\mathbf{g}_t = (\mathbf{x}_t - \mathbf{x}_{t+1})/\gamma$).

$$\mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{1}{\gamma} (\mathbf{x}_t - \mathbf{x}_{t+1})^\top (\mathbf{x}_t - \mathbf{x}^\star).$$

- Apply $2\mathbf{v}^\top \mathbf{w} = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - \|\mathbf{v}-\mathbf{w}\|^2$ (cosine theorem) to rewrite

$$\begin{aligned} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) &= \frac{1}{2\gamma} \left(\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \right) \\ &= \frac{\gamma}{2} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \left(\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \right) \end{aligned}$$

- Sum this up over the first T iterations:

$$\sum_{t=0}^{T-1} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) = \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \left(\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \|\mathbf{x}_T - \mathbf{x}^\star\|^2 \right)$$

Vanilla analysis II

pure

- Remember:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star)$$

- Plug this lower bound into Vanilla Analysis:

$$\begin{aligned}
\sum_{t=0}^{T-1} \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) &\leq \sum_{t=0}^{T-1} \mathbf{g}_t^\top (\mathbf{x}_t - \mathbf{x}^\star) \\
&= \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \left(\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 - \|\mathbf{x}_T - \mathbf{x}^\star\|^2 \right) \\
&\leq \frac{\gamma}{2} \sum_{t=0}^{T-1} \|\mathbf{g}_t\|^2 + \frac{1}{2\gamma} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2
\end{aligned}$$

Vanilla analysis III

Result: upper bound for the average error $f(\mathbf{x}_t) - f(\mathbf{x}^\star)$ over the first T iterations:

$$\frac{1}{T}\sum_{t=0}^{T-1}\big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \leq \frac{1}{T}\big(\frac{\gamma}{2}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2\big)$$

Last iterate $\mathbf{x}_{T-1}$ is not necessarily the best one.

Open questions:

- Can we control the $\|\mathbf{g}_t\|^2$?
- How do we choose γ?

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps

Assume that all gradients of f are bounded in norm.

- Equivalent to f being Lipschitz (Theorem 2.10; Exercise 20).
- Rules out many interesting functions (for example, the "supermodel" $f(x) = x^2$)

Theorem 3.1

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and $\|\nabla f(\mathbf{x})\| \leq B$ for all $\mathbf{x}$. Choosing the stepsize

$$\gamma := \frac{R}{B\sqrt{T}},$$

this is the theoritical analysis about how to choose the learning rate.

gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{RB}{\sqrt{T}}.$$

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps II ($\frac{1}{T}\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{RB}{\sqrt{T}}$)

Proof.

- Plug $\|\mathbf{x}_0 - \mathbf{x}^\star\| \leq R$ and $\|\mathbf{g}_t\| \leq B$ into Vanilla Analysis II:

$$\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1}\|\mathbf{g}_t\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \leq \frac{\gamma}{2}B^2T + \frac{1}{2\gamma}R^2.$$

- choose γ such that

$$q(\gamma) = \frac{\gamma}{2}B^2T + \frac{R^2}{2\gamma}$$

  is minimized.
- Solving $q'(\gamma) = 0$ yields the minimum $\gamma = \frac{R}{B\sqrt{T}}$, and $q(R/(B\sqrt{T})) = RB\sqrt{T}$.
- Dividing by T, the result follows.

□

Lipschitz convex functions: $\mathcal{O}(1/\varepsilon^2)$ steps III ($\frac{1}{T}\sum_{t=0}^{T-1}(f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{RB}{\sqrt{T}}$)

$$T \geq \frac{R^2B^2}{\varepsilon^2} \quad \Rightarrow \quad \text{average error} \leq \frac{RB}{\sqrt{T}} \leq \varepsilon.$$

Advantages:

- dimension-independent (no d in the bound)!
- holds for both average, or best iterate

Disadvantages:

- R, B might be large and/or depend on the dimension d
- Slow: $10,000 \cdot R^2B^2$ iterations for error 0.01

In Practice:

What if we don't know R and B? $\rightarrow$ Exercise 24 (having to know R can't be avoided)

Smooth functions

"Not too curved"

Definition 3.2

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be differentiable, $X \subseteq \mathbf{dom}(f)$ convex, $L \in \mathbb{R}_+$. f is called smooth (with parameter L) over X if

just require convex set

$$f(\mathbf{y}) \leq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{L}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

this is the upper bound of f(y)

