Unconstrained optimization I

Topics we'll cover

1. Optimization by local search
2. The problem of multiple local optima
3. Gradient descent
4. Taking the derivative of a function of many variables

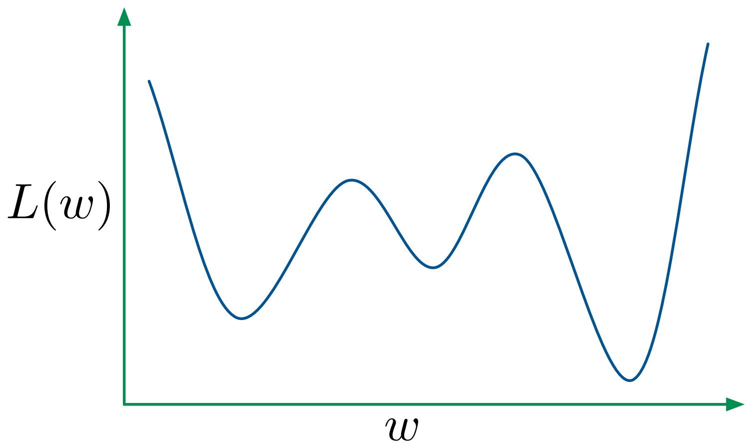
Minimizing a loss function

Usual setup in machine learning: choose a model w by minimizing a loss function $L(w)$ that depends on the data.

- Linear regression: $L(w) = \sum_i (y^{(i)} - (w \cdot x^{(i)}))^2$
- Logistic regression: $L(w) = \sum_i \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})})$

Default way to solve this minimization: **local search**.

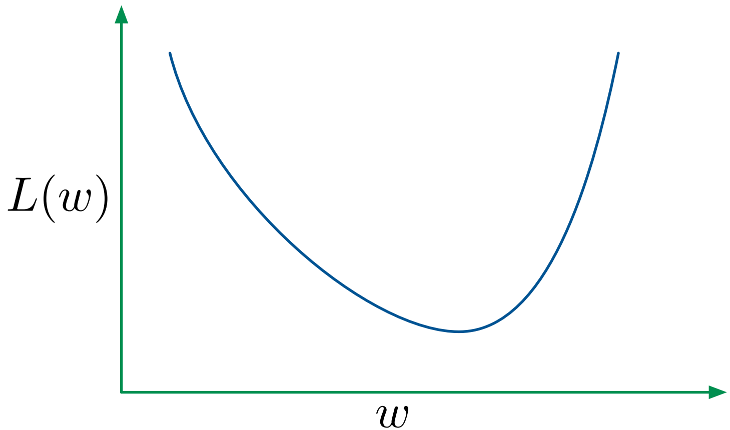
Local search



- Initialize w arbitrarily
- Repeat until w converges:
  - Find some w' close to w with $L(w') < L(w)$.
  - Move w to w'.

A good situation for local search

When the loss function is **convex**:



Idea for picking search direction:
Look at the **derivative** of $L(w)$ at the current point w.

Gradient descent

For minimizing a function $L(w)$:

- $w_o = 0, t = 0$
- while $\nabla L(w_t) \not\approx 0$:
  - $w_{t+1} = w_t - \eta_t \nabla L(w_t)$
  - $t = t + 1$

Here η_t is the *step size* at time t.

Multivariate differentiation

Example: $w \in \mathbb{R}^3$ and $F(w) = 3w_1 w_2 + w_3$.

Example: $w \in \mathbb{R}^d$ and $F(w) = w \cdot x$.

Example: $w \in \mathbb{R}^d$ and $F(w) = \|w\|^2$.

Gradient descent

For minimizing a function $L(w)$:

- $w_o = 0, t = 0$
- while $\nabla L(w_t) \not\approx 0$:
  - $w_{t+1} = w_t - \eta_t \nabla L(w_t)$
  - $t = t + 1$

Here η_t is the *step size* at time t.

