

CME 252: Gradient Descent

AJ Friend
ICME, Stanford University

Gradient Descent

Outline

Gradient Descent

Introduction

minimize $f(x)$

- ▶ initial assumptions on f :
 - ▶ convex
 - ▶ twice differentiable
 - ▶ unconstrained; domain of f is \mathbf{R}^n
 - ▶ assume min is attained: $p^\star = \inf_x f(x)$

Iterative Methods

- ▶ iterative methods produce a sequence of points, x^k for $k = 1, 2, \dots$ such that

$$f(x^k) \rightarrow p^\star$$

- ▶ also consider the algebraic problem of finding x^\star such that

$$\nabla f(x^\star) = 0$$

Optimization Master Algorithm

- ▶ solve “hard” problem via sequence of “easier” problems
 - ▶ approximate locally as “easier” problem
 - ▶ solve easy problem
 - ▶ move to new point
 - ▶ repeat until “solved”
- ▶ quadratic optimization problems via linear algebra
- ▶ unconstrained convex opt. via quadratic opt.
- ▶ constrained convex opt. via unconstrained convex opt.
- ▶ nonconvex opt. via convex opt.

Gradients

- ▶ $f(x) : \mathbf{R}^n \rightarrow \mathbf{R}$
- ▶ gradient $\nabla f(x) : \mathbf{R}^n \rightarrow \mathbf{R}^n$ given by

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

- ▶ **intuition:** $\nabla f(x)$ points in direction of steepest **ascent**

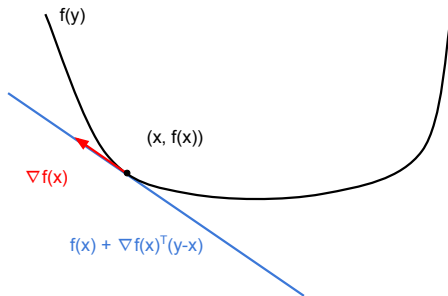
Affine Underestimator

- ▶ f is convex if and only if $\text{dom}(f)$ is convex and

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

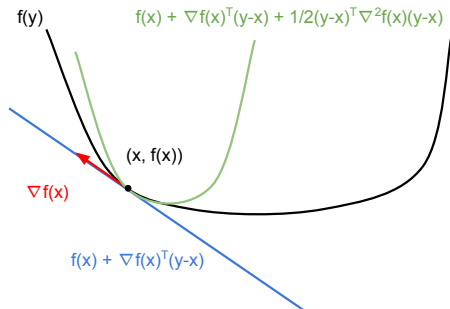
for all $x, y \in \text{dom}(f)$

- ▶ first-order Taylor approximation is a **global underestimator** of f



Quadratic Estimator

- ▶ second-order Taylor $f(y) \approx f(x) + \nabla f(x)^T(y - x) + 1/2(y - x)\nabla^2 f(x)(y - x)$
may not be over- or under-estimator
- ▶ estimator is convex as $\nabla^2 f(x) \succeq 0$



Positive semidefinite matrices

- ▶ a matrix $A \in \mathbf{R}^{n \times n}$ is **positive semidefinite** ($A \succeq 0$) if
 - ▶ A is **symmetric**: $A = A^T$
 - ▶ $x^T A x \geq 0$ for all $x \in \mathbf{R}^n$
- ▶ $A \succeq 0$ if and only if all **eigenvalues** of A are nonnegative
- ▶ intuition: graph of **convex** $f(x) = x^T A x$ looks like a bowl
- ▶ example: ρI for any $\rho \geq 0$ is PSD
- ▶ if $A = 100I$, $\{x \mid x^T A x \leq 1\}$ is a ball of radius 0.1
- ▶ if $A = 0.01I$, $\{x \mid x^T A x \leq 1\}$ is a ball of radius 10

Computing Gradients

- ▶ $f(x) = a^T x$
 - ▶ $\nabla f(x) = a$
- ▶ $f(x) = x^T B x$
 - ▶ $\nabla f(x) = (B + B^T)x$
 - ▶ $\nabla f(x) = 2Bx$ (if B symmetric)
 - ▶ $\nabla^2 f(x) = (B + B^T)$ (symmetric part of B)
- ▶ chain rule: $f(x) = g(Ax + b) : \mathbf{R}^n \rightarrow \mathbf{R}$
 - ▶ $\nabla f(x) = A^T \nabla g(Ax + b)$
 - ▶ $\nabla^2 f(x) = A^T \nabla^2 g(Ax + b) A$
- ▶ reference: “The Matrix Cookbook”