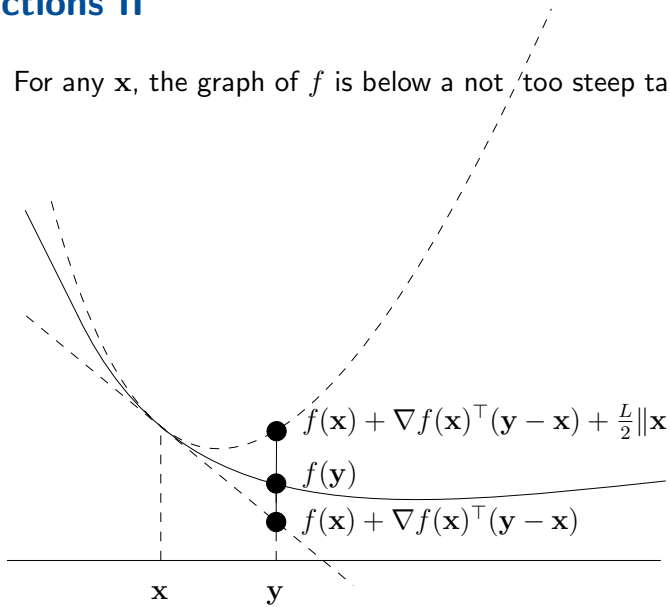
f smooth $:\Leftrightarrow$ f smooth over $\mathbb{R}^d$.

Definition does not require convexity (useful later)

Not to be confused with smooth functions in mathematical analysis (where smooth means infinitely often differentiable).

Smooth functions II

Smoothness: For any $\mathbf{x}$, the graph of f is below a not too steep tangent paraboloid at $(\mathbf{x}, f(\mathbf{x}))$:



this term is qudratic

Smooth functions: Simple characterization

$$\text{Smoothness of } f(\mathbf{x}) \quad = \quad \text{convexity of } \frac{L}{2}\mathbf{x}^\top\mathbf{x} - f(\mathbf{x}).$$

Exercise 21

Suppose that $\mathbf{dom}(f)$ is open and convex, and that $f : \mathbf{dom}(f) \to \mathbb{R}$ is differentiable. Let $L \in \mathbb{R}_+$. Then the following two statements are equivalent.

(i) f is smooth with parameter L.

(ii) g defined by $g(\mathbf{x}) = \frac{L}{2}\mathbf{x}^\top\mathbf{x} - f(\mathbf{x})$ is convex over $\mathbf{dom}(g) := \mathbf{dom}(f)$.

this one is convex, might be used to prove the convexity of the function.

Smooth functions III

Example: $f(x) = x^2$.

$$f(y) = y^2 \quad = \quad x^2 + 2x(y-x) + (x-y)^2 = f(x) + f'(x)(y-x) + \frac{2}{2}(x-y)^2$$

Hence, f is smooth with parameter 2.

- In general: quadratic functions are smooth (Exercise 22).
- Operations that preserve smoothness: same as for convexity, except the pointwise maximum. . . which is in general not differentiable, so smoothness doesn't apply.

Exercise 25

(i) Let $f_1, f_2, \ldots, f_m$ be functions that are smooth with parameters $L_1, L_2, \ldots, L_m$, and let $\lambda_1, \lambda_2, \ldots, \lambda_m \in \mathbb{R}_+$. Then the function $f := \sum_{i=1}^m \lambda_i f_i$ is smooth with parameter $\sum_{i=1}^m \lambda_i L_i$.

(ii) Let f be smooth with parameter L, and let $g(\mathbf{x}) = A\mathbf{x} + \mathbf{b}$, for $A \in \mathbb{R}^{d \times m}$ and $\mathbf{b} \in \mathbb{R}^d$. Then the function $f \circ g$ is smooth with parameter $L\|A\|^2$, where is $\|A\|$ is the spectral norm of A.

this definition is not clear, hard to picture.

Smooth functions IV

this means the bounded gradients of f

- Bounded gradients $\Leftrightarrow$ Lipschitz continuity of f
- Smoothness $\Leftrightarrow$ Lipschitz continuity of ∇f (in the convex case).

this means the smoothness

Lemma 3.5

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable. The following two statements are equivalent.

(i) f is smooth with parameter L.

(ii) $\|\nabla f(\mathbf{x}) - \nabla f(\mathbf{y})\| \leq L\|\mathbf{x} - \mathbf{y}\|$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$.

there are uper and lower bounds for the gradients of f.

