Lab 13

DSGA-1014: Linear Algebra and Optimization

CDS at NYU
Zahra Kadkhodaie

Fall 2021

Suppose $A \in R^{2\times 2}$ is symmetric with positive eigenvalues. Describe geometrically the contour lines of $f : R^2 \to R$ given by $f(x) = x^T Ax$. Recall that the contour line for value γ is given by $\{x \in R^2 : f(x) = \gamma\}$

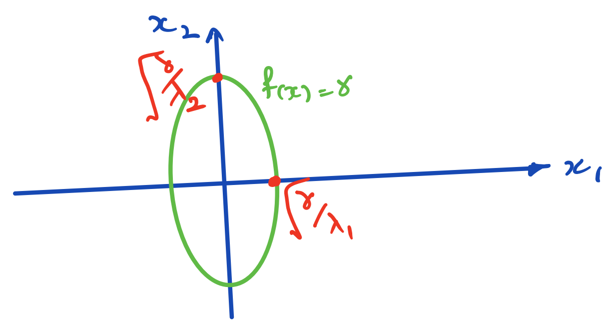
First, suppose A is diagonal $A = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}$

Then $x^T A x = \lambda_1 x_1^2 + \lambda_2 x_2^2$. Thus solving $f(x) = \gamma$ is the equation for an elipse. By the spectral theorem, any symmetric matrix is diagonal up to a rotation. Thus, generally we obtain a rotated ellipse centered at zero.

$$\lambda_1 x_1^2 + \lambda_2 x_2^2 = \gamma$$

$$\underbrace{\frac{x_1^2}{\gamma / \lambda_1}}_{\text{minor radius}} + \underbrace{\frac{x_2^2}{\gamma / \lambda_2}}_{\text{major radius}} = 1$$

if $\lambda_1 > \lambda_2$



Suppose $f : R^2 \to R$ is defined by $f(x) = x_1^2 + 100x_2^2$. Explain what issues this may pose for gradient descent.

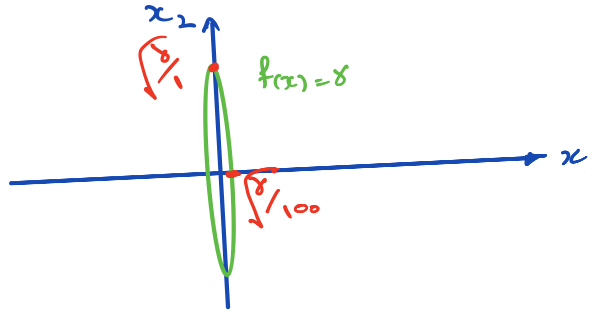
$$A = \begin{bmatrix} 100 & 0 \\ 0 & 1 \end{bmatrix}$$

$$f = x^T A x$$

The contour lines of f are very eccentric ellipses.

$$\nabla f(x) = \begin{bmatrix} 2x_1 \\ 200x_2 \end{bmatrix}$$

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



If α_t is small $\Rightarrow$ GD takes many steps to converge

If α_t is large $\Rightarrow$

$-\alpha_t \nabla f(x_t)$ overshoots in x_1 direction $\Rightarrow$ It doesn't point to the descent direction

Assume we use gradient descent when minimizing the least-square cost $f(x) = ||Ax - y||^2$. Write the gradient step update for this problem.

$x \in R^m$ $A \in R^{n \times m}$ $x^* = \arg\min f(x) \iff \nabla f(x^*) = 0$

features n = number of data points.

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

$$f(x) = \langle Ax - y, Ax - y \rangle \Rightarrow \nabla f(x) = 2A^T(Ax_t - y)$$

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

exact solution:

$(A^TA)^{-1}A^Ty$ for linearly independent columns

$A^+y = V\Sigma' U^T y$ for linearly dependent columns ← when does GD converge to this?