Gradient Example

- ▶ log-sum-exp function $g(y) : \mathbf{R}^n \rightarrow \mathbf{R}$ with

$$g(y) = \log \sum_{i=1}^n e^{y_i}, \quad \nabla g(y) = \frac{1}{\sum_{i=1}^n e^{y_i}} \begin{bmatrix} e^{y_1} \\ \vdots \\ e^{y_n} \end{bmatrix}$$

- ▶ $f(x) = g(Ax + b)$
- ▶ let $z_i = \exp(a_i^T x + b_i)$ (a_i^T : i th row of A)

$$\nabla f(x) = A^T \nabla g(Ax + b) = \frac{1}{\mathbf{1}^T z} A^T z$$

Least-squares Example

- ▶ minimize $f(x) = \|Ax - b\|_2^2$ for $A \in \mathbf{R}^{m \times n}$
- ▶ possibly no x such that $Ax = b$
- ▶ A “skinny”, $m \geq n$, and full rank (least-squares has unique solution)
- ▶ note:

$$\|Ax - b\|_2^2 = (Ax - b)^T(Ax - b) = x^T A^T A x - 2(A^T b)^T x + b^T b$$

- ▶ solution via the **normal equations**
- ▶ i.e., $\nabla f(x^\star) = 0$ implies

$$A^T A x^\star = A^T b$$

- ▶ call a linear system solver to find x^\star

Convex Quadratic Problems

- ▶ least-squares is an example of an unconstrained **convex quadratic** problem

$$\text{minimize} \quad 1/2x^T Qx + b^T x,$$

where $Q \succeq 0$ (find lowest point in a bowl)

- ▶ prototypical convex function/problem
- ▶ without linear algebra software, how would you solve?

Gradient Descent:

- ▶ $-\nabla f(x)$ is a **descent direction**
- ▶ repeatedly move a little bit in that direction

$$x^{k+1} = x^k - \alpha^k \nabla f(x^k)$$

- ▶ α^k is a stepsize (to be chosen)
- ▶ repeat until “solved” (when?)

Stopping Criteria

assume $f(x^*) = p^*$ is attained

- ▶ $\|\nabla f(x^k)\| \leq \epsilon$
- ▶ $\|f(x) - p^*\| \leq \epsilon$ (if known or estimated)
- ▶ $\|x - x^*\|_2 \leq \epsilon$ (if known or estimated)

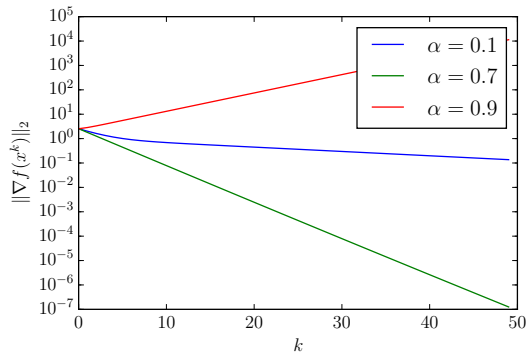
Gradient Descent in Python

with α^k fixed:

```
def f(x):  
    return x.dot(A).dot(x)/2 - b.dot(x)  
  
def g(x):  
    return A.dot(x) - b  
  
for i in range(100):  
    gr = g(x)  
    x = x - alpha*gr
```

Gradient Descent Example

- quadratic problem: minimize $1/2x^T Ax - b^T x$, $A \succ 0$



- how to choose α^k without hand-tuning?

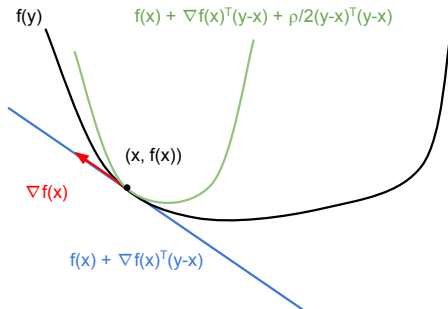
Re-interpret Gradient Descent

- ▶ what if you were given a **quadratic upper-bound** on f ?
- ▶ at x^k had **guarantee** that

$$f(y) \leq q(y) = f(x^k) + \nabla f(x^k)^T (y - x^k) + 1/2(y - x^k)^T(\rho I)(y - x^k),$$

for some $\rho > 0$

- ▶ what could you do with it?