Sufficient decrease

Lemma 3.7

Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and smooth with parameter L. With stepsize

$$\gamma := \frac{1}{L},$$

gradient descent satisfies

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2, \quad t \geq 0.$$

this is the theoritical indicator

Remark

(i) *This doesn't require convexity.*

(ii) *This already holds if f is smooth with parameter L over the line segment connecting $\mathbf{x}_t$ and $\mathbf{x}_{t+1}$.*

Sufficient decrease II

$$f(\mathbf{x}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2.$$

Proof.

Use smoothness and definition of gradient descent ($\mathbf{x}_{t+1} - \mathbf{x}_t = -\nabla f(\mathbf{x}_t)/L$):

$$\begin{aligned} f(\mathbf{x}_{t+1}) &\leq f(\mathbf{x}_t) + \nabla f(\mathbf{x}_t)^\top(\mathbf{x}_{t+1} - \mathbf{x}_t) + \frac{L}{2}\|\mathbf{x}_t - \mathbf{x}_{t+1}\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{L}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2 \\ &= f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2. \end{aligned}$$

□

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps

this is bigger.

Theorem 3.8

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L. Choosing stepsize

we have convex in this one

$$\gamma := \frac{1}{L},$$

gradient descent yields

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps II

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Proof.

Vanilla Analysis II:

$$\sum_{t=0}^{T-1} \big(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\big) \leq \frac{\gamma}{2}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2\gamma}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

This time, we can bound the squared gradients by sufficient decrease:

$$\frac{1}{2L}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 \leq \sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}_{t+1})) = f(\mathbf{x}_0) - f(\mathbf{x}_T).$$

□

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps II ($f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$)

Putting it together with $\gamma = 1/L$:

$$\begin{aligned}\sum_{t=0}^{T-1} (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) &\leq \frac{1}{2L}\sum_{t=0}^{T-1} \|\nabla f(\mathbf{x}_t)\|^2 + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2 \\ &\leq f(\mathbf{x}_0) - f(\mathbf{x}_T) + \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.\end{aligned}$$

Rewriting:

$$\sum_{t=1}^{T} (f(\mathbf{x}_t) - f(\mathbf{x}^\star)) \leq \frac{L}{2}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

As last iterate is the best (sufficient decrease!):

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{1}{T}\left(\sum_{t=1}^{T} (f(\mathbf{x}_t) - f(\mathbf{x}^\star))\right) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

Smooth convex functions: $\mathcal{O}(1/\varepsilon)$ steps III ($f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T}\|\mathbf{x}_0 - \mathbf{x}^\star\|^2$)

$R^2 := \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$.

$$T \geq \frac{R^2 L}{2\varepsilon} \quad \Rightarrow \quad \text{error} \leq \frac{L}{2T} R^2 \leq \varepsilon.$$

- $50 \cdot R^2 L$ iterations for error $0.01 \ldots$
- $\ldots$ as opposed to $10,000 \cdot R^2 B^2$ in the Lipschitz case

In Practice:

What if we don't know the smoothness parameter L?

$\rightarrow$ Exercise 26

Smooth convex functions: less than $\mathcal{O}(1/\varepsilon)$ steps?

we need to know the value of L and R.

Fixing L and $R = \|\mathbf{x}_0 - \mathbf{x}^\star\|$, the error of gradient descent after T steps is $O(1/T)$.

Lee and Wright [LW19]:

- A better upper bound of $o(1/T)$ holds.
- A lower bound of $\Omega(1/T^{1+\delta})$ also holds, for any fixed $\delta > 0$.

So, gradient descent is slightly faster on smooth functions than what we proved, but not significantly.

First-order methods: less than $\mathcal{O}(1/\varepsilon)$ steps?

Maybe gradient descent is not the best possible algorithm?

After all, it is just some algorithm that uses gradient information.

First-order method:

- An algorithm that gains access to f only via an oracle that is able to return values of f and ∇f at arbitrary points.
- Gradient descent is a specific first-order method.

What is the best first-order method for smooth convex functions, the one with the smallest upper bound on the number of oracle calls in the worst case?

Nemirovski and Yudin 1979 [NY83]: every first-order method needs in the worst case $\Omega(1/\sqrt{\varepsilon})$ steps (gradient evaluations) in order to achieve an additive error of ε on smooth functions.

