SDSC6015 Stochastic Optimization for Machine Learning

Lu Yu

Department of Data Science, City University of Hong Kong

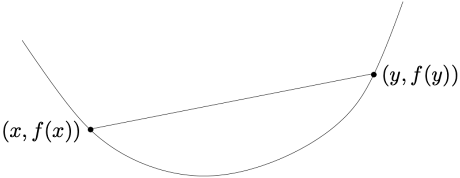
September 11, 2025

Convex Function and Convex Optimization

Recap: Convex Functions

Definition: A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if
(i) $\mathbf{dom}(f)$ is a convex set and
(ii) for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbf{dom}(f)$ and λ with $0 \leqslant \lambda \leqslant 1$, we have

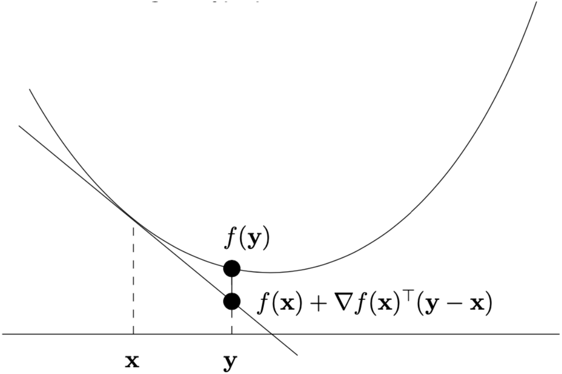
$$f(\lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{y}) \leqslant \lambda f(\boldsymbol{x}) + (1-\lambda) f(\boldsymbol{y}).$$



Geometrically: The line segment between $(\boldsymbol{x}, f(\boldsymbol{x}))$ and $(\boldsymbol{y}, f(\boldsymbol{y}))$ lies above the graph of f.

$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y}-\boldsymbol{x}), \qquad \boldsymbol{x},\boldsymbol{y} \in \mathbf{dom}(f).$$

Graph of f is above all its tangent hyperplanes.



Recap: Differentiable Functions

- A function $f : \mathbb{R}^d \to \mathbb{R}$ is differentiable at a point $\boldsymbol{x}_0$ if it can be well-approximated by a linear function near that point.
- There exists a gradient $\nabla f(\boldsymbol{x}_0)$ such that

$$f(\boldsymbol{x}_0 + \mathbf{h}) \approx f(\boldsymbol{x}_0) + \nabla f(\boldsymbol{x}_0) \cdot \mathbf{h}\,,$$

where $\mathbf{h}$ is a small change around $\boldsymbol{x}_0$.

Differentiable Functions

- If f is differentiable at every point in its domain, it is called a differentiable function.
- The graph of a differentiable function has a non-vertical tangent line at each interior point in its domain.

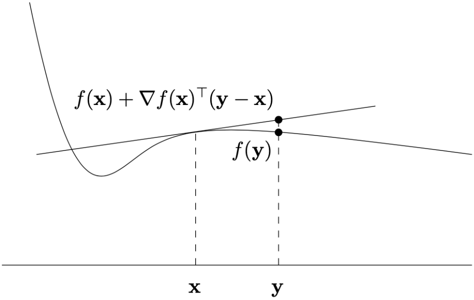


Figure: Graph of the affine function $f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top(\boldsymbol{y} - \boldsymbol{x})$ is a tangent hyperplane to the graph of f at $(\boldsymbol{x}, f(\boldsymbol{x}))$.

Recap: Convex Optimization

Convex Optimization Problems are of the form

$$\min_{x \in \mathbb{R}^d} f(x),$$

where

- f is a **convex** and **differentiable** function
- $\mathbb{R}^d$ is convex
- $\boldsymbol{x}^*$ is the minimizer of function f :

$$\boldsymbol{x}^* = \arg\min_{\boldsymbol{x} \in \mathbb{R}^d} f(\boldsymbol{x})$$

Note: there can be several global minima $\boldsymbol{x}_1^* \neq \boldsymbol{x}_2^*$ with $f(\boldsymbol{x}_1^*) = f(\boldsymbol{x}_2^*)$.

The Algorithm

- **Assumptions**: $f : \mathbb{R}^d \to \mathbb{R}$ is convex, differentiable, has a global minimum $\boldsymbol{x}^*$.
- **Goal**: Find $\boldsymbol{x} \in \mathbb{R}^d$ such that

$$f(\boldsymbol{x}) - f(\boldsymbol{x}^*) < \varepsilon\,,$$

where $\varepsilon > 0$ is small.

