DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 5 | Part 1

Stochastic Gradient Descent

Recall: (Sub)gradient descent

- **Goal:** minimize function $f(\vec{x})$.
- Iterative procedure that takes small steps in direction of steepest descent.

Recall: Gradient Descent

- Pick arbitrary starting point $\vec{x}^{(0)}$, **learning rate** parameter $\eta > 0$.

- Until convergence, repeat:
  - Compute gradient of f at $\vec{x}^{(i)}$; that is, compute $\vec{\nabla} f(\vec{x}^{(i)})$.
  - Update $\vec{x}^{(i+1)} = \vec{x}^{(i)} - \eta \vec{\nabla} f(\vec{x}^{(i)})$.

- When do we stop?
  - When difference between $\vec{x}^{(i)}$ and $\vec{x}^{(i+1)}$ is negligible.
  - I.e., when $\|\vec{x}^{(i)} - \vec{x}^{(i+1)}\|$ is small.

```python
def gradient_descent(
        gradient, x, learning_rate=.01,
        threshold=.1e-4
):
    while True:
        x_new = x - learning_rate * gradient(x)
        if np.linalg.norm(x - x_new) < threshold:
            break
        x = x_new
    return x
```

Gradient Descent for Minimizing Risk

- In ML, we often want to minimize a **risk function**:

$$R(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} L(H(\vec{x}^{(i)}; \vec{w}), y_i)$$

Observation

- The gradient of the risk function is a sum of gradients:

$$\vec{\nabla} R(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} \vec{\nabla} L(H(\vec{x}^{(i)}; \vec{w}), y_i)$$

- One term for each point in training data.

Problem

- In machine learning, the number of training points n can be **very large**.
- Computing the gradient can be **expensive** when n is large.
- Therefore, each step of gradient descent can be **expensive**.

Idea

- The (full) gradient of the risk uses all of the training data:

$$\nabla R(\vec{w}) = \frac{1}{n} \sum_{i=1}^{n} \nabla L(H(\vec{x}^{(i)}; \vec{w}), y_i)$$

- It is an average of n gradients.

- **Idea:** instead of using all n points, randomly choose $\ll n$.

Stochastic Gradient

- Choose a random subset (**mini-batch**) B of the training data.

- Compute a **stochastic gradient**:

$$\nabla R(\vec{w}) \approx \sum_{i \in B} \vec{\nabla} L(H(\vec{x}^{(i)}; \vec{w}), y_i)$$

Stochastic Gradient

$$\nabla R(\vec{w}) \approx \sum_{i \in B} \vec{\nabla} L(H(\vec{x}^{(i)}; \vec{w}), y_i)$$

- **Good:** if $|B| \ll n$, this is much faster to compute.
- **Bad:** it is a (random) approximation of the full gradient, noisy.

Stochastic Gradient Descent (SGD) for ERM

- Pick arbitrary starting point $\vec{x}^{(0)}$, **learning rate** parameter $\eta > 0$, batch size $m \ll n$.

- Until convergence, repeat:
  - Randomly sample a batch B of m training data points.
  - Compute stochastic gradient of f at $\vec{x}^{(i)}$:

$$\vec{g} = \sum_{i \in B} \vec{\nabla} L(H(\vec{x}^{(i)}; \vec{w}), y_i)$$

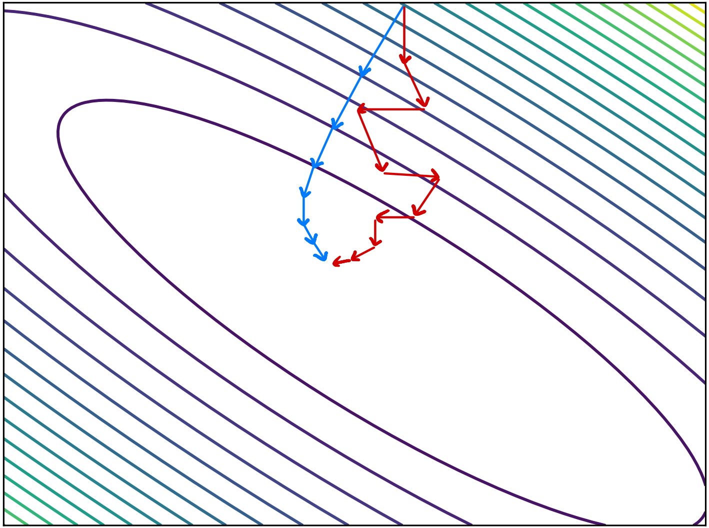
- Update $\vec{x}^{(i+1)} = \vec{x}^{(i)} - \eta \vec{g}$

Idea

- In practice, a stochastic gradient often works well enough.
- It is better to take many noisy steps quickly than few exact steps slowly.

Batch Size

- Batch size m is a parameter of the algorithm.
- The larger m, the more reliable the stochastic gradient, but the more time it takes to compute.
- Extreme case when $m = 1$ will still work.



Usefulness of SGD

- SGD allows learning on **massive** data sets.
- Useful even when exactly solutions available.
  - E.g., least squares regression / classification.

DSC 140A

Probabilistic Modeling & Machine Learning

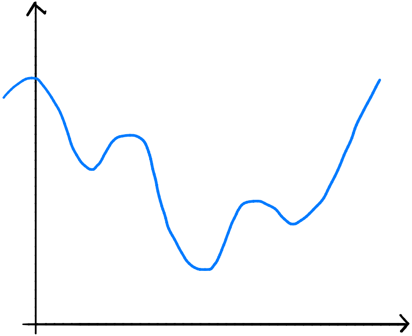
Lecture 5 | Part 2

Convexity

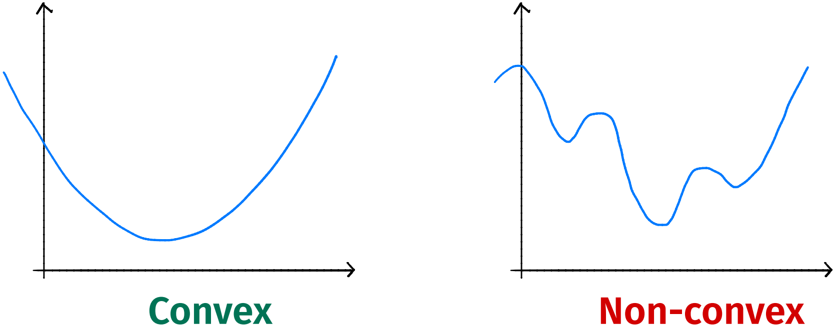
Question

- When is gradient descent guaranteed to work?

Not here...



Convex Functions

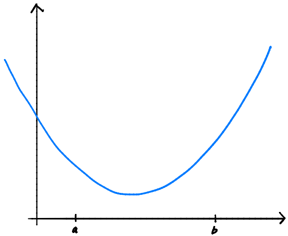


Convexity: Definition

- f is **convex** if for **every** a, b the line segment between

$$(a, f(a)) \qquad \text{and} \qquad (b, f(b))$$

does not go below the plot of f.

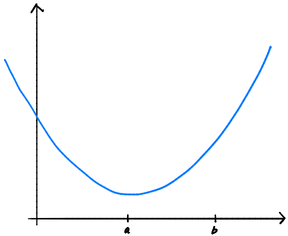


Convexity: Definition

- f is **convex** if for **every** a, b the line segment between

$$(a, f(a)) \qquad \text{and} \qquad (b, f(b))$$

does not go below the plot of f.

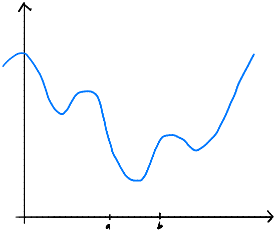


Convexity: Definition

- f is **convex** if for **every** a, b the line segment between

$$(a, f(a)) \qquad \text{and} \qquad (b, f(b))$$

does not go below the plot of f.