There is a gap between $O(1/\varepsilon)$ (gradient descent) and the lower bound!

Acceleration for smooth convex functions: $\mathcal{O}(1/\sqrt{\varepsilon})$ steps

Nesterov 1983 [Nes83, Nes18]: There is a first-order method that needs only $O(1/\sqrt{\varepsilon})$ steps on smooth convex functions, and by the lower bound of Nemirovski and Yudin, this is a best possible algorithm!

The algorithm is known as (Nesterov's) accelerated gradient descent.

A number of (similar) optimal algorithms with other proofs of the $\mathcal{O}(1/\sqrt{\varepsilon})$ upper bound are known, but there is no well-established "simplest proof".

Here: a proof based on potential functions [BG17]. Proof is simple but not very instructive (it works, but it's not clear why).

Nesterov's accelerated gradient descent

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex, differentiable, and smooth with parameter L. Choose $\mathbf{z}_0 = \mathbf{y}_0 = \mathbf{x}_0$ arbitrary. For $t \geq 0$, set

$$\begin{aligned} \mathbf{y}_{t+1} &:= \mathbf{x}_t - \frac{1}{L}\nabla f(\mathbf{x}_t) \\ \mathbf{z}_{t+1} &:= \mathbf{z}_t - \frac{t+1}{2L}\nabla f(\mathbf{x}_t) \\ \mathbf{x}_{t+1} &:= \frac{t+1}{t+3}\mathbf{y}_{t+1} + \frac{2}{t+3}\mathbf{z}_{t+1}. \end{aligned}$$

- Perform a "smooth step" from $\mathbf{x}_t$ to $\mathbf{y}_{t+1}$.
- Perform a more aggressive step from $\mathbf{z}_t$ to $\mathbf{z}_{t+1}$.
- Next iterate $\mathbf{x}_{t+1}$ is a weighted average of $\mathbf{y}_{t+1}$ and $\mathbf{z}_{t+1}$, where we compensate for the more aggressive step by giving $\mathbf{z}_{t+1}$ a relatively low weight.

Why should this work??

Nesterov's accelerated gradient descent: Error bound

Theorem 3.9

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; furthermore, suppose that f is smooth with parameter L. Accelerated gradient descent yields

$$f(\mathbf{y}_T) - f(\mathbf{x}^\star) \leq \frac{2L \|\mathbf{z}_0 - \mathbf{x}^\star\|^2}{T(T+1)}, \quad T > 0.$$

To reach error at most ε, accelerated gradient descent therefore only needs $O(1/\sqrt{\varepsilon})$ steps instead of $O(1/\varepsilon)$.

Recall the bound for gradient descent:

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2T} \|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Nesterov's accelerated gradient descent: The potential function

Idea: assign a potential $\Phi(t)$ to each time t and show that $\Phi(t+1) \leq \Phi(t)$.

Out of the blue: let's define the potential as

$$\Phi(t) := t(t+1)\left(f(\mathbf{y}_t) - f(\mathbf{x}^\star)\right) + 2L\left\|\mathbf{z}_t - \mathbf{x}^\star\right\|^2.$$

If we can show that the potential always decreases, we get

$$\underbrace{T(T+1)\left(f(\mathbf{y}_T) - f(\mathbf{x}^\star)\right) + 2L\left\|\mathbf{z}_T - \mathbf{x}^\star\right\|^2}_{\Phi(T)} \leq \underbrace{2L\left\|\mathbf{z}_0 - \mathbf{x}^\star\right\|^2}_{\Phi(0)}.$$

Rewriting this, we get the claimed error bound.

Potential function decrease: Three Ingredients

Sufficient decrease for the smooth step from $\mathbf{x}_t$ to $\mathbf{y}_{t+1}$:

$$f(\mathbf{y}_{t+1}) \leq f(\mathbf{x}_t) - \frac{1}{2L}\|\nabla f(\mathbf{x}_t)\|^2; \tag{1}$$

Vanilla analysis for the more aggressive step from $\mathbf{z}_t$ to $\mathbf{z}_{t+1}$: ($\gamma = \frac{t+1}{2L}$, $\mathbf{g}_t = \nabla f(\mathbf{x}_t)$):

$$\mathbf{g}_t^\top(\mathbf{z}_t - \mathbf{x}^\star) = \frac{t+1}{4L}\|\mathbf{g}_t\|^2 + \frac{L}{t+1}\left(\|\mathbf{z}_t - \mathbf{x}^\star\|^2 - \|\mathbf{z}_{t+1} - \mathbf{x}^\star\|^2\right); \tag{2}$$