Unconstrained optimization II

Topics we'll cover

1. Why does gradient descent work?
2. Setting the step size
3. Gradient descent for logistic regression

Gradient descent

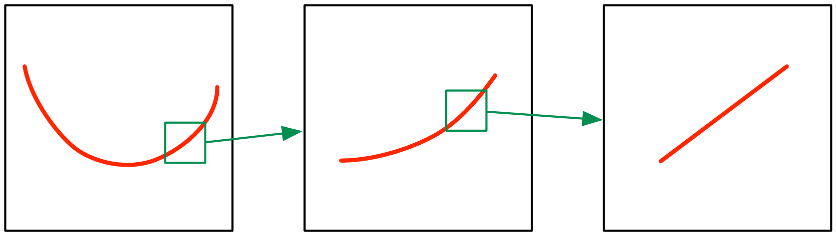
For minimizing a function $L(w)$, over $w \in \mathbb{R}^d$:

- $w_o = 0, t = 0$
- while $\nabla L(w_t) \not\approx 0$:
  - $w_{t+1} = w_t - \eta_t \nabla L(w_t)$
  - $t = t + 1$

Here η_t is the *step size* at time t.

Gradient descent: rationale

"Differentiable" $\Longrightarrow$ "locally linear".



For *small* displacements $u \in \mathbb{R}^d$,

$$L(w+u) \;\approx\; L(w) + u \cdot \nabla L(w) \qquad\qquad .$$

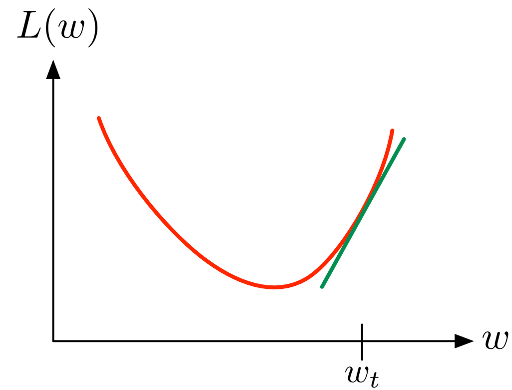
Therefore, if $u = -\eta \nabla L(w)$ is small,

$$L(w+u) \;\approx\; L(w) - \eta \|\nabla L(w)\|^2 \;<\; L(w)$$

The step size matters

Update rule: $w_{t+1} = w_t - \eta_t \nabla L(w_t)$

- Step size η_t too small: not much progress
- Too large: overshoot the mark



Some choices:

- Set η_t according to a fixed schedule, like $1/t$
- Choose by line search to minimize $L(w_{t+1})$

Example: logistic regression

For $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)}) \in \mathbb{R}^d \times \{-1, +1\}$, loss function

$$L(w) \;=\; \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})})$$

What is the derivative?

Gradient descent for logistic regression

- Set $w_0 = 0$
- For $t = 0, 1, 2, \ldots$, until convergence:

$$w_{t+1} \; = \; w_t \; + \; \eta_t \sum_{i=1}^{n} y^{(i)} x^{(i)} \Pr_{w_t}(-y^{(i)} | x^{(i)})$$

Unconstrained optimization III

Topics we'll cover

1. Stochastic gradient descent for logistic regression
2. Stochastic gradient descent more generally

Recall: gradient descent for logistic regression

- Set $w_0 = 0$
- For $t = 0, 1, 2, \ldots$, until convergence:

$$w_{t+1} = w_t + \eta_t \sum_{i=1}^{n} y^{(i)} x^{(i)} \Pr_{w_t}(-y^{(i)} | x^{(i)})$$

Each update involves the entire data set, which is inconvenient.

Stochastic gradient descent: update based on just one point:

- Get next data point (x, y) by cycling through data set
- $w_{t+1} = w_t + \eta_t \, y \, x \Pr_{w_t}(-y|x)$

Decomposable loss functions

Loss function for logistic regression:

$$L(w) = \sum_{i=1}^{n} \ln(1 + e^{-y^{(i)}(w \cdot x^{(i)})}) = \sum_{i=1}^{n} (\text{loss of } w \text{ on } (x^{(i)}, y^{(i)}))$$

Most ML loss functions are like this: for training set $(x^{(1)}, y^{(1)}), \ldots, (x^{(n)}, y^{(n)})$,

$$L(w) = \sum_{i=1}^{n} \ell(w; x^{(i)}, y^{(i)})$$

where $\ell(w; x, y)$ captures the loss on a single point.