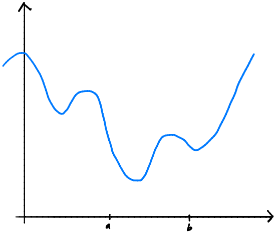


Convexity: Definition

- f is **convex** if for **every** a, b the line segment between

$$(a, f(a)) \qquad \text{and} \qquad (b, f(b))$$

does not go below the plot of f.



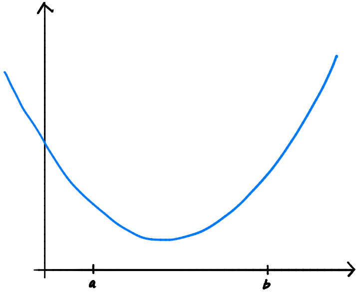
Other Terms

- If a function is not convex, it is **non-convex**.
- **Strictly convex**: the line lies strictly above curve.
- **Concave**: the line lines on or below curve.

Convexity: Formal Definition

- A function $f : \mathbb{R} \to \mathbb{R}$ is **convex** if for every choice of $a, b \in \mathbb{R}$ and $t \in [0, 1]$:

$$(1 - t)f(a) + tf(b) \geq f((1 - t)a + tb).$$

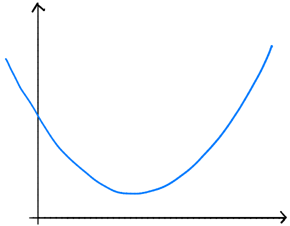


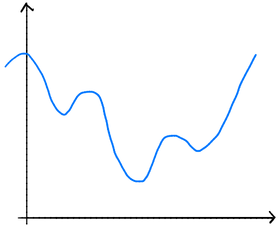
Example

Is $f(x) = |x|$ convex?

Another View: Second Derivatives

- If $\frac{d^2 f}{dx^2}(x) \geq 0$ for all x, then f is convex.
- Example: $f(x) = x^4$ is convex.
- **Warning!** Only works if f is twice differentiable!



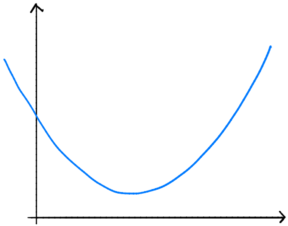


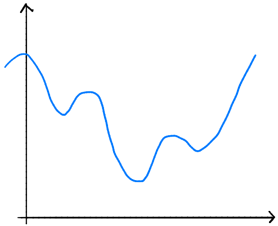
Another View: Second Derivatives

- "Best" straight line at x_0:
  - $h_1(z) = f'(x_0) \cdot z + b$
- "Best" parabola at x_0:
  - At x_0, f looks likes $h_2(z) = \frac{1}{2} f''(x_0) \cdot z^2 + f'(x_0) z + c$
  - Possibilities: upward-facing, downward-facing.

Convexity and Parabolas

- Convex if for **every** x_0, parabola is upward-facing.
  - That is, $f''(x_0) \geq 0$.





Proving Convexity Using Properties

Suppose that $f(x)$ and $g(x)$ are convex. Then:

- $w_1 f(x) + w_2 g(x)$ is convex, provided $w_1, w_2 \geq 0$
  - Example: $3x^2 + |x|$ is convex
- $g(f(x))$ is convex, provided g is non-decreasing.
  - Example: e^{x^2} is convex
- $\max\{f(x), g(x)\}$ is convex
  - Example: $\begin{cases} 0, & x < 0 \\ x, & x \geq 0 \end{cases}$ is convex

Convexity and Gradient Descent

- Convex functions are (relatively) easy to optimize.
- **Theorem**: if $f(x)$ is convex and "not too steep"[1] then (stochastic) (sub)gradient descent converges to a **global optimum** of f *provided* that the step size is small enough[2].