Convexity (graph of f is above the tangent hyperplane at $\mathbf{x}_t$):

$$f(\mathbf{x}_t) - f(\mathbf{w}) \leq \mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{w}), \quad \mathbf{w} \in \mathbb{R}^d. \tag{3}$$

Potential function decrease: Proof

Definition of potential:

$$\begin{aligned}\Phi(t) &= t(t+1)\left(f(\mathbf{y}_t) - f(\mathbf{x}^\star)\right) + 2L\left\|\mathbf{z}_t - \mathbf{x}^\star\right\|^2,\\ \Phi(t+1) &= t(t+1)\left(f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star)\right) + 2(t+1)\left(f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star)\right) + 2L\left\|\mathbf{z}_{t+1} - \mathbf{x}^\star\right\|^2.\end{aligned}$$

Now, prove that $\Delta := (\Phi(t+1) - \Phi(t))/(t+1) \leq 0$:

$$\begin{aligned}\Delta &= t\left(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)\right) + 2\left(f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star)\right) + \frac{2L}{t+1}\left(\left\|\mathbf{z}_{t+1} - \mathbf{x}^\star\right\|^2 - \left\|\mathbf{z}_t - \mathbf{x}^\star\right\|^2\right)\\ &\overset{(2)}{=} t\left(f(\mathbf{y}_{t+1}) - f(\mathbf{y}_t)\right) + 2\left(f(\mathbf{y}_{t+1}) - f(\mathbf{x}^\star)\right) + \frac{t+1}{2L}\|\mathbf{g}_t\|^2 - 2\mathbf{g}_t^\top(\mathbf{z}_t - \mathbf{x}^\star)\\ &\overset{(1)}{\leq} t\left(f(\mathbf{x}_t) - f(\mathbf{y}_t)\right) + 2\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) - \frac{1}{2L}\|\mathbf{g}_t\|^2 - 2\mathbf{g}_t^\top(\mathbf{z}_t - \mathbf{x}^\star)\\ &\leq t\left(f(\mathbf{x}_t) - f(\mathbf{y}_t)\right) + 2\left(f(\mathbf{x}_t) - f(\mathbf{x}^\star)\right) - 2\mathbf{g}_t^\top(\mathbf{z}_t - \mathbf{x}^\star)\\ &\overset{(3)}{\leq} t\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{y}_t) + 2\mathbf{g}_t^\top(\mathbf{x}_t - \mathbf{x}^\star) - 2\mathbf{g}_t^\top(\mathbf{z}_t - \mathbf{x}^\star)\\ &= \mathbf{g}_t^\top\left((t+2)\mathbf{x}_t - t\mathbf{y}_t - 2\mathbf{z}_t\right) \overset{\text{(algo)}}{=} \mathbf{g}_t^\top\mathbf{0} = 0. \quad \square\end{aligned}$$

Interlude

Recall: the "supermodel" $f(x) = x^2$ is smooth with parameter $L = 2$.

Consequently, gradient descent attains additive error ε after at most $O(1/\varepsilon)$ steps, and accelerated gradient descent does so after at most $O(1/\sqrt{\varepsilon})$ steps.

Concretely (our result for gradient descent on smooth functions, using $x^* = 0$):

$$f(x_T) \leq \frac{1}{T} x_0^2.$$

Reality (gradient descent with the smooth stepsize $\gamma = 1/L = 1/2$):

$$x_{t+1} = x_t - \frac{1}{2}\nabla f(x_t) = x_t - \frac{1}{2} 2 x_t = 0.$$

Done after one step!

But this is only because f is very beautiful, and we have picked the best possible smoothness parameter for it. Let's look at a more realistic scenario...

Interlude II

Didn't look too closely: the "supermodel" $f(x) = x^2$ is smooth with parameter $L = 4$.

Concretely (our result for gradient descent on smooth functions, using $x^* = 0$):

$$f(x_T) \leq \frac{2}{T} x_0^2.$$

Reality (gradient descent with the smooth stepsize $\gamma = 1/L = 1/4$):

$$x_{t+1} = x_t - \frac{1}{4}\nabla f(x_t) = x_t - \frac{1}{4} 2 x_t = \frac{x_t}{2}.$$

Hence,

$$f(x_T) = f\left(\frac{x_0}{2^T}\right) = \frac{1}{2^{2T}} x_0^2.$$

Exponentially better than our result for smooth functions! Which additional property of the "supermodel" is responsible for this? Can we generalize this property and also handle other smooth functions much faster?