Gradient Descent

Gradient Descent

Goal: minimizing the convex and differentiable function $f : \mathbb{R}^d \to \mathbb{R}$

- **Fact**: $\nabla f(\boldsymbol{x})$ provides the direction and rate of the **fastest increase** of $f(\boldsymbol{x})$
- Minimizing the function $f(\boldsymbol{x})$ via moving against $\nabla f(\boldsymbol{x})$

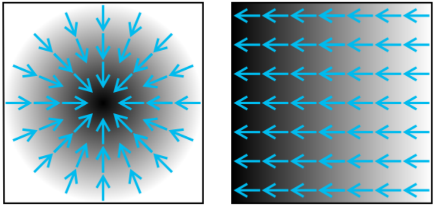


Figure: The gradient, represented by the blue arrows, denotes the direction of greatest change of a scalar function. The values of the function are represented in greyscale and increase in value from white (low) to dark (high).

Gradient Descent

Update rule for gradient descent:

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_{k+1} \nabla f(\boldsymbol{x}_k)$$

- $\boldsymbol{x}_k$: current point (parameters or variables).
- η_k : step size (learning rate), a positive scalar determining how far we move in the gradient direction.
- $\boldsymbol{x}_{k+1}$: next point after the update.

Example:

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

This is a convex function with its minimum at $x = 0$.

- gradient $\nabla f(x) = x$
- GD update with a fixed step size:

$$x_{t+1} = x_t - \eta x_t = x_t(1 - \eta).$$

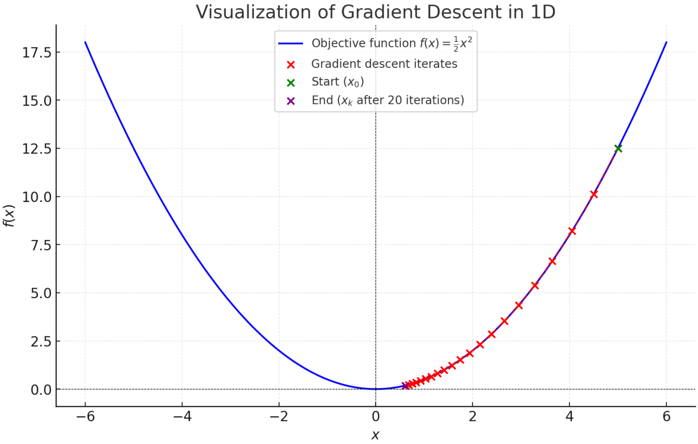
- In general, after k iterations, the GD iterate is:

$$x_t = x_{t-1}(1-\eta) = x_{t-2}(1-\eta)^2 = \cdots = x_0(1-\eta)^k$$

  As $t \to \infty$, if $0 < \eta < 1$, the iterates converge to 0, which is the minimum of the function.

Gradient Descent

- Step size $\eta = 0.1$
- Starting point $x_0 = 5$



Gradient Descent - Vanilla Analysis

How to bound $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)$?

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

$$\boldsymbol{g}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}^*) = \frac{1}{\eta}(\boldsymbol{x}_t - \boldsymbol{x}_{t+1})^\top(\boldsymbol{x}_t - \boldsymbol{x}^*)$$

- Apply $2\boldsymbol{v}^\top\boldsymbol{w} = \|\boldsymbol{v}\|^2 + \|\boldsymbol{w}\|^2 - \|\boldsymbol{v} - \boldsymbol{w}\|^2$ to rewrite

$$\begin{aligned}\boldsymbol{g}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}^*) &= \frac{1}{2\eta}(\|\boldsymbol{x}_t - \boldsymbol{x}_{t+1}\|^2 + \|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2)\\ &= \frac{\eta}{2}\|\boldsymbol{g}_t\|^2 + \frac{1}{2\eta}(\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|^2)\end{aligned}$$

- Sum this up over the first T iterations:

$$\sum_{t=0}^{T-1}\boldsymbol{g}_t^\top(\boldsymbol{x}_t - \boldsymbol{x}^*) = \frac{\eta}{2}\sum_{t=0}^{T-1}\|\boldsymbol{g}_t\|^2 + \frac{1}{2\eta}(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2)$$