Gradient descent and stochastic gradient descent

For minimizing

$$L(w) = \sum_{i=1}^{n} \ell(w; x^{(i)}, y^{(i)})$$

Gradient descent:

- $w_o = 0$
- while not converged:
  - $w_{t+1} = w_t - \eta_t \sum_{i=1}^{n} \nabla \ell(w_t; x^{(i)}, y^{(i)})$

Stochastic gradient descent:

- $w_o = 0$
- Keep cycling through data points (x, y):
  - $w_{t+1} = w_t - \eta_t \nabla \ell(w_t; x, y)$

Variant: mini-batch stochastic gradient descent

Stochastic gradient descent:

- $w_o = 0$
- Keep cycling through data points (x, y):
  - $w_{t+1} = w_t - \eta_t \nabla \ell(w_t; x, y)$

Mini-batch stochastic gradient descent:

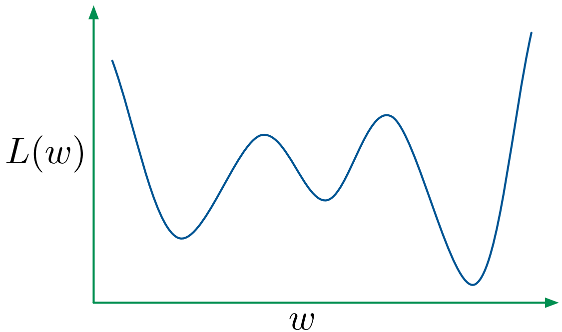
- $w_o = 0$
- Repeat:
  - Get the next batch of points B
  - $w_{t+1} = w_t - \eta_t \sum_{(x,y)\in B} \nabla \ell(w_t; x, y)$

Convexity I

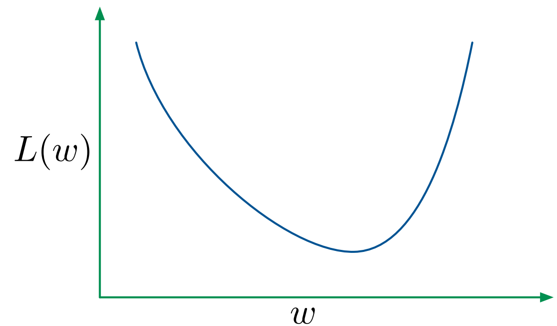
Topics we'll cover

1. Definition of convexity
2. The second-derivative test for convexity

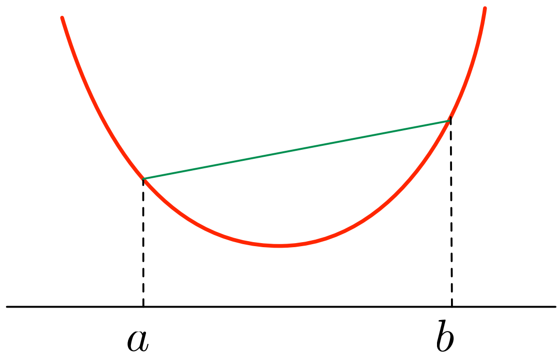
Is our loss function convex?



versus



Convexity



A function $f : \mathbb{R}^d \to \mathbb{R}$ is **convex** if for all $a, b \in \mathbb{R}^d$ and $0 < \theta < 1$,

$$f(\theta a + (1-\theta)b) \;\leq\; \theta f(a) + (1-\theta) f(b).$$

It is **strictly convex** if strict inequality holds for all $a \neq b$.

f is **concave** $\Leftrightarrow$ $-f$ is convex

Checking convexity for functions of one variable

A function $f : \mathbb{R} \to \mathbb{R}$ is convex if its second derivative is ≥ 0 everywhere.

Example: $f(z) = z^2$

Checking convexity

Function of one variable

$$F : \mathbb{R} \to \mathbb{R}$$

- Value: number
- Derivative: number
- Second derivative: number

Convex if second derivative is always ≥ 0

Function of d variables

$$F : \mathbb{R}^d \to \mathbb{R}$$

- Value: number
- Derivative: d-dimensional vector
- Second derivative: $d \times d$ matrix

Convex if second derivative matrix is always positive semidefinite

First and second derivatives of multivariate functions

For a function $f : \mathbb{R}^d \to \mathbb{R}$,

- the first derivative is a vector with d entries:

$$\nabla f(z) = \begin{pmatrix} \frac{\partial f}{\partial z_1} \\ \vdots \\ \frac{\partial f}{\partial z_d} \end{pmatrix}$$

- the second derivative is a $d \times d$ matrix, the **Hessian** $H(z)$:

$$H_{jk} = \frac{\partial^2 f}{\partial z_j \partial z_k}$$

Example

Find the second derivative matrix of $f(z) = \|z\|^2$.

