Lecture 13: Unconstrained Optimization

Sondre Pedersen

Feb 20, 2025

Now we are going back to the basics. This could have been taught before constrained optimization, but the order was switched in order to be prepared for the labs in time.

Outline

- Optimality conditions for unconstrained optimization.
- Ingredients in gradient descent algorithms for unconstrained optimization
  - Descent directions (steepest descent, Newton, Quasi-Newton)
  - How far to walk in descent direction (line search, trust region)
  - Termination criteria
- Scaling

Example use - Machine Learning

Unconstrained optimization can be used to learn from data and make predictions. Linear regression for example, the most basic ML algorithm. Linear least squares has an explicit solution. Nonlinear least squares is a little more complicated. The point is that least squares can be formulated as an optimization problem. Given a bunch of data, we want to find the best coefficients a, b to get the lowest squared distance to actual data points.

$$\min_{a,b} \quad \sum_{i=1}^{n} (y_i - ax_i - b)^2$$

When the line is linear, we can write the objective function as

$$\min_{\theta} \quad (y - x\theta)^\top (y - x\theta)$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \; x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}, \; \theta = \begin{bmatrix} a \\ b \end{bmatrix}$$

This is nice and easy, so we can find a direct solution: $\theta = (x^\top x)^{-1} x^\top y$. In machine learning however, the functions are not linear, so more powerful methods are required.

1 Unconstrained Optimization

Before we had constrained optimization

$$\min_{x \in \mathbb{R}^n} f(x) \text{ s.t. } \begin{cases} c_i(x) = 0, \; i \in \mathcal{E} \\ c_i(x) \geq 0, \; i \in \mathcal{I} \end{cases}$$

Now we take away all the constraints, and only look at $\min_{x \in \mathbb{R}^n} f(x)$. Note: from now on me assume is $f(x)$ is 'smooth': $f \in C^1$ or $f \in \mathcal{C}^2$. That is ∇f exists and is continuous and ∇f^2 exists and is continuous.
What is a solution? Same as before. x^* is a global solution : $f(x^*) \leq f(x), \; \forall x$. The local solution is also the same as before. x^* is a local solution: $f(x^*) \leq f(x), \; \forall x \in \mathcal{N}(x)$.

Necessary condition for optimality - x^* is a local solution and $f \in C^1 \implies \nabla f(x^*) = 0$.

Taylor expansions The form familiar from calculus:

$$f(x) = f(a) + (x-a)f'(x) + \frac{1}{2}(x-a)^2 f''(a) + \dots.$$

More useful in this course:

$$f(x+h) = f(x) + hf'(x) + \frac{1}{2}h^2 f''(x) + \dots.$$

These are equivalent by change of variables. Taylor's theorem is that if f is continuously differentiable, then

$$f(x+p) = f(x) + \nabla_f(x+tp)^\top p, \qquad \text{for some } t \in (0,1).$$

For second order (f is twice continuously differentiable):

$$f(x+p) = f(x) + \nabla f(x)^\top p + \frac{1}{2}p^\top \nabla^2 f(x+tp)p, \qquad \text{for some } t \in (0,1).$$

Sufficient conditions for optimality - $\nabla f(x^*) = 0$ and $\nabla^2 f(x^*) > 0 \implies x^*$ strict local solution.

2 General algorithm

For solving $\min_x f(x)$.

```
1. Initial guess x_0, k = 0
2. While termination criteria not fulfilled:
3.  Find descent direction p_k (in x_k)
4.  Step along p_k to x_{k+1}: x_{k+1} = x_k + α_k p_k
5.  k = k + 1
6. end
```

Termination criteria - Given a small tolerance $\epsilon > 0$. Here is a list of different criteria:

- $||x_k - x^*|| \leq \epsilon$ or $|f(x_k) - f(x^*)| \leq \epsilon$ (not possible, we don't know x^*)
- $||\nabla f(x_k)|| \leq \epsilon$
- $||x_k - x_{k-1}|| \leq \epsilon$
- $|f(x_n) - f(x_{n-1})| \leq \epsilon$
- $k > k_{max}$

Descent directions - There are different ways to pick a direction to move:

- Steepest descent: $p = -\nabla f(x_k)$
- Newton: approximate $f(x)$ around x_k. $f(x_k + p) \approx +\nabla f(x_k)^\top p + \frac{1}{2}p^\top \nabla^2 f(x_k)p := m_k(p)$.
- Quasi-Newton: the same as Newton, but use an approximation for the Hessian.

We call the function $m_k(p)$ the model function. We want to find the p that minimize the model function:

$$\begin{aligned} \min_p \qquad & m_k(p) \to \nabla_p m_k(p) = 0 \\ & \nabla_p m_k(p) = \nabla f(x_k) + \nabla^2 f(x_k)p = 0 \\ \to & p = -[\nabla^2 f(x_k)]^{-1} \nabla f(x_k) \end{aligned}$$

How does the Newton direction differ from the steepest descent? When the function is quadratic, the Newton direction goes directly towards the local optimum for a quadratic approximation. T

Step length - How far should we walk along p_k? This is also known as globalization strategies - making things work when we are far from the optimum.

- Line search: Find α that approximately solve $\min_\alpha f(x_k + \alpha p_n) \to \alpha_k$
- Trust region (not covered in this course)

It's hard to say which one of these methods is the best. It depends on the problem.