Spring 26 IntroToML Lab 2

Convexity, Gradient Descent, and Logistic Regression

Patrick Shen

NYU Center for Data Science

January 29, 2026

Agenda

- Review of GD and SGD
- GD on logistic regression
- convexity theory
- Notebook demo

Review of Gradient Descent

Consider empirical risk

$$f(x) = \frac{1}{n}\sum_{i=1}^{n} f_i(x), \qquad x \in \mathbb{R}^d.$$

GD update:

$$\theta_k = \theta_{k-1} - \eta_k \nabla_{\theta_{k-1}} f(x).$$

If $\theta_k = \theta_{k-1}$, then θ_{k-1} is a fixed point of the update, and (for convex f) a global minimizer.

Stochastic Gradient Descent

SGD update:

$$\theta_k = \theta_{k-1} - \eta_k \nabla f_{\theta_{k-1}}(x_{i_k}),$$

where i_k is sampled uniformly from $\{1, 2, \ldots, n\}$.

Unbiasedness:

$$\mathbb{E}\left[\nabla f_{\theta_{k-1}}(x_{i_k})\right] = \nabla_{\theta_{k-1}} f(x).$$

Logistic Regression: model

Outputs are in $[0,1]$.

Let $X \in \mathbb{R}^{m\times d}$, $y \in \{0,1\}^m$, parameters $w \in \mathbb{R}^d$, $b \in \mathbb{R}$.

$$z = Xw + b\mathbf{1}, \qquad \hat{y} = \sigma(z), \qquad \sigma(z) = \frac{1}{1+e^{-z}}.$$

Logistic loss

$$J(w, b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)\right].$$

$$J(w,b) = -\frac{1}{m}\sum_{i=1}^{m}\left[y_i \log(\hat{y}_i) + (1-y_i)\log(1-\hat{y}_i)\right].$$

Exercise.

- Write $\frac{\partial J}{\partial z}$, $\frac{\partial J}{\partial w}$, $\frac{\partial J}{\partial b}$.
- Write one step of GD for (w, b).

Gradients and one-step GD (standard form)

For logistic regression with $\hat{y} = \sigma(z)$ and $z = Xw + b\mathbf{1}$:

$$\frac{\partial J}{\partial z} = \frac{1}{m}(\hat{y} - y), \qquad \nabla_w J(w, b) = \frac{1}{m}X^\top(\hat{y} - y), \qquad \frac{\partial J}{\partial b} = \frac{1}{m}\mathbf{1}^\top(\hat{y} - y).$$

Gradients and one-step GD (standard form)

For logistic regression with $\hat{y} = \sigma(z)$ and $z = Xw + b\mathbf{1}$:

$$\frac{\partial J}{\partial z} = \frac{1}{m}(\hat{y} - y), \qquad \nabla_w J(w, b) = \frac{1}{m} X^\top (\hat{y} - y), \qquad \frac{\partial J}{\partial b} = \frac{1}{m}\mathbf{1}^\top (\hat{y} - y).$$

One GD step:

$$w_k = w_{k-1} - \eta_k \nabla_w J(w_{k-1}, b_{k-1}), \qquad b_k = b_{k-1} - \eta_k \frac{\partial J}{\partial b}(w_{k-1}, b_{k-1}).$$

Notebook exercise

Optimization: the key question

The key question in optimization is to find the smallest (or largest) value of a function.

$$\min f(x) \quad \text{s.t.} \quad x \in \Omega \subseteq \mathbb{R}^d,$$

where:

- Ω is the **domain / feasible set**
- $f(x)$ is the **objective function**

Convex set

Definition (Convex set). A set Ω is convex if for all $x, y \in \Omega$ and all $\lambda \in [0, 1]$,

$$\lambda x + (1 - \lambda) y \in \Omega.$$

Convex set

Definition (Convex set). A set Ω is convex if for all $x, y \in \Omega$ and all $\lambda \in [0, 1]$,

$$\lambda x + (1 - \lambda) y \in \Omega.$$

Exercise. Let

$$C = \left\{ x \in \mathbb{R}^n : \|x\|_2 \leq 1 \right\}.$$

Prove that C is a convex set.

Characterization of convexity

Theorem. The following statements are equivalent:

1. f is convex.
2. For any $\lambda \in [0,1]$ and any x, y in the domain of f,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y).$$

First-order characterization of convexity

Assume $f : \Omega \to \mathbb{R}$ is **differentiable** and the domain Ω is convex.

Theorem (First-order condition). The following are equivalent:

1. f is convex.
2. For all $x, y \in \Omega$,
$$f(y) \geq f(x) + \langle \nabla f(x), \, y - x \rangle.$$

First-order characterization of convexity

Assume $f : \Omega \to \mathbb{R}$ is **differentiable** and the domain Ω is convex.

Theorem (First-order condition). The following are equivalent:

1. f is convex.
2. For all $x, y \in \Omega$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle.$$

Geometric meaning: the tangent hyperplane at x is a global underestimator of f.

Second-order characterization of convexity

Assume $f : \Omega \to \mathbb{R}$ is **twice differentiable** and the domain Ω is convex.

Theorem (Second-order condition). The following are equivalent:

1. f is convex.
2. For all $x \in \Omega$, the Hessian is positive semidefinite:

$$\nabla^2 f(x) \succeq 0 \quad \Longleftrightarrow \quad v^\top \nabla^2 f(x)\, v \geq 0 \;\; \forall v \in \mathbb{R}^d.$$

Special case: if $\nabla^2 f(x) \succ 0$ for all x, then f is **strictly convex**.

Useful lemmas: why GD can work (intuition)

Lemma 1. For a convex function $f(x)$, all local minima (if they exist) are global minima.

- Proof idea: contradiction.

Lemma 2. Let $\{g_i\}_{i=1}^n$ be convex functions. Then

$$g(x) := \frac{1}{n}\sum_{i=1}^{n} g_i(x)$$

is convex.

First-order optimality for differentiable convex f

Theorem. If f is differentiable and convex, then the following are equivalent:

1. $x^\star$ is a (local/global) minimizer of f
2. $\nabla f(x^\star) = 0$

Exercise: quadratic objective

Let

$$f(w) = \frac{1}{2}\|w\|_2^2, \qquad w \in \mathbb{R}^d.$$

1. Compute $\nabla f(w)$.
2. Write the gradient update $w_k = w_{k-1} - \eta \nabla f(w_k)$.

Exercise: convexity of logistic loss

Exercise. Show that the logistic loss function $J(w, b)$ is convex.

Hint. Compute the Hessian and show it is positive semidefinite (PSD).

Wrap-up

- Convex sets/functions and key properties
- GD and SGD updates for empirical risk minimization
- Logistic regression: model, loss, gradients, and GD steps