Lecture 11: Optimization

DS GA 1002 Statistical and Mathematical Models
http://www.cims.nyu.edu/~cfgranda/pages/DSGA1002_fall15

Carlos Fernandez-Granda

11/30/2015
Motivation

In machine learning and statistics we propose models to explain the data. Optimization allows to fit our models by choosing the parameters that maximize/minimize a cost function such as:

- the likelihood of the data given the parameters
- the error achieved by the model on a training dataset

Understanding optimization algorithms used to achieve this is crucial.
Derivatives and convexity

Optimization algorithms in 1D

Multiple dimensions

Optimization algorithms
Derivatives of a 1D function

The derivative of \( f : \mathbb{R} \rightarrow \mathbb{R} \) at \( x \) is

\[
f'(x) := \lim_{h \to 0} \frac{f(x + h) - f(x)}{h}
\]

Higher order derivatives are defined recursively,

\[
f''(x) := (f'(x))'
\]
\[
f'''(x) := (f''(x))'
\]
The first-order or linear approximation of $f$ at $x$ is

$$f_x^1 (y) := f (x) + f' (x) (y - x)$$

Lemma:

$$\lim_{y \to x} \frac{f (y) - f_x^1 (y)}{y - x} = 0$$
Local monotonicity

The first derivative encodes the local monotonicity of the function

- If $f'(x) > 0$, $f$ is increasing at $x$
- If $f'(x) < 0$, $f$ is decreasing at $x$
- If $f'(x) = 0$, $f$ reaches a local minimum or a local maximum at $x$
Convexity

A function is convex if for any \( x, y \in \mathbb{R} \) and any \( \theta \in (0, 1) \),

\[
\theta f(x) + (1 - \theta) f(y) \geq f(\theta x + (1 - \theta) y)
\]

Equivalently, \( f \) is convex if and only if for every \( x \in \mathbb{R} \)

\[
f(y) \geq f(x) + f'(x)(y - x)
\]

If \( f'(x) = 0 \), \( x \) is a global minimum
Proof

Rearranging the terms we have

\[ f(y) \geq \frac{f(x + \theta(y - x)) - f(x)}{\theta} + f(x) \]

Set \( h = \theta(y - x) \) and take limit \( h \to 0 \)
Proof

Let $z = \theta x + (1 - \theta) y$, then

\[
\begin{align*}
  f(x) &\geq f'(z)(x - z) + f(z) \\
    &= f'(z)(1 - \theta)(x - y) + f(z) \\
  f(y) &\geq f'(z)(y - z) + f(z) \\
    &= f'(z)\theta(y - x) + f(z)
\end{align*}
\]
Strict convexity

A function is strictly convex if for any $x, y \in \mathbb{R}$ and any $\theta \in (0, 1)$,

$$\theta f(x) + (1 - \theta) f(y) > f(\theta x + (1 - \theta) y)$$

Equivalently, $f$ is strictly convex if and only if for every $x \in \mathbb{R}$

$$f(y) > f(x) + f'(x)(y - x)$$

If $f'(x) = 0$, $x$ is the only global minimum
Second derivative

\[ f \text{ is convex if and only if } f''(x) \geq 0 \text{ for all } x \in \mathbb{R} \]

\[ f \text{ is strictly convex if and only if } f''(x) > 0 \text{ for all } x \in \mathbb{R} \]
Proof

If $f$ is convex and $y > x$,

\[
f(x) \geq f'(y)(x - y) + f(y),
\]

\[
f(y) \geq f'(x)(y - x) + f(x).
\]

Rearranging

\[
f'(y)(y - x) \geq f(y) - f(x) \geq f'(x)(y - x).
\]

Since $y - x > 0$, $f'(y) \geq f'(x)$. 

