EE270
Large scale matrix computation, optimization and learning

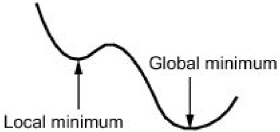
Instructor : Mert Pilanci

Stanford University

Tuesday, Feb 11 2020

Randomized Linear Algebra and Optimization
Lecture 12: Gradient Descent

Optimization: Gradient Descent



- Consider unconstrained minimization of $f : \mathbb{R}^d \to \mathbb{R}$, differentiable function
- we want to solve

$$\min_{x \in \mathbb{R}^d} f(x)$$

- **Gradient descent:** choose initial $x_0 \in \mathbb{R}^d$ and repeat

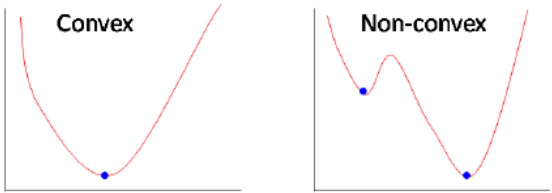
$$x_{t+1} = x_t - \mu_t \nabla f(x_t)$$

- for $t = 1, ..., T$

Convex vs Non-convex functions

- a function f is called **convex** if

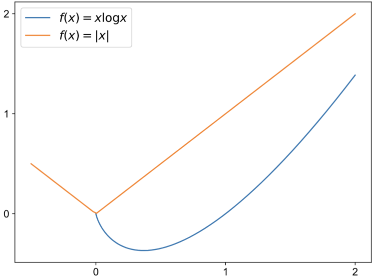
$$\forall x_1, x_2 \in \mathcal{X}, \forall t \in [0,1] : \quad f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$



Convex vs Non-convex functions

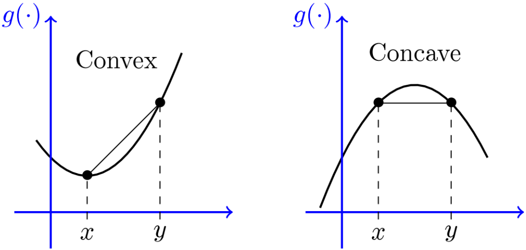
- a function f is called **strictly convex** if

$$\forall x_1 \neq x_2 \in \mathcal{X}, \forall t \in [0,1] : \quad f(tx_1 + (1-t)x_2) < tf(x_1) + (1-t)f(x_2)$$



Concave functions

- a function f is called (strictly) **concave** if $-f$ is (strictly) convex



Differentiable functions

- A one dimensional function $f : \mathbb{R} \to \mathbb{R}$ is differentiable if the derivative
  $f'(x) := \lim_{h \to 0} \frac{f(x+h)-f(x)}{h}$ exists
- Suppose that all partial derivatives of $f : \mathbb{R}^d \to \mathbb{R}$ exists
  The gradient $\nabla f(x)$ is the vector of partial derivatives
  $[\nabla f(x)]_i = \frac{\partial}{\partial x_i} f(x)$

Alternative definitions of convexity

- Assume that $f(x) : \mathbb{R}^d \to \mathbb{R}$ is differentiable. Then f is convex, if and only if for every x, y the inequality

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

is satisfied

Twice differentiable functions

- Suppose that all second derivatives of $f : \mathbb{R}^d \to \mathbb{R}$
  $\frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} f(x)$ exists
  The Hessian $\nabla^2 f(x)$ is the matrix of partial derivatives
  $[\nabla^2 f(x)]_{ij} = \frac{\partial}{\partial x_i}\frac{\partial}{\partial x_j} f(x)$

Twice differentiable convex functions

- A twice differentiable function $f(x)$ is convex if and only if the Hessian $\nabla^2 f(x)$ is positive semi-definite for all $x \in \mathbb{R}^d$
- Suppose that f is convex and differentiable, then x^* is a global minimizer of f if and only if $\nabla f(x^*) = 0$

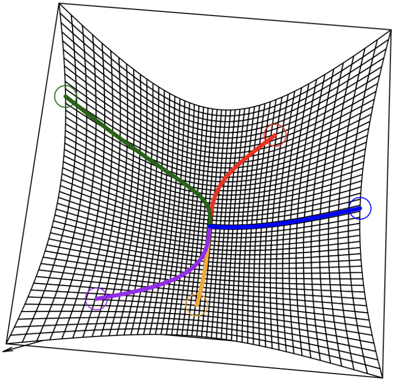
Gradient descent for differentiable functions