Gradient Descent - Vanilla Analysis

Using first-order characterization of convexity:

$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \forall \boldsymbol{x}, \boldsymbol{y}.$$

- with $\boldsymbol{x} = \boldsymbol{x}_t, \boldsymbol{y} = \boldsymbol{x}^*$:

  $$f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leqslant \boldsymbol{g}_t^\top (\boldsymbol{x}_t - \boldsymbol{x}^*)$$

  giving

  $$\sum_{t=0}^{T-1} (f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \leqslant \frac{\eta}{2} \sum_{t=0}^{T-1} \|\boldsymbol{g}_t\|^2 + \frac{1}{2\eta} (\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2), \tag{1}$$

  an upper bound for the average error $f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)$ over steps
- Stepsize η is crucial!

Gradient Descent for Lipschitz Convex Functions

Assume that all gradients of f are bounded in norm.

- Equivalent to f being Lipschitz (see notes)
- Rules out many interesting functions (for example, $f(x) = x^2$)

Theorem 1

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

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

gradient descent yields

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leqslant \frac{RB}{\sqrt{T}}.$$

Gradient Descent on Lipschitz Convex Functions

Proof.

- Plugging $\|\boldsymbol{x}_0 - \boldsymbol{x}^*\| \leqslant R$ and $\|\boldsymbol{g}_t\| \leqslant B$ into display (1) in Vanilla Analysis

$$\sum_{t=0}^{T-1}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \leqslant \frac{\eta}{2}\sum_{t=0}^{T-1}\|\boldsymbol{g}_t\|^2 + \frac{1}{2\eta}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 \leqslant \frac{\eta}{2}B^2T + \frac{1}{2\eta}R^2$$

- Choosing η such that

$$h(\eta) := \frac{\eta}{2}B^2T + \frac{R^2}{2\eta}$$

is minimized.

- Solving $h'(\eta) = 0$ yields the minimum $\eta = \frac{R}{B\sqrt{T}}$ and $h(\frac{R}{B\sqrt{T}}) = RB\sqrt{T}$
- Dividing by T, the result follows.

Gradient Descent for Lipschitz Convex Functions

$$T \geqslant \frac{R^2 B^2}{\varepsilon^2} \Rightarrow \text{ average error } \leqslant \frac{RB}{\sqrt{T}} \leqslant \varepsilon$$

Advantages:

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

In Practice: What if we don't know R and B?

Gradient Descent for Lipschitz Convex Functions

Practical Recommendation

If B and R are unknown:

- Start with a small, constant step size (e.g., $\eta = 0.01$)
- Monitor the convergence behavior; if the method oscillates or diverges, reduce η.
- Alternatively, use a decreasing step size (e.g., $\eta_t = \frac{\eta_0}{\sqrt{t+1}}$) or an adaptive method (e.g., Adam).

Questions?

Definition

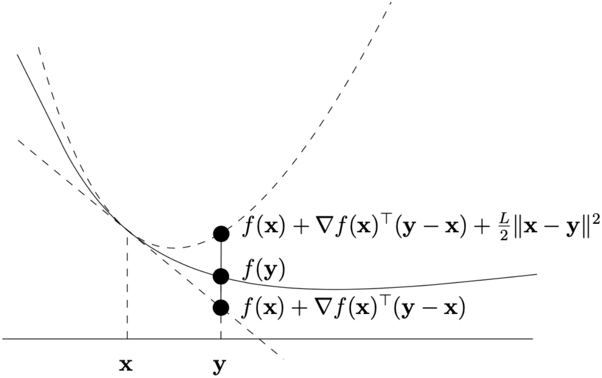
Let $f : \mathbf{dom}(f) \to \mathbb{R}$ be differentiable, $X \subseteq \mathbf{dom}(f)$, $L > 0$. f is called smooth (with parameter L) over X if

$$f(\boldsymbol{y}) \leqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}) + \frac{L}{2}\|\boldsymbol{x} - \boldsymbol{y}\|^2, \;\; \forall \boldsymbol{x}, \boldsymbol{y} \in X.$$

- **"Not too curved"**
- L quantifies how fast the gradient can change (see later)

Smooth Functions

Smoothness: f can be bounded above by a quadratic (paraboloid-shaped) function near any point.