Proof

For \( y > x \) and \( 0 < \theta < 1 \), let \( z = \theta y + (1 - \theta) x \)

Since \( y > z > x \), there exist \( \gamma_1 \in [x, z] \) and \( \gamma_2 \in [z, y] \) such that

\[
\begin{align*}
f'(\gamma_1) &= \frac{f(z) - f(x)}{z - x}, \\
f'(\gamma_2) &= \frac{f(y) - f(z)}{y - z}
\end{align*}
\]

Since \( \gamma_1 < \gamma_2 \), if \( f' \) is nondecreasing

\[
\frac{f(y) - f(z)}{y - z} \geq \frac{f(z) - f(x)}{z - x},
\]

which implies

\[
\frac{z - x}{y - x} f(y) + \frac{y - z}{y - x} f(x) \geq f(z)
\]
Quadratic approximation

The second-order or quadratic approximation of $f$ at $x$ is

$$f_x^2 (y) := f(x) + f'(x)(y - x) + \frac{1}{2}f''(x)(y - x)^2$$

Lemma:

$$\lim_{x \to y} \frac{f(y) - f_x^2 (y)}{(y - x)^2} = 0$$
Derivatives and convexity

Optimization algorithms in 1D

Multiple dimensions

Optimization algorithms
**Derivative descent**

**Intuition:** If $f'(x) > 0$ decrease $x$, if $f'(x) > 0$ increase $x$

1. Choose a random initialization $x \in \mathbb{R}$

2. For $i = 1, 2, \ldots$ compute

   $$x_i = x_{i-1} - \alpha f'(x_{i-1})$$

   until $|f'(x_{i-1})| \leq \epsilon$

The step size $\alpha$ is a fixed constant
Derivative descent, $\alpha = 7$
Derivative descent, $\alpha = 2$
Quadratic function

Derivative descent finds the global minimum of a convex quadratic

\[ q(x) := \frac{a}{2}x^2 + bx + c \]

in one step if we set

\[ \alpha = \frac{1}{f''(x)} = \frac{1}{a} \]
Newton’s method

**Intuition:** Iteratively minimize the local quadratic approximation of $f$

1. Choose a random initialization $x \in \mathbb{R}$

2. For $i = 1, 2, \ldots$ compute

   $$x_i = x_{i-1} - \frac{f'(x_{i-1})}{f''(x_{i-1})}$$

   until $|f'(x_{i-1})| \leq \epsilon$
Newton’s method
Derivative descent, $\alpha = 0.2$
Derivative descent, $\alpha = 0.05$
Newton’s method
Newton’s method
Derivative descent
Newton’s method
Derivatives and convexity

Optimization algorithms in 1D

Multiple dimensions

Optimization algorithms
Directional and partial derivatives

The directional derivative of a function \( f : \mathbb{R}^n \to \mathbb{R} \) is

\[
f'_u(x) := \lim_{h \to 0} \frac{f(x + hu) - f(x)}{h}.
\]

where \( ||u||_2 = 1 \)

Higher-order dir. der. in the same direction are computed recursively

\[
f''_u(x) := \lim_{h \to 0} \frac{f'(x + hu) - f'(x)}{h}.
\]
Partial derivatives

The partial derivative of a function \( f : \mathbb{R}^n \rightarrow \mathbb{R} \) is

\[
\frac{\partial f (x)}{\partial x_i} := \lim_{h \to 0} \frac{f \left( \begin{bmatrix} x_1 \\
x_2 \\
\vdots \\
x_i + h \\
\vdots \\
x_n \end{bmatrix} \right) - f \left( \begin{bmatrix} x_1 \\
x_2 \\
\vdots \\
x_i \\
\vdots \\
x_n \end{bmatrix} \right)}{h} = f'_{e_i} (x)
\]

Higher-order directional derivatives are computed recursively

\[
\frac{\partial^2 f (x)}{\partial x_j \partial x_i} = \frac{\partial}{\partial x_j} \frac{\partial f (x)}{\partial x_i} \quad \quad \frac{\partial^2 f (x)}{\partial x_i^2} = \frac{\partial}{\partial x_i} \frac{\partial f (x)}{\partial x_i}
\]
Gradient