Show that if f is a μ-strongly convex function with minimizer x^*, then for all $x \in R^n$

exersice from lecture 13

$$f(x) - f(x^*) \leq \frac{1}{2\mu} ||\nabla f(x)||^2$$

we learned in lecture 13 (slide 10):

$$f(x) + \langle h, \nabla f(x) \rangle + \frac{\mu}{2} \|h\|^2 \leq f(x+h)$$

$$\min_h \; f(x) + \langle h, \nabla f(x) \rangle + \frac{\mu}{2} \|h\|^2$$

$$\nabla_h \left(f(x) + \langle h, \nabla f(x) \rangle + \frac{\mu}{2} \|h\|^2 \right) = \nabla f(x) + \mu h = 0$$

$$\Rightarrow h = -\tfrac{1}{\mu} \nabla f(x)$$

plug in this h:

$$f(x) - \tfrac{1}{\mu} \|\nabla f(x)\|^2 + \frac{\mu}{2} \left(\tfrac{1}{\mu^2} \|\nabla f(x)\|^2 \right)$$

$$= f(x) - \frac{1}{2\mu} \|\nabla f(x)\|^2$$ $\rightsquigarrow$ This is left-hand-side minimized w.r.t. h

Now if $h = x^* - x$:

$$\min\left\{ f(x) + \langle h, \nabla f(x) \rangle + \frac{\mu}{2} \|h\|^2 \right\} = \underbrace{f(x) - \frac{1}{2\mu} \|\nabla f(x)\|^2}_{} \leq$$

$$f(x) + \langle x^* - x, \nabla f(x) \rangle + \frac{\mu}{2} \|x^* - x\|^2 \leq \underbrace{f(x^*)}_{}$$

$$\Rightarrow f(x) - \frac{1}{2\mu} \|\nabla f(x)\|^2 \leq f(x^*)$$

$$\Rightarrow f(x) - f(x^*) \leq \frac{1}{2\mu} \|\nabla f(x)\|^2$$

Assume that we are doing gradient descent to minimize the least-square cost $f(x) = ||Ax - y||^2$. Assume that the columns of A are linearly dependent, meaning that $Ker(A) \neq \{0\}$. At which speed should gradient descent converge to the minimum? If now $Ker(A) = \{0\}$, at which speed should gradient descent converge? Note: By speed, we only ask about the dependence in t, the number of iterations, of the gap $f(x_t) - \min f$, where x_t is the position of gradient descent t iterations.

If $A \in \mathbb{R}^{n \times m}$ has linearly dependent columns $\Rightarrow Rank(A) < m$

and $\dim(Ker(A)) = m - Rank(A) > 0$

$$f(x) = \underbrace{x^T A^T A x}_{\text{quadratic form}} - \underbrace{2y^T A x}_{\text{in the } Im(A)} + y^T y$$

$A^T A$ PSD matrix

$\Rightarrow$ recall from HW 9: $f(x)$ admits a minimizer (is convex)

Recall from HW9, problem 9.4: if rank $(A) < m$, then f is not strictly convex. $\Rightarrow$ smallest singular value of $A = 0$

$\Rightarrow$ A is not strongly convex

From proposition on page 11 of lecture 13:

If α_t is constant & $\alpha_t = \frac{1}{L} \Rightarrow f(x_t) - f(x^*) \leq \frac{2L\|x_0 - x^*\|^2}{t+4}$

$\Rightarrow O\left(\frac{L}{t}\right)$

If A has linearly independent columns $\Leftrightarrow \text{rank}(A) = m$

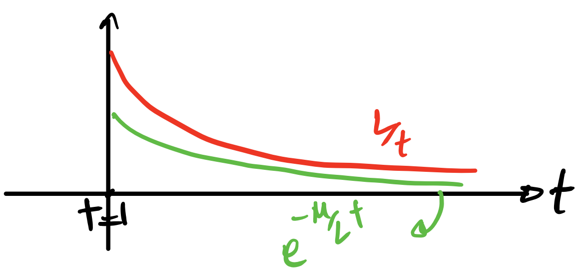
$\Rightarrow$ from HW 9 : $f(x)$ is strongly convex $\Rightarrow \mu > 0$ exist.

