Gradient Descent with fixed step size

Lipschitz continuity:

$$\|f(x) - f(y)\|_2 \leq L \|x-y\|_2 \quad \forall \ x, y \in X$$

Theorem 6.1 (Ryan Tibshirani)

Suppose $f: \mathbb{R}^n \to \mathbb{R}$ is convex and differentiable and that its gradient is Lipschitz continuous with constant L, then after running gradient descent with step size $\alpha \leq \frac{1}{L}$ for k iterations

$$f(x^{(k)}) - f(x^*) \leq \frac{\|x_0 - x^*\|_2^2}{2\alpha k}$$

$\alpha = \frac{1}{L}$ is too small in practice

Two General Strategies

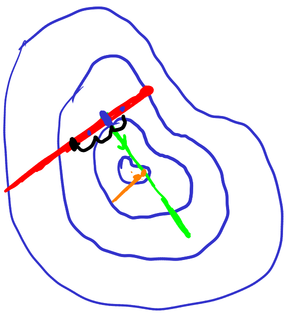
1. **Line Search** Today!

Choose direction $\hat{d}$

choose a step size

$$\alpha^* = \underset{\alpha}{\arg\min} \, f(\vec{x} + \alpha \hat{d})$$

$$\vec{x}_k \leftarrow \vec{x}_{k-1} + \alpha^* \hat{d}$$

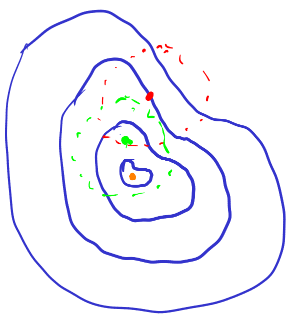


2. **Trust Region**

Choose $\hat{f}$ + trust region for $\hat{f}$

$$\vec{x}_k \leftarrow \underset{\vec{x}}{\arg\min} \, \hat{f}(\vec{x})$$

subject to $d(\vec{x}, \vec{x}_{k-1}) \leq \delta$

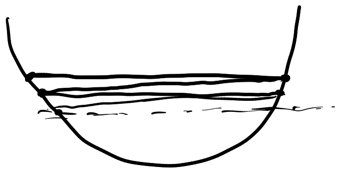


Line Search

When to stop searching the line?

Condition 1: Decrease

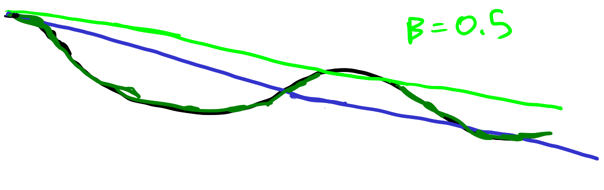
$$f(\vec{x} + \alpha \hat{d}) < f(\vec{x})$$



Not Sufficient!

Condition 2: Sufficient Decrease (Armijo)

$$f(\vec{x} + \alpha \hat{d}) \leq f(\vec{x}) + \beta \alpha \nabla_{\hat{d}} f(\vec{x}) \qquad \beta \in (0,1]$$



Condition 3: Curvature

$$\nabla_{\hat{d}} f(\vec{x} + \alpha \hat{d}) \geq \sigma \nabla_{\hat{d}} f(\vec{x})$$

2 and 3 together: Wolfe

Other: Strong Wolfe
Goldstein

Theorem 3.2 (N+W)

Consider any iteration of the form

$$\vec{x}^{(k+1)} \leftarrow x^{(k)} + \alpha^{(k)} \hat{d}^{(k)}$$

where $\hat{d}$ is a descent direction and $\alpha^{(k)}$ satisfies Wolfe Conditions. Suppose f is bounded below and continuously differentiable on an open set N containing the level set $\{x : f(x) \leq f(x^0)\}$ and ∇f is Lipschitz continuous with constant L on N, then

$$\sum_{k=0}^{\infty} \cos^2 \theta^{(k)} \|\nabla f(x^{(k)})\|^2 < \infty$$

where $\theta^{(k)}$ is the angle between $\hat{d}^{(k)}$ and $\nabla f(x^{(k)})$

$$\cos \theta^{(k)} = -\frac{\nabla f(x^{(k)})^T \hat{d}^{(k)}}{\|\nabla f(x^{(k)})\|}$$

---

1 Will a grad descent algorithm that chooses α that meets the Wolfe conditions find a local minimum? No

---

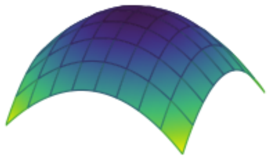
Can the algorithm always find an α that meets the Wolfe conditions? Yes

Lemma 3.1 (N+W)

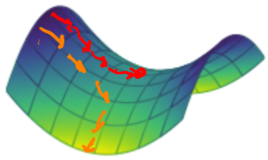
Suppose that $f: \mathbb{R}^n \rightarrow \mathbb{R}$ is continuously differentiable, Let $\hat{d}$ be a descent direction at $\vec{x}$ and assume that f is bounded below. Then $0 < \beta < \sigma < 1$, then there exist intervals satisfying Wolfe.

$$\sum_{k=0}^{\infty} \cos^2 \theta^{(k)} \|\nabla f(x^{(k)})\|^2 \Longrightarrow \nabla f(x^{(k)}) \longrightarrow 0$$

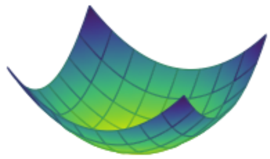
Critical Points in $\mathbb{R}^n$



A *local maximum*. The gradient at the center is zero, but the Hessian is negative definite.



A *saddle*. The gradient at the center is zero, but it is not a local minimum.



A *bowl*. The gradient at the center is zero and the Hessian is positive definite. It is a local minimum.

Local Minimum

Not sufficient

FONC $\quad \nabla f(\vec{x}) = 0$

SONC $\quad \nabla^2 f(\vec{x}) \succeq 0$ positive semidefinite

A matrix A is positive semidefinite if

$$\vec{x}^T A \vec{x} \geq 0 \quad \forall \ \vec{x} \in \mathbb{R}^n$$

equivalent to

A is symmetric and all eigenvalues ≥ 0

$$\begin{bmatrix} 1 & 0 & \\ & 1 & \\ & & 3 \end{bmatrix}$$

SOSC: $\nabla^2 f(\vec{x}) \succ 0 \quad$ positive definite

Not necessary for strong local

consider $f(x) = x^4$

$f''(0) = 0$

How to find α that satisfies conditions

Simple + Effective:

Backtracking

$\alpha \leftarrow \alpha_0$

$\rho \in (0,1)$

while Armijo is not satisfied

$\alpha \leftarrow \rho\alpha$

return α

More complicated, but satisfies Wolfe

Alg 3.5 + 3.6 in N+W

- Line search is a practical way to choose step size
- Wolfe cond on step size guarantee that we make enough progress
- If line search finds α meets Wolfe, gauranteed convergence to critical pt.

  Theorem 3.2
- Backtracking practical way to choose α