Smooth Function

- In general: quadratic functions are smooth e.g. $f(x) = x^2$.
- Operations that preserve smoothness (the same that preserve convexity):

Lemma 1

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

(ii) Let f be a smooth function with parameter L, and let $g : \mathbb{R}^m \to \mathbb{R}^d$ an affine function, meaning that $g(\boldsymbol{x}) = A\boldsymbol{x} + \mathbf{b}$, for some matrix $A \in \mathbb{R}^{d\times m}$ and some vector $\mathbf{b} \in \mathbb{R}^d$. Then the function $f \circ g$ (that maps $\boldsymbol{x}$ to $f(A\boldsymbol{x} + \mathbf{b})$ is smooth with parameter $L\|A\|^2$, where is $\|A\|$ is the spectral norm of A.[1]

[1] the largest singular value of A

Convex Function v.s. Smooth Function

In the convex case:

- Bounded gradient $\Leftrightarrow$ Lipschitz continuity of f
- Smoothness $\Leftrightarrow$ Lipschitz continuity of ∇f .

Lemma 2

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

(i) f is smooth with parameter L.

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

Proof in lecture slides of L. Vandenberghe, http://www.seas.ucla.edu/ vandenbe/236C/lectures/gradient.pdf.

Lemma 3 (sufficient decrease)
Let $f : \mathbb{R}^d \to \mathbb{R}$ be differentiable and smooth with parameter L. With stepsize $\eta = \frac{1}{L}$, gradient descent satisfies

$$f(\boldsymbol{x}_{t+1}) \leqslant f(\boldsymbol{x}_t) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|^2, \qquad t \geqslant 0.$$

This implies: $f(\boldsymbol{x}_0) \geqslant f(\boldsymbol{x}_1) \geqslant f(\boldsymbol{x}_2) \geqslant \cdots$

$$f(\boldsymbol{x}_{t+1}) \leqslant f(\boldsymbol{x}_t) - \frac{1}{2L}\|\nabla f(\boldsymbol{x}_t)\|^2, \qquad t \geqslant 0.$$

Proof.

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

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

Gradient Descent on Convex Smooth Functions

Theorem 2

Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and differentiable with a global minimum x^*; furthermore, suppose that f is smooth with parameter L. Choosing stepsize $\eta = \frac{1}{L}$, gradient descent yields

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leqslant \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \qquad T > 0.$$

Gradient Descent on Convex Smooth Functions

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leqslant \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2, \qquad T > 0.$$

Proof.

Inequality (1) in Vanilla Analysis:

$$\sum_{t=0}^{T-1}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \leqslant \frac{\eta}{2}\sum_{t=0}^{T-1}\|\nabla f(\boldsymbol{x}_t)\|^2 + \frac{1}{2\eta}(\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 - \|\boldsymbol{x}_T - \boldsymbol{x}^*\|^2),$$

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

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

Gradient Descent on Convex Smooth Functions

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

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

Rewriting:

$$\sum_{t=1}^{T}(f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*)) \leqslant \frac{L}{2}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 .$$

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

$$f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*) \leqslant \frac{1}{T}\Big(\sum_{t=1}^{T}(f(\boldsymbol{x}_T) - f(\boldsymbol{x}^*)\Big) \leqslant \frac{L}{2T}\|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 .$$

Gradient Descent on Convex Smooth Functions

$R^2 = \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2$.

$$T \geqslant \frac{R^2 L}{2\varepsilon} \qquad \Rightarrow \qquad \text{error} \leqslant \frac{L}{2T} R^2 \leqslant \varepsilon\,.$$

- $50 \cdot R^2 L$ iterations for error $\varepsilon = 0.01$
- 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?

Gradient Descent on Convex Smooth Functions

Solution: The idea is to start by guessing L

- **Initial Guess:** Start with a guess for L:

$$L = \frac{2\varepsilon}{R^2}.$$

  If this guess is correct, we can achieve the desired error in just **1 iteration**.

- **Refining the Guess:** If the guess is too small, double L and try again. We keep doubling L until the guess is large enough. This process works because eventually, the guessed L will be larger than or equal to the true smoothness parameter.

- **Checking if a Guess is Correct:** A guess for L is correct if the following condition holds:

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

  This condition can be checked directly during optimization.