$\rightarrow$ By theorem μ-strongly convex functions from lecture

$\frac{L}{\mu}$ = condition number

for $\alpha = \frac{1}{L}$

$$f(x_t) - f(x^*) \leq \left(1 - \frac{\mu}{L}\right)^t (f(x_0) - f(x^*)) = O\left(e^{-\mu/L\, t}\right)$$



speed of convergence decreases if condition number increases.

Show that for a small Δx backtracking line search algorithm eventually terminates.

The goal is to reduce f enough along the ray $\{x + t\Delta x \mid t \geq 0\}$ by choosing the best t.

Algorithm:

choose $0 < \alpha < 0.5$ and $0 < \beta < 1$

given a descent direction Δx for f at x:

$t = 1$

while $f(x + t\Delta x) > f(x) + \alpha t \nabla f(x)^T \Delta x$

$t = \beta t$

This algorithm is called backtracking because it starts with $t = 1$ and then iteratively reduces t.

Stopping condition :

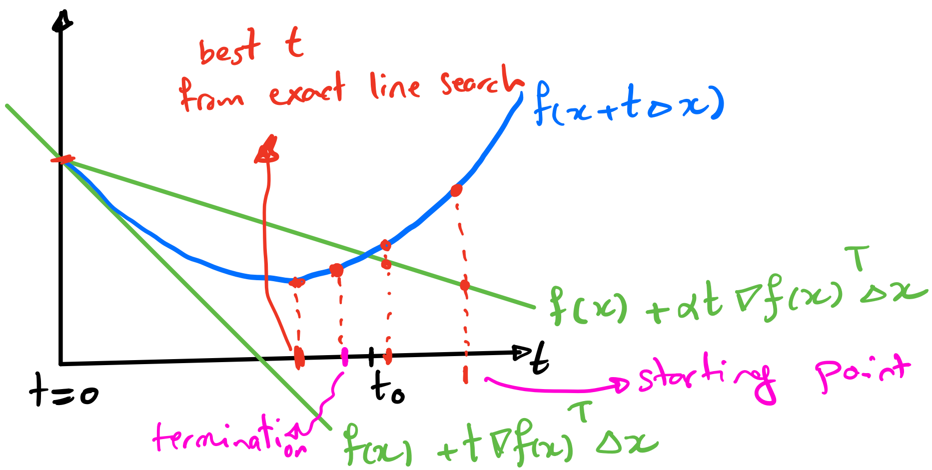
$$f(x + t\Delta x) \leq f(x) + \alpha t \nabla f(x)^T \Delta x$$

Since Δx is a descent direction: $\nabla f^T(x) \Delta x < 0$

So for small enough t :

$$f(x + t\Delta x) \approx f(x) + t \nabla f(x)^T \Delta x < f(x) + \alpha t \nabla f(x)^T \Delta x$$

which shows that the algorithm eventually terminates.



The backtracking condition is that f lies below the upper green line i.e. $0 \leq t \leq t_0$

It follows that the algorithm stops with a step length t that statisfies

$t = 1$ or $t \in (\beta t_0, t_0]$

if $1 \leq t_0$

we can say $\underbrace{t}_{\text{step size obtained from algorithm.}} \geq \min\{1, \beta t_0\}$

typically chosen: $0.01 \leq \alpha \leq 0.3$

$\underbrace{0.1}_{\text{a very crude search}} \leq \beta \leq \underbrace{0.8}_{\text{less crude search}}$

What is the update line if we use Newton Method to minimize the least-square cost $f(x) = ||Ax - y||^2$.

$$\nabla f(x_t) = 2A^T(Ax_t - y)$$

$$H_f(x_t) = 2\underset{m\times n}{A^T}\underset{n\times m}{A}$$

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

$$= x_t - (A^TA)^{-1}A^T(Ax_t - y)$$

what's the advantage & disadvantage?