- $-\nabla f(x)$ is the direction of largest instantaneous decrease
- Gradient Descent (GD):

$$x_{t+1} = x_t - \mu_t \nabla f(x_t)$$

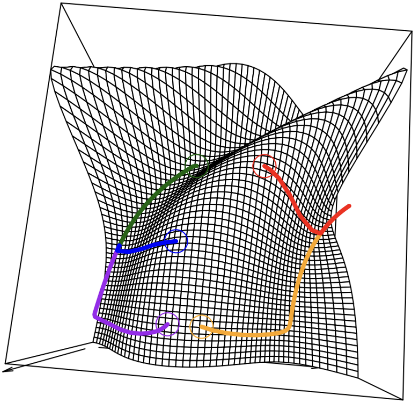
- where μ_t is the step size at iteration t.
- if μ_t is sufficiently small and $\nabla f(x_t) \neq 0$, guaranteed to decrease the value of f
- If f is convex, converges to **global minimum** under mild conditions

Gradient descent for convex functions



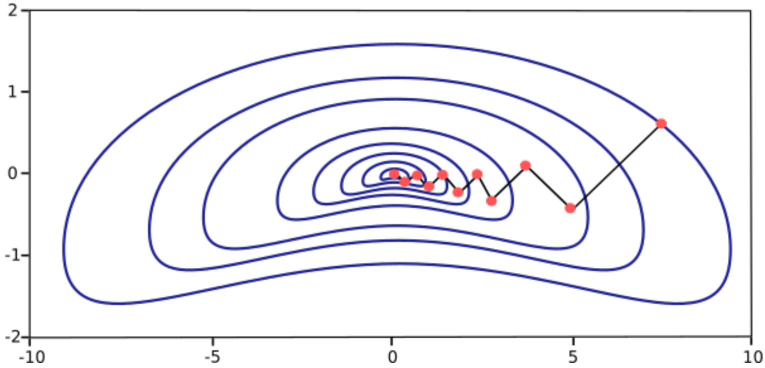
slide credit: R. Tibshirani

Gradient descent for non-convex functions



slide credit: R. Tibshirani

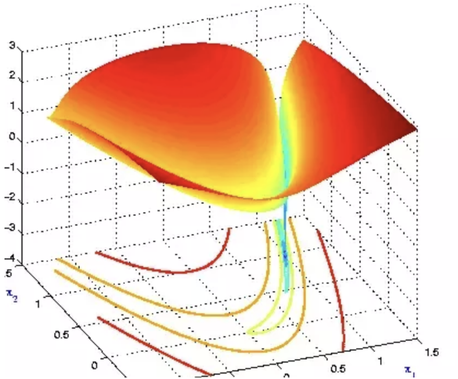
Gradient descent iterations



slide credit: A. Quesada

Gradient descent on highly curved functions

- Rosenbrock function (non-convex)
  $f(x_1, x_2) = (a - x_1)^2 + b(x_2 - x_1^2)^2$
  where a and b are parameters, e.g., $a = 1, b = 100$
  has a global minimum at $(x_1, x_2) = (a, a^2)$



Optimizing convex least squares cost

- Consider
$$\min_x \underbrace{\frac{1}{2}\|Ax - b\|_2^2}_{f(x)}$$
- gradient $\nabla f(x) = A^T(Ax - b)$
- Gradient Descent:
$$x_{t+1} = x_t - \mu A^T(Ax_t - b)$$
- fixed step size $\mu_t = \mu$

Optimizing convex least squares cost

- Basic (in)equality method

  (1) x^* minimizes $f(x)$, hence $\nabla f(x^*) = A^T(Ax^* - b) = 0$

  (2) $x_{t+1} = x_t - \mu A^T(Ax_t - b)$

  (3) define error $\Delta_t = x_t - x^*$

Optimizing convex least squares cost

- Basic (in)equality method
  (1) x^* minimizes $f(x)$, hence $\nabla f(x^*) = A^T(Ax^* - b) = 0$
  (2) $x_{t+1} = x_t - \mu A^T(Ax_t - b)$
  (3) define error $\Delta_t = x_t - x^*$

- $\Delta_{t+1} = \Delta_t - \mu A^T A \Delta_t$

Optimizing convex least squares cost

- run gradient descent M iterations, i.e., $t = 1, ..., M$
- $\Delta_M = (I - \mu A^T A)^M \Delta_0$
- $\|\Delta_M\|_2 \leq \sigma_{\max}\left((I - \mu A^T A)^M\right) \|\Delta_0\|_2$

  $\sigma_{\max}\left(I - \mu A^T A\right)^M = \max_{i=1,..,d} \left|1 - \lambda_i(A^T A)\right|^d$

  where λ_i is the i-th eigenvalue in decreasing order

Optimizing convex least squares cost

Questions?