The gradient of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is

$$\nabla f (x) = \begin{bmatrix}
\frac{\partial f(x)}{\partial x_1} \\
\frac{\partial f(x)}{\partial x_2} \\
\vdots \\
\frac{\partial f(x)}{\partial x_n}
\end{bmatrix}$$

The gradient encodes variation in every direction

$$f_u' (x) = \nabla f (x)^T u$$

where $\|u\|_2 = 1$

The gradient is the direction of maximum variation
Gradient
Linear approximation

The first-order or linear approximation of $f$ at $x$ is

$$f_x^1(y) := f(x) + \nabla f(x)(y - x)$$

Lemma:

$$\lim_{y \to x} \frac{f(y) - f_x^1(y)}{\|y - x\|_2} = 0$$
Convexity

A function $f : \mathbb{R}^n \to \mathbb{R}$ is convex if for any $x, y \in \mathbb{R}^n$ and any $\theta \in (0, 1)$,

$$\theta f(x) + (1 - \theta) f(y) \geq f(\theta x + (1 - \theta) y)$$

Equivalently, $f$ is convex if and only if it is convex along any line, i.e. if and only if $g : \mathbb{R} \to \mathbb{R}$

$$g(\alpha) := f(y + \alpha u)$$

is convex for any $y \in \mathbb{R}^n$ and $\|u\|_2 = 1$

Equivalently, $f$ is convex if and only if for every $x, y \in \mathbb{R}^n$

$$f(y) \geq f(x) + \nabla f(x)(y - x)$$

If $\nabla f(x) = 0$, $x$ is a global minimum
Strict convexity

A function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is strictly convex if for any $x, y \in \mathbb{R}^n$ and any $\theta \in (0, 1)$,

$$\theta f(x) + (1 - \theta) f(y) > f(\theta x + (1 - \theta) y)$$

Equivalently, if and only if it is strictly convex along any line, i.e. if and only if $g : \mathbb{R} \rightarrow \mathbb{R}$

$$g(\alpha) := f(y + \alpha u)$$

is strictly convex for any $y \in \mathbb{R}^n$ and $\|u\|_2 = 1$

Equivalently, if and only if for every $x, y \in \mathbb{R}^n$

$$f(y) > f(x) + \nabla f(x)(y - x)$$

If $\nabla f(x) = 0$, $x$ is the only global minimum.
The **Hessian matrix** of a function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ is

$$\nabla^2 f (x) = \begin{bmatrix}
\frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\
\frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \frac{\partial^2 f(x)}{\partial x_1^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\
\vdots & \vdots & \cdots & \vdots \\
\frac{\partial^2 f(x)}{\partial x_1 \partial x_n} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2}
\end{bmatrix}$$

The Hessian matrix encodes **curvature** in every direction

$$f''_u (x) = u^T \nabla^2 f (x) u$$

where $\|u\|_2 = 1$
A quadratic form \( q : \mathbb{R}^n \) is a function of the form

\[
q(x) := \frac{1}{2} x^T A x + b^T x + c
\]

where \( A \) is symmetric

Generalization of second-order polynomial to multiple dimensions
Clarification about symmetric matrices

All symmetric matrices have an eigendecomposition of the form

\[ A = UDU^T \]

where \( U = [u_1 \ u_2 \ \cdots \ u_n] \) is an orthogonal matrix

The eigenvalues \( \lambda_1, \lambda_2, \ldots, \lambda_n \) can be positive, negative or zero

For any vector \( x \)

\[ x^T A x = \sum_{i=1}^{n} \lambda_i \left( u_i^T x \right)^2 \]
Clarification about symmetric matrices

\[ \lambda_1 = \max_{\|u\|_2 = 1} u^T A u \]
\[ u_1 = \arg \max_{\|u\|_2 = 1} u^T A u \]

\[ \lambda_2 = \max_{\|u\|_2 = 1, u \perp u_1} u^T A u \]
\[ u_2 = \arg \max_{\|u\|_2 = 1, u \perp u_1} u^T A u \]

\[ \lambda_i = \max_{\|u\|_2 = 1, u \perp u_1, \ldots, u_{i-1}} u^T A u \]
\[ u_i = \arg \max_{\|u\|_2 = 1, u \perp u_1, \ldots, u_{i-1}} u^T A u \]
Hessian matrix