Gradient Descent on Convex Smooth Functions

- **Number of Iterations:**
  - Once the correct L is found, the number of iterations needed to reach the desired error is:
$$\frac{2R^2L}{2\varepsilon}.$$
  - The total number of iterations, considering all the guesses (doubling the initial guess), is at most:
$$\frac{4R^2L}{2\varepsilon}.$$

This ensures the error bound ε is achieved efficiently.

Convergence Rate of Gradient Descent

Summary: For gradient descent with constant step size to achieve an average error bound:

$$\frac{1}{T}\sum_{t=0}^{T-1} f(\boldsymbol{x}_t) - f(\boldsymbol{x}^*) \leqslant \varepsilon$$

- Lipschitz convex functions: need $T = \mathcal{O}(1/\varepsilon^2)$ steps
- Smooth convex functions: need $T = \mathcal{O}(1/\varepsilon)$ steps.

Questions?

Subgradient Method

Subgradient

Recall: for convex and differentiable $f : \mathbb{R}^d \to \mathbb{R}$

$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + \nabla f(\boldsymbol{x})^\top (\boldsymbol{y} - \boldsymbol{x}), \qquad \forall \boldsymbol{x}, \boldsymbol{y}.$$

Definition

A subgradient of a convex function $f : \mathbb{R}^d \to \mathbb{R}$ at $\boldsymbol{x}$ is any $g \in \mathbb{R}^d$ such that

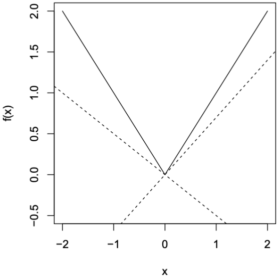
$$f(\boldsymbol{y}) \geqslant f(\boldsymbol{x}) + g^\top (\boldsymbol{y} - \boldsymbol{x}), \qquad \forall \boldsymbol{y}\,.$$

- Always exists (at any point in the interior of the domain of f)
- If f differentiable at $\boldsymbol{x}$, then $g = \nabla f(\boldsymbol{x})$ uniquely

Subgradient

Example 1: Consider $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = |x|$$

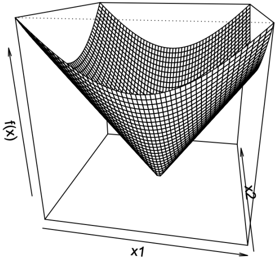


- For $x \neq 0$, unique subgradient $g = \text{sign}(x)$
- For $x = 0$, subgradient g is any element of $[-1, 1]$

Subgradient

Example 2: Consider $f : \mathbb{R}^d \to \mathbb{R}$

$$f(\boldsymbol{x}) = \|\boldsymbol{x}\|_2$$

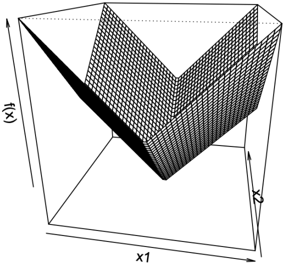


- For $\boldsymbol{x} \neq \mathbf{0}$, unique subgradient $g = \boldsymbol{x}/\|\boldsymbol{x}\|_2$
- For $\boldsymbol{x} = \mathbf{0}$, subgradient g is any element of $\{\boldsymbol{z} \in \mathbb{R}^d : \|\boldsymbol{z}\|_2 \leqslant 1\}$

Subgradient

Example 3: Consider $f : \mathbb{R}^d \to \mathbb{R}$

$$f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1 = \sum_{i=1}^{d} |x_i|$$



- For $x_i \neq 0$, unique i-th component $g_i = \text{sign}(x_i)$
- For $x_i = 0$, i-th component g_i is any element of $[-1, 1]$

Example 4: For the convex set $X \subset \mathbb{R}^d$, consider the indicator function $1_{\mathrm{X}} : \mathbb{R}^{\mathrm{d}} \to \mathbb{R}$

$$1_{\mathrm{X}}(\boldsymbol{x}) = \begin{cases} 0 & \text{if } \boldsymbol{x} \in X \\ +\infty & \text{if } \boldsymbol{x} \notin X \end{cases}$$