Strongly convex functions

"Not too flat"

Definition 3.10

Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be a convex and differentiable function, $X \subseteq \mathbf{dom}(f)$ convex and $\mu \in \mathbb{R}_+, \mu > 0$. Function f is called strongly convex (with parameter μ) over X if

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \nabla f(\mathbf{x})^\top (\mathbf{y} - \mathbf{x}) + \frac{\mu}{2}\|\mathbf{x} - \mathbf{y}\|^2, \quad \forall \mathbf{x}, \mathbf{y} \in X.$$

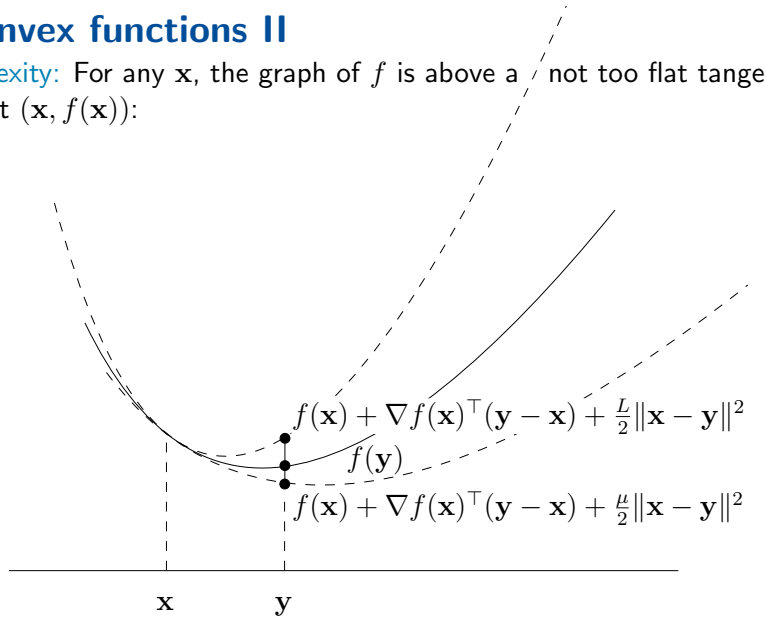
f strongly convex :$\Leftrightarrow$ f strongly convex over $\mathbb{R}^d$.

Exercise 29

If f is strongly convex with parameter $\mu > 0$, then f is strictly convex and has a unique global minimum.

Strongly convex functions II

Strong convexity: For any $\mathbf{x}$, the graph of f is above a not too flat tangent paraboloid at $(\mathbf{x}, f(\mathbf{x}))$:



Strongly convex functions: Simple characterization

$$\text{Strong convexity of } f(\mathbf{x}) \quad = \quad \text{convexity of } f(\mathbf{x}) - \frac{\mu}{2}\mathbf{x}^\top \mathbf{x}.$$

Exercise 28

Suppose that $\mathbf{dom}(f)$ is open and convex, and that $f : \mathbf{dom}(f) \to \mathbb{R}$ is differentiable. Let $\mu \in \mathbb{R}_+$. Then the following two statements are equivalent.

(i) f is strongly convex with parameter μ.

(ii) g defined by $g(\mathbf{x}) = f(\mathbf{x}) - \frac{\mu}{2}\mathbf{x}^\top \mathbf{x}$ is convex over $\mathbf{dom}(g) := \mathbf{dom}(f)$.

Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps

Want to show: $\lim_{t\to\infty} \mathbf{x}_t = \mathbf{x}^\star$

Vanilla Analysis:

$$\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star) = \frac{\gamma}{2}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{2\gamma}\left(\|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right)$$

Now use stronger lower bound on left hand side, coming from strong convexity:

$$\nabla f(\mathbf{x}_t)^\top(\mathbf{x}_t - \mathbf{x}^\star) \geq f(\mathbf{x}_t) - f(\mathbf{x}^\star) + \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2$$

Putting it together:

$$f(\mathbf{x}_t) - f(\mathbf{x}^\star) \leq \frac{1}{2\gamma}\left(\gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \|\mathbf{x}_t - \mathbf{x}^\star\|^2 - \|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2\right) - \frac{\mu}{2}\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

Rewriting:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq 2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + (1 - \mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps II

$$\underline{\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2} \leq 2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \underline{(1-\mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2}.$$

Squared distance to $\mathbf{x}^\star$ goes down by a constant factor, up to some "noise".