Checking convexity

Function of one variable

$$F : \mathbb{R} \to \mathbb{R}$$

- Value: number
- Derivative: number
- Second derivative: number

Convex if second derivative is always ≥ 0

Function of d variables

$$F : \mathbb{R}^d \to \mathbb{R}$$

- Value: number
- Derivative: d-dimensional vector
- Second derivative: $d \times d$ matrix

Convex if second derivative matrix is always positive semidefinite

Linear algebra IV
Positive semidefinite matrices

Topics we'll cover

1. Positive semidefinite matrices
2. Properties of PSD matrices
3. Checking if a matrix is PSD
4. A hierarchy of square matrices

When is a square matrix “positive”?

- A superficial notion: when all its entries are positive
- A deeper notion: **when the quadratic function defined by it is always positive**

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

Positive semidefinite matrices

Recall: every **square** matrix M encodes a **quadratic function**:

$$x \mapsto x^T M x = \sum_{i,j=1}^{d} M_{ij} x_i x_j$$

(M is a $d \times d$ matrix and x is a vector in $\mathbb{R}^d$)

A symmetric matrix M is **positive semidefinite (psd)** if:

$$x^T M x \geq 0 \text{ for all vectors } x$$

A symmetric matrix M is **positive semidefinite (psd)** if:

$$x^T M x \geq 0 \;\text{ for all vectors } x$$

We saw that $\begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$ is PSD. What about $\begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix}$?

A symmetric matrix M is **positive semidefinite (psd)** if:

$$x^T Mx \geq 0 \text{ for all vectors } z$$

When is a diagonal matrix PSD?

A symmetric matrix M is **positive semidefinite (psd)** if:

$$x^T M x \geq 0 \quad \text{for all vectors } z$$

If M is PSD, must cM be PSD for a constant c?

A symmetric matrix M is **positive semidefinite (psd)** if:

$$x^T M x \geq 0 \quad \text{for all vectors } z$$

If M, N are of the same size and PSD, must $M + N$ be PSD?

Checking if a matrix is PSD

A matrix M is PSD if and only if it can be written as $M = UU^T$ for some matrix U.

Quick check: say $U \in \mathbb{R}^{r \times d}$ and $M = UU^T$.

1. M is square.
2. M is symmetric.
3. Pick any $x \in \mathbb{R}^r$. Then

$$
\begin{aligned}
x^T M x = x^T U U^T x &= (x^T U)(U^T x) \\
&= (U^T x)^T (U^T x) \\
&= \|U^T x\|^2 \geq 0.
\end{aligned}
$$

Another useful fact: any covariance matrix is PSD.

A hierarchy of square matrices

Square	$M \in \mathbb{R}^{d \times d}$
Symmetric	$M = M^T$
Positive semidefinite	$x^T M x \geq 0$ for all $x \in \mathbb{R}^d$

Convexity II

Topics we'll cover

1. Second derivative test for convexity
2. Convexity examples

Second-derivative test for convexity

A function of several variables, $F(z)$, is convex if its second-derivative matrix $H(z)$ is positive semidefinite for all z.

More formally:
Suppose that for $f : \mathbb{R}^d \to \mathbb{R}$, the second partial derivatives exist everywhere and are continuous functions of z. Then:

1. $H(z)$ is a symmetric matrix
2. f is convex $\Leftrightarrow$ $H(z)$ is positive semidefinite for all $z \in \mathbb{R}^d$

Example

Is $f(x) = \|x\|^2$ convex?

Example

Fix any vector $u \in \mathbb{R}^d$. Is this function $f : \mathbb{R}^d \to \mathbb{R}$ convex?

$$f(z) = (u \cdot z)^2$$

Least-squares regression

Recall loss function: for data points $(x^{(i)}, y^{(i)}) \in \mathbb{R}^d \times \mathbb{R}$,

$$L(w) = \sum_{i=1}^{n} (y^{(i)} - (w \cdot x^{(i)}))^2$$