---

[1] Technically, c-Lipschitz

[2] step size related to steepness, should decrease like $1/\sqrt{t}$, where t is step number

Nonconvexity and Gradient Descent

- Nonconvex functions are (relatively) hard to optimize.
- Gradient descent can still be useful.
- But not guaranteed to converge to a global minimum.

DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 5 | Part 3

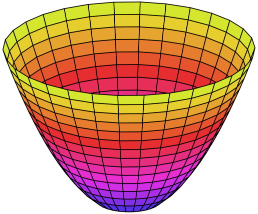
Convexity in Many Dimensions

Convexity: Definition

- $f(\vec{x})$ is **convex** if for **every** $\vec{a}, \vec{b}$ the line segment between

$$(\vec{a}, f(\vec{a})) \qquad \text{and} \qquad (\vec{b}, f(\vec{b}))$$

does not go below the plot of f.



Convexity: Formal Definition

- A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if for every choice of $\vec{a}, \vec{b} \in \mathbb{R}^d$ and $t \in [0, 1]$:

$$(1 - t)f(\vec{a}) + tf(\vec{b}) \geq f((1 - t)\vec{a} + t\vec{b}).$$

The Second Derivative Test

- For 1-d functions, convex if second derivative ≥ 0.
- For 2-d functions, convex if ???

Second Derivatives in 2-d

- In 2-d, there are 4 second derivatives of $f(\vec{x})$:
  - $\frac{\partial f^2}{\partial x_1^2}, \frac{\partial f^2}{\partial x_2^2}, \frac{\partial f^2}{\partial x_1 x_2}, \frac{\partial f^2}{\partial x_2 x_1}$

Convexity in 2-d

- "Best" quadratic function approximating f at $\vec{x}$:

$$
\begin{aligned}
h_2(z_1, z_2) &= az_1^2 + bz_2^2 + cz_1 z_2 + \ldots \\
&= \frac{1}{2}\frac{\partial f^2}{\partial x_1^2}(\vec{x}) \cdot z_1 + \frac{1}{2}\frac{\partial f^2}{\partial x_2^2}(\vec{x}) \cdot z_2 + \frac{\partial f^2}{\partial x_1 x_2}(\vec{x}) \cdot z_1 z_2 + \ldots
\end{aligned}
$$

- a, b, c determine rough shape. Possibilities:
  - Upward-facing bowl.
  - Downward-facing bowl.
  - "Saddle"

Convexity in 2-d

- Convex if at any $\vec{x}$, for any z_1, z_2:

$$\frac{1}{2}\frac{\partial f^2}{\partial x_1^2}(\vec{x}) \cdot z_1 + \frac{1}{2}\frac{\partial f^2}{\partial x_2^2}(\vec{x}) \cdot z_2 + \frac{\partial f^2}{\partial x_1 x_2}(\vec{x}) \cdot z_1 z_2 \geq 0$$

The Hessian Matrix

- Create the **Hessian** matrix of second derivatives:

$$H(\vec{x}) = \begin{pmatrix} \frac{\partial f^2}{\partial x_1^2}(\vec{x}) & \frac{\partial f^2}{\partial x_1 x_2}(\vec{x}) \\ \frac{\partial f^2}{\partial x_2 x_1}(\vec{x}) & \frac{\partial f^2}{\partial x_2^2}(\vec{x}) \end{pmatrix}$$

In General

- If $f : \mathbb{R}^d \to \mathbb{R}$, the **Hessian** at $\vec{x}$ is:

$$H(\vec{x}) = \begin{pmatrix} \frac{\partial f^2}{\partial x_1^2}(\vec{x}) & \frac{\partial f^2}{\partial x_1 x_2}(\vec{x}) & \cdots & \frac{\partial f^2}{\partial x_1 x_d}(\vec{x}) \\ \frac{\partial f^2}{\partial x_2 x_1}(\vec{x}) & \frac{\partial f^2}{\partial x_2^2}(\vec{x}) & \cdots & \frac{\partial f^2}{\partial x_2 x_d}(\vec{x}) \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f^2}{\partial x_d x_1}(\vec{x}) & \frac{\partial f^2}{\partial x_d^2}(\vec{x}) & \cdots & \frac{\partial f^2}{\partial x_d^2}(\vec{x}) \end{pmatrix}$$

Observations

- H is square.
- H is symmetric.