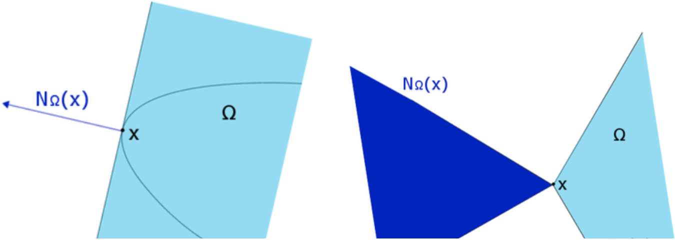
- Normal cone: given a convex set X and a point $\boldsymbol{x} \in X$, the normal cone to X to $\boldsymbol{x}$ is defined as

$$\mathcal{N}_X(\boldsymbol{x}) = \{g \in \mathbb{R}^d : g^\top \boldsymbol{x} \geqslant g^\top \boldsymbol{y} \text{ for all } \boldsymbol{y} \in X\}.$$

- For $\boldsymbol{x} \in X$, it holds that $\partial 1_{\mathrm{X}}(\boldsymbol{x}) = \mathcal{N}_{\mathrm{X}}(\boldsymbol{x})$ (see notes)

Normal Cone

The normal cone is the set of vectors pointing outward from a convex set at a specific point.



Subdifferential

Set of all subgradients of convex $f : \mathbb{R}^d \to \mathbb{R}$ is called the subdifferential:

$$\partial f(\boldsymbol{x}) = \{g \in \mathbb{R}^d : g \text{ is a subgradient of } f \text{ at } \boldsymbol{x}\}\,.$$

- ∂f is closed and convex
- If f is differentiable at $\boldsymbol{x}$, then $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$
- If $\partial f(\boldsymbol{x}) = \{g\}$, then f is differentiable at $\boldsymbol{x}$ and $\nabla f(\boldsymbol{x}) = g$

Why subgradients?

If you can compute subgradients, then you can minimize any convex function.

Optimality Condition

For any convex function $f : \mathbb{R}^d \to \mathbb{R}$

$$f(\boldsymbol{x}^*) = \min_{\boldsymbol{x}} f(\boldsymbol{x}) \qquad \Leftrightarrow \qquad \mathbf{0} \in \partial f(\boldsymbol{x}^*)$$

- $\boldsymbol{x}^*$ is a minimizer if and only if $\mathbf{0}$ is a subgradient of f at $\boldsymbol{x}^*$ (see notes)
- This is called the subgradient optimality condition
- Note the implication for a convex and differentiable function f, with $\partial f(\boldsymbol{x}) = \{\nabla f(\boldsymbol{x})\}$

Constrained Minimization

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) \qquad \text{subject to} \qquad \boldsymbol{x} \in X$$

Lemma 4 (from Lecture 1)

Suppose that $f : \mathbf{dom}(f) \to \mathbb{R}$ is convex and differentiable over an open domain $\mathbf{dom}(f) \subseteq \mathbb{R}^d$, and let $X \subseteq \mathbf{dom}(f)$ be a convex set. Point $\boldsymbol{x}^* \in X$ is a minimizer of f over X if and only if

$$\nabla f(\boldsymbol{x}^*)^\top (\boldsymbol{x} - \boldsymbol{x}^*) \geqslant 0, \quad \forall \boldsymbol{x} \in X.$$

Proof. (see notes)

Step 1: Recast the problem as

$$\min_{\boldsymbol{x}} f(\boldsymbol{x}) + 1_{\mathrm{X}}(\boldsymbol{x})$$

Step 2: Apply subgradient optimality

$$\mathbf{0} \in \partial(f(\boldsymbol{x}^*) + 1_{\mathrm{X}}(\boldsymbol{x}^*))$$

Questions?

Subgradient Method

Now consider convex function $f : \mathbb{R}^d \to \mathbb{R}$ convex, but not necessarily differential.

Subgradient method: like gradient descent, but replacing gradients with subgradients

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_{k+1} g_k$$

- $\boldsymbol{x}_k$: current point
- $g_k \in \nabla f(\boldsymbol{x}_k)$: any subgradient of f at $\boldsymbol{x}_k$
- $\eta_k > 0$: step size
- $\boldsymbol{x}_{k+1}$: next point after the update.

Caveat: Subgradient method is not necessarily a descent method!
e.g. $f(x) = |x|$ (non-smoothness causes oscillation)

Subgradient Method

Theorem 3: Assume $f : \mathbb{R}^d \to \mathbb{R}$ is convex and L-Lipschitz.