Theorem 3.14

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum $\mathbf{x}^\star$; suppose that f is smooth with parameter L and strongly convex with parameter $\mu > 0$. Choosing $\gamma := \frac{1}{L}$, gradient descent with arbitrary $\mathbf{x}_0$ satisfies the following two properties.

(i) Squared distances to $\mathbf{x}^\star$ are geometrically decreasing:

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)\|\mathbf{x}_t - \mathbf{x}^\star\|^2, \quad t \geq 0.$$

(ii) The absolute error after T iterations is exponentially small in T:

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2, \quad T > 0.$$

Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps III

$$\underline{\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2} \leq 2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 + \underline{(1-\mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2}.$$

Proof of (i) ($\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)\|\mathbf{x}_t - \mathbf{x}^\star\|^2$).

Bounding the noise: $\gamma = 1/L$, sufficient decrease

$$\begin{aligned} 2\gamma(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \gamma^2\|\nabla f(\mathbf{x}_t)\|^2 &= \frac{2}{L}(f(\mathbf{x}^\star) - f(\mathbf{x}_t)) + \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 \\ &\leq \frac{2}{L}(f(\mathbf{x}_{t+1}) - f(\mathbf{x}_t)) + \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 \\ &\leq -\frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 + \frac{1}{L^2}\|\nabla f(\mathbf{x}_t)\|^2 = 0. \end{aligned}$$

Hence, the noise is nonpositive, and we get (i):

$$\|\mathbf{x}_{t+1} - \mathbf{x}^\star\|^2 \leq (1-\mu\gamma)\|\mathbf{x}_t - \mathbf{x}^\star\|^2 = \left(1 - \frac{\mu}{L}\right)\|\mathbf{x}_t - \mathbf{x}^\star\|^2.$$

Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps III

Proof of (ii) ($f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$).

From (i):

$$\|\mathbf{x}_T - \mathbf{x}^\star\|^2 \leq \left(1 - \frac{\mu}{L}\right)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

Smoothness together with $\nabla f(\mathbf{x}^\star) = \mathbf{0}$:

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \nabla f(\mathbf{x}^\star)^\top (\mathbf{x}_T - \mathbf{x}^\star) + \frac{L}{2}\|\mathbf{x}_T - \mathbf{x}^\star\|^2 = \frac{L}{2}\|\mathbf{x}_T - \mathbf{x}^\star\|^2.$$

Putting it together:

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2}\|\mathbf{x}_T - \mathbf{x}^\star\|^2 \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

□

Smooth and strongly convex functions: $\mathcal{O}(\log(1/\varepsilon))$ steps IV

$$f(\mathbf{x}_T) - f(\mathbf{x}^\star) \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T \|\mathbf{x}_0 - \mathbf{x}^\star\|^2.$$

$R^2 := \|\mathbf{x}_0 - \mathbf{x}^\star\|^2$.

$$T \geq \frac{L}{\mu} \ln\left(\frac{R^2 L}{2\varepsilon}\right) \quad \Rightarrow \quad \text{error } \leq \frac{L}{2}\left(1 - \frac{\mu}{L}\right)^T R^2 \leq \varepsilon.$$

- $\frac{L}{\mu} \ln(50 \cdot R^2 L)$ iterations for error 0.01 ...
- ... as opposed to $50 \cdot R^2 L$ in the smooth case

In Practice:

What if we don't know the smoothness parameter L?

$\rightarrow$ (similar to) Exercise 26

Bibliography


Nikhil Bansal and Anupam Gupta.
Potential-function proofs for first-order methods.
CoRR, abs/1712.04581, 2017.


Ching-Pei Lee and Stephen Wright.
First-order algorithms converge faster than $o(1/k)$ on convex problems.
In *ICML - Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pages 3754–3762, Long Beach, California, USA, 2019.


Yurii Nesterov.
A method of solving a convex programming problem with convergence rate $o(1/k^2)$.
Soviet Math. Dokl., 27(2), 1983.


Yurii Nesterov.
Lectures on Convex Optimization, volume 137 of *Springer Optimization and Its Applications*.
Springer, second edition, 2018.


Arkady. S. Nemirovsky and D. B. Yudin.
Problem complexity and method efficiency in optimization.
Wiley, 1983.