Quadratic Over-estimator

- ▶ $f(y) \leq q(y)$ everywhere
- ▶ $f(x^k) = q(x^k)$
- ▶ $\nabla f(x^k) = \nabla q(x^k) \neq 0$
- ▶ move to the minimum of $q(y)$

$$x^{k+1} = \operatorname{argmin} q(y)$$

- ▶ guaranteed descent: $f(x^{k+1}) \leq q(x^{k+1}) < q(x^k) = f(x^k)$

Gradient Step

$$q(y) = f(x^k) + \nabla f(x^k)^T (y - x^k) + 1/2(y - x^k)(\rho I)(y - x^k)$$

- ▶ move to

$$x^{k+1} = \underset{y}{\operatorname{argmin}} q(y)$$

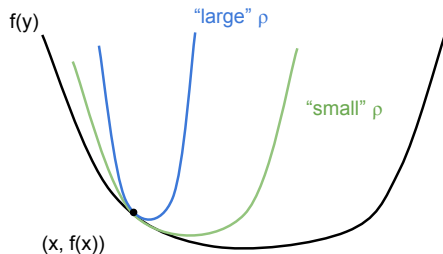
- ▶ solve algebraically: $\nabla q(y) = \nabla f(x^k) + \rho(y - x^k) = 0$ implies

$$x^{k+1} = y = x^k - \frac{1}{\rho} \nabla f(x^k)$$

- ▶ **exactly** the gradient step with $\alpha^k = 1/\rho$

Gradient Step

- ▶ gradient descent is equivalent to minimizing a quadratic upper bound on f
- ▶ “small” ρ : low curvature, large step $\alpha^k = 1/\rho$
- ▶ “large” ρ : high curvature, more conservative step



Quadratic Upper Bound

- ▶ how to **guarantee** quadratic upper bound?
- ▶ Taylor theorem: for all $x, y \in \mathbf{dom}(f)$

$$f(y) = f(x) + \nabla f(x)^T(y - x) + 1/2(y - x)\nabla^2 f(z)(y - x)$$

for some z

- ▶ if some $M > 0$ such that $\nabla^2 f(x) \preceq MI$ for all x then

$$f(y) \leq f(x) + \nabla f(x)^T(y - x) + M/2\|y - x\|_2^2$$

is an upper bound

Quadratic Upper Bound

intuition for $\nabla^2 f(x) \preceq MI$:

- ▶ $MI - \nabla^2 f(x)$ is PSD (all nonnegative eigenvalues)
- ▶ M larger than any eigenvalue of $\nabla^2 f(x)$
- ▶ MI has stronger curvature than $\nabla^2 f(x)$ in every direction
- ▶ change in gradient is bounded:

$$\|\nabla f(x^{k+1}) - \nabla f(x^k)\|_2 \leq M\|x^{k+1} - x^k\|_2$$

if you know M :

- ▶ take $\alpha^k = 1/M$
- ▶ taking smallest possible M gives larger steps and potentially faster convergence

Other Over-estimators

- ▶ anything special about our quadratic over-estimator?
- ▶ not really; just that minimizing it was easy
- ▶ other estimators are possible
 - ▶ collect many gradients instead of just one (bundle methods)
 - ▶ other convex functions to over-estimate or just estimate
- ▶ leads to other methods, but minimizing at each step may be harder
- ▶ recurring theme: approximate “hard” problem with simple model; solve; repeat

Line Search

- ▶ M usually not known in practice, so how to choose α^k ?
- ▶ line search!
- ▶ exact: $\alpha^k = \operatorname{argmin}_t f(x^k - t\nabla f(x^k))$ (usually expensive)
- ▶ approximate quadratic model search (assuming $\nabla^2 f(x) \preceq MI$ exists)
 - ▶ $y = x^k - t\nabla f(x^k)$
 - ▶ start with some M
 - ▶ decrease t (increase $M = 1/t$) until $f(x^{k+1}) \leq q(x^{k+1})$
 - ▶ that is, until quadratic upper bound “looks” accurate
 - ▶ but how to keep the step size from getting too small?
 - ▶ re-estimate M as $\|\nabla f(x^{k+1}) - \nabla f(x^k)\|/\|x^{k+1} - x^k\|$
- ▶ many other line search options

curvature

- ▶ strongly convex means minimum eigenvalue
- ▶ lipschitz continuous means max eigenvalue of hessian

condition numbers

quadratic under and over estimators

- ▶ convergence analysis

line search

- ▶ probably not too much
- ▶ simple linesearch based on quadratics

Stochastic gradient descent

- ▶ and in parallel
- ▶ decreasing stepsize (show example of hopping)
- ▶ give an example stepsize but skip analysis

Nonconvex problems

- ▶ restarting
- ▶ monitoring
- ▶ babysitting

Newton's method

- ▶ better quadratic approximation
- ▶ but what if not PSD. . . .

Problems with Constraints

- ▶ simple barrier methods