Deterministic Optimization

Unconstrained Optimization: Derivative Based

Shabbir Ahmed

Anderson-Interface Chair and Professor

School of Industrial and Systems Engineering

Gradient Descent

Gradient Descent

Learning objective:

- Examine the gradient descent method

Unconstrained Optimization: Derivative Based

$$(P): \quad \min f(x) \quad \text{s.t. } x \in \mathbb{R}^n$$

where $f : \mathbb{R}^n \to \mathbb{R}$ is continuous and twice differentiable.

- Lesson 1: Optimality Conditions
- Lesson 2: Gradient Descent
- Lesson 3: Newton's Method

Descent Methods

$$(P): \quad \min f(x) \;\; \text{s.t.} \;\; x \in \mathbb{R}^n.$$

Basic paradigm of descent methods:
- Choose an initial solution x^0.
- Choose a descent direction d^0.
- Choose a step size α_0.
- Update the solution $x^1 = x^0 + \alpha_0 d^0$.
- If some stopping criteria is met, STOP; else repeat with current solution.

Gradient Descent

- Let x^k be the current iterate, and we want to chose a "downhill direction" d^k and a step size α such that $f(x^k + \alpha d^k) < f(x^k)$.

- By Taylor's expansion:

$$f(x^k + \alpha d^k) \approx f(x^k) + \alpha \nabla f(x^k)^\top d_k.$$

  So we want $\nabla f(x^k)^\top d^k < 0$. The steepest descent direction is $d^k = -\nabla f(x^k)$.

Gradient Descent

- Step size
  - Line search: Define $g(\alpha) := f(x^k + \alpha d^k)$. Choose α to minimize g.
  - Fixed step size: Fix α a priori (may not converge if α is too big)
- Update the iterate as $x^{k+1} \leftarrow x^k - \alpha \nabla f(x_k)$.
- Stop if $||\nabla f(x_k)|| \leq \epsilon$.

Example: Gradient Descent Iteration

$$\min f(x) = (x_1 + 1)^4 + x_1 x_2 + (x_2 + 1)^4$$

- Let $x^0 = [0, 1]^\top$, and $f(x^0) = 17.0$.
- The gradient $\nabla f(x) = [4(x_1 + 1)^3 + x_2, x_1 + 4(x_2 + 1)^3]$. At x^0, $\nabla f(x^0) = [5, 32]^\top$.
- The next iterate $x^1 = x^0 - \alpha \nabla f(x^0) = [-5\alpha,\ 1 - 32\alpha]^\top$.

Example: Gradient Descent Iteration

- Then $g(\alpha) = f(x^1) = (-5\alpha + 1)^4 - 5\alpha(1 - 32\alpha) + (1 - 32\alpha + 1)^4$.

- Minimizing $g(\alpha)$, we get $\alpha = 0.0527$.

- Therefore $x^1 = [-0.2635, -0.6864]^\top$ and $f(x^1) = 0.4848$.

Behavior of Gradient Descent

- At any point x^k with $\nabla f(x^k) \neq 0$, the gradient descent produces the most rapid convergence (locally).

- Initial progress is good, but near a stationary point, the convergence behavior is bad.

Behavior of Gradient Descent

- "Zig-zags," i.e., each successive direction (of move) is perpendicular to the previous direction.

  Let d^k be the gradient descent direction and α_k be the optimum step length at step k, i.e. $0 = \frac{dg(\alpha)}{d\alpha}|_{\alpha=\alpha_k} = \nabla f(x^k + \alpha_k d^k)^\top d^k = \nabla f(x^{k+1})^\top d^k$.

  Since $d^{k+1} = -\nabla f(x^{k+1})$, we have that ${d^{k+1}}^\top d^k = 0$, i.e. two successive directions are perpendicular.

- Very small step sizes near stationary point.

Summary

- The gradient descent method moves from on iteration to the next by moving along the negative of the gradient direction in order to minimize the function