Consider the eigendecomposition of $\nabla^2 f(x) = UDU^T$

Maximum curvature at $x$

$$\max_u q''_u (x) = \max_u u^T \nabla^2 f(x) u = \max_i \lambda_i$$

Minimum curvature at $x$

$$\min_u q''_u (x) = \min_u u^T \nabla^2 f(x) u = \min_i \lambda_i$$
Positive (semi)definite matrices

If all the eigenvalues $\lambda_i \geq 0$ then the matrix is positive semidefinite.

If $\nabla^2 f$ is always positive semidefinite, $f$ is convex.

If all the eigenvalues $\lambda_i > 0$ then the matrix is positive definite.

If $\nabla^2 f$ is always positive definite, $f$ is strictly convex.
Quadratic form with positive-definite Hessian matrix
Quadratic form with negative-definite Hessian matrix
Quadratic form with Hessian matrix that is not pos./neg.-semidefinite
The second-order or quadratic approximation of $f$ at $x$ is

$$f_x^2(y) := f(x) + \nabla f(x)(y - x) + \frac{1}{2} (y - x)^T \nabla^2 f(x) (y - x)$$

Lemma:

$$\lim_{y \to x} \frac{f(y) - f_x^2(y)}{\|y - x\|_2^2} = 0$$
Derivatives and convexity

Optimization algorithms in 1D

Multiple dimensions

Optimization algorithms
Gradient descent

**Intuition:** Move in the direction of maximum negative variation

1. Choose a random initialization $x \in \mathbb{R}$

2. For $i = 1, 2, \ldots$ compute

$$x_i = x_{i-1} - \alpha \nabla f (x_{i-1})$$

until $\|\nabla f (x_{i-1})\|_2 \leq \epsilon$

The step size $\alpha$ is a fixed constant
Gradient descent, $\alpha = 1$
Gradient descent, $\alpha = 6$
Quadratic function

The global minimum $x^*$ of a convex quadratic form

$$q(x) := \frac{1}{2} x^T A x + b^T x + c$$

can be reached from $x$ by

$$x^* = x - \nabla^2 f (x_{i-1})^{-1} \nabla f (x_{i-1}) = x - A^{-1} (Ax + b)$$
Newton’s method

**Intuition:** Iteratively minimize the local quadratic approximation of $f$

1. Choose a random initialization $x \in \mathbb{R}$

2. For $i = 1, 2, \ldots$ compute

$$x_i = x_{i-1} - \nabla^2 f (x_{i-1})^{-1} \nabla f (x_{i-1})$$

until $\|\nabla f (x_{i-1})\|_2 \leq \epsilon$
Newton’s method
Gradient descent
Newton’s method
Stochastic gradient descent

If a function $f$ can be decomposed into

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$$

We can apply stochastic gradient descent to minimize it

1. Choose a random initialization $x \in \mathbb{R}$

2. For $i = 1, 2, \ldots$ compute

$$x_i = x_{i-1} - \alpha \nabla f_i(x_{i-1})$$

until $\|x_i - x_{i-1}\|_2 \leq \epsilon$

Intuition: If the $f_i$ are similar on average we converge
Stochastic gradient descent

Model $\mathcal{M}(x)$ with parameters $x$, data $d_1, \ldots, d_m$

**Aim:** Minimize average error

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} (\mathcal{M}(x) - d_i)^2 = \frac{1}{m} \sum_{i=1}^{m} f_i(x)$$

**Example:** Fit linear model

$$f(x) = \frac{1}{m} \sum_{i=1}^{m} (x^T d_i - y_i)^2$$