Convexity in 2-d

- Convex if at any $\vec{x}$, for any z_1, z_2:

$$\frac{1}{2}\frac{\partial f^2}{\partial x_1^2}(\vec{x}) \cdot z_1 + \frac{1}{2}\frac{\partial f^2}{\partial x_2^2}(\vec{x}) \cdot z_2 + \frac{\partial f^2}{\partial x_1 x_2}(\vec{x}) \cdot z_1 z_2 \geq 0$$

- Equivalently, convex if for any $\vec{x}$ and any $\vec{z}$:

$$\vec{z}^T H(\vec{x})\, \vec{z} \geq 0$$

Positive Semi-Definite

- A square, $d \times d$ symmetric matrix X is **positive semi-definite** (PSD) if for any $\vec{u}$:

$$\vec{u}^T X \vec{u} \geq 0$$

The Second Derivative Test

- A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if for any $\vec{x} \in \mathbb{R}^d$, the Hessian matrix $H(\vec{x})$ is positive semi-definite.

But wait...

- How can we tell if a matrix is positive semi-definite?

Example

$$M = \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$$

Example

$$M = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$$

Example

Is $f(x, y) = x^2 + 4xy + y^2$ convex?

Sums of Convex Functions

- Suppose that $f(\vec{x})$ and $g(\vec{x})$ are convex. Then $w_1 f(\vec{x}) + w_2 g(\vec{x})$ is convex, provided $w_1, w_2 \geq 0$.

Affine Composition

- Suppose that $f(x)$ is convex. Let A be a matrix, and $\vec{x}$ and $\vec{b}$ be vectors. Then

  $$g(\vec{x}) = f(A\vec{x} + \vec{b})$$

  is convex as a function of $\vec{x}$.

- Useful!

DSC 140A

Probabilistic Modeling & Machine Learning

Lecture 5 | Part 4

Convex Loss Functions

Convexity and Gradient Descent

- Convex functions are (relatively) easy to optimize.
- **Theorem**: if $f(x)$ is convex and "not too steep"[3] then (stochastic) (sub)gradient descent converges to a **global optimum** of f *provided* that the step size is small enough[4].

---

[3] Technically, c-Lipschitz

[4] step size related to steepness, should decrease like $1/\sqrt{t}$, where t is step number

Convex Loss

- **Recall:** sums of convex functions are convex.
- **Implication:** if loss function is convex as a function of $\vec{w}$, so is the risk.
- Convex losses are **nice**.

Example

- Recall the square loss:

$$L(H(\vec{x}, \vec{w}), y) = (\vec{x} \cdot \vec{w} - y)^2$$

- Is this convex as a function of $\vec{w}$?

Mean Squared Error

- The square loss is a convex function of $\vec{w}$.
- We had an explicit solution for the best $\vec{w}$:

$$\vec{w} = (X^T X)^{-1} X^T \vec{y}$$

- But we could also have used gradient descent.

Perceptron Loss

- The perceptron loss is:

$$L_{\text{tron}}(H(\vec{x};\vec{w}), y) = \begin{cases} 0, & \operatorname{sign}(\vec{w}\cdot\vec{x}) = \operatorname{sign}(y) \\ |\vec{w}\cdot\vec{x}|, & \operatorname{sign}(\vec{w}\cdot\vec{x}) \neq \operatorname{sign}(y) \end{cases}$$

- Is it convex as a function of $\vec{w}$?

Summary

- We learned what it means for a function to be **convex**.
- Convex functions are (relatively) **easy** to optimize with gradient descent.
- We like **convex loss functions**, like the square loss.