- For a fixed step size scheme

$$\eta_k = \eta, \qquad k = 1, 2, 3, \dots ,$$

subgradient method satisfies

$$\lim_{k\to\infty} f(\boldsymbol{x}_{\text{best}}^{(k)}) \leqslant f^* + L^2\eta/2 \,.$$

- For diminishing step sizes, satisfying

$$\sum_{k=1}^{\infty} \eta_k^2 < \infty, \qquad \sum_{k=1}^{\infty} \eta_k = \infty \,,$$

subgradient method satisfies

$$\lim_{k\to\infty} f(\boldsymbol{x}_{\text{best}}^{(k)}) \leqslant f^* \,.$$

Note: $f(\boldsymbol{x}_{\text{best}}^{(k)}) = \min_{i=0,\dots,k} f(\boldsymbol{x}_i), \quad f^* = f(\boldsymbol{x}^*)$

Subgradient Method

Can prove both results from the same basic inequality. Key steps:

- Using the definition of subgradient

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \leqslant \|\boldsymbol{x}_{k-1} - \boldsymbol{x}^*\|^2 - 2\eta_k(f(\boldsymbol{x}_{k-1}) - f^*) + \eta_k^2\|g_{k-1}\|^2$$

- Iterating last inequality

$$\|\boldsymbol{x}_k - \boldsymbol{x}^*\|^2 \leqslant \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|^2 - 2\sum_{i=1}^{k}\eta_i(f(\boldsymbol{x}_{i-1}) - f^*) + \sum_{i=1}^{k}\eta_i^2\|g_{i-1}\|^2$$

Subgradient Method

- Using $\|\boldsymbol{x}_k - \boldsymbol{x}^*\| \geqslant 0$ and letting $R = \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$,

$$0 \leqslant R^2 - 2\sum_{i=1}^{k} \eta_i (f(\boldsymbol{x}_{i-1}) - f^*) + L^2 \sum_{i=1}^{k} \eta_i^2$$

- Introducing $f(\boldsymbol{x}_{\text{best}}^{(k)}) = \min_{i=0,\dots,k} f(\boldsymbol{x}_i)$, and rearranging, we have the basic inequality

$$f(\boldsymbol{x}_{\text{best}}^{(k)}) - f^* \leqslant \frac{R^2 + L^2 \sum_{i=1}^{k} \eta_i^2}{2\sum_{i=1}^{k} \eta_i}$$

For different step size choices, convergence results can be directly obtained from this bound.

With fixed step size η,

$$f(\boldsymbol{x}_{\text{best}}^{(k)}) - f^* \leqslant \frac{R^2}{2k\eta} + \frac{L^2\eta}{2}\,.$$

To make $f(\boldsymbol{x}_{\text{best}}^{(k)}) - f^* \leqslant \varepsilon$, let's make each term $\leqslant \varepsilon/2$, by choosing

$$\eta = \frac{\varepsilon}{L^2} \qquad \text{and} \qquad k = \frac{R^2 L^2}{\varepsilon^2}\,.$$

Thus, the subgradient method has convergence rate $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$
...compare this to $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ rate of gradient descent

Questions?

Summary

f	Algorithm	Convergence	# Iterations
Convex L-Lipschitz	GD	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{RL}{\sqrt{T}}$	$\frac{R^2L^2}{\varepsilon^2}$
Convex L-Smooth	GD	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{R^2L}{2T}$	$\frac{R^2L}{2\varepsilon}$
Convex L-Lipschitz	Subgrad	$f(\boldsymbol{x}_{\text{best}}^{(T)}) - f(\boldsymbol{x}^*) \leqslant \frac{LR}{\sqrt{T}}$	$\frac{R^2L^2}{\varepsilon^2}$

- Time horizon $T > 0$ is given
- $R := \|\boldsymbol{x}_0 - \boldsymbol{x}^*\|$
- $\boldsymbol{x}_{\text{best}}^{(T)} := \arg\min_{i=0,1,\ldots,T} f(\boldsymbol{x}_i)$.

Thus, the subgradient method has convergence rate $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$
...compare this to $\mathcal{O}\left(\frac{1}{\varepsilon}\right)$ rate of gradient descent

References

Stephen P Boyd, *Lecture notes for ee 264b,stanford university (2010-2011)*.

Sébastien Bubeck, *Convex optimization: Algorithms and complexity*, Foundations and Trends in Machine Learning **8** (2015), no. 3-4, 231–357.

Stephen P Boyd and Lieven Vandenberghe, *Convex optimization*, Cambridge University Press, 2004.