**Math ProbSet 6**
**Nonlinear Optimization**
**Chris Rytting**
**Kendra Robbins**

### 9.1

Suppose $L$ is a linear objective function with a minimizer $x^*$.
Suppose further that $L$ is not constant.
Then there exists $y$ such that $Ly \neq Lx^*$.
Since $x^*$ is a minimizer we must have $Ly > Lx^*$.
Then $Lx^* - Ly = L(x^* - y) < 0$.
Consider $L(x^* + x^* - y)$:

$L(x^* + x^* - y) = Lx^* + L(x^* - y) < Lx^*$

$\Rightarrow x^*$ is not a minimizer

This is a contradiction, so given that L has a minimum, L must be a constant function.
If an unconstrained linear objective function M has no minimum, then M trivially cannot be a constant function, because the single value that a constant function takes on would be its minimum.
So an unconstrained linear objective function is either constant or has no minimum.

### 9.2

Suppose $x \in \mathbb{R}^m$ and $A \in M_{mxn}(\mathbb{R})$. Using that inner products over $\mathbb{R}$ are symmetric, specifically $< Ax, b >= x^T A^T b = b^T Ax =< b, Ax >$, we get:

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

$$= x^T A^T Ax - x^T A^T b - b^T Ax + 2b^T b = x^T A^T Ax - 2b^T Ax + 2b^T b$$

In minimizing this, we find that the FOC is: $2A^T Ax - 2A^T b = 0$, which is equivalent to $2A^T Ax = 2A^T b$

Additionally, in minimizing $x^T A^T Ax - 2b^T Ax$, FOC gives: $2A^T Ax - 2A^T b = 0$ which is equivalent to $2A^T Ax = 2A^T b$

Further, since $A^T A$ is positive definite, the second order-condition $2A^T A > 0$ is satisfied.

Since minimizing $||Ax - b||_2$ and $x^T A^T Ax - 2b^T Ax$ are both equivalent to solving $2A^T Ax = 2A^T b$, these two minimization problems are equivalent.

### 9.3

Gradient Decent:

- moves in the direction of greatest decrease, which is the negative of the gradient.

Different methods can be used to choose how far to move in the direction of greatest decrease.

- converges quickly for some problems and slowly for others
- converges linearly for quadratic problems with a special characteristic

Newton's Method:

- use when dimension is not too big and Hessian is positive definite and easy to compute
- finds roots of the problem
- uses the inverse of the Hessian and the gradient
- is a descent method and local approximation method
- converges quickly for nice quadratic functions
- may fail to converge when initial point is far from the solution
- may be very expensive for large problems
- not a good choice of optimization method if the Hessian is not positive definite or its inverse is expensive to compute

Quasi Newton's Methods:

- various methods have been created in efforts to implement the ideas in Newton's Method in ways that are computationally less expensive
- Broyden-Fletcher-Goldfarb-Shanno is one example

Conjugate Gradient:

- may take many steps to converge, but each step is inexpensive
- will optimize a quadratic of $n$ variables in $n$ steps
- moves toward a minimizer by moving along Q-conjugate directions
- good for solving large quadratic problems of a nice form

Nonlinear Least Squares:

- use when objective function can be expressed as the sum of squares of residuals

**9.4**
Let $f(x) = \frac{1}{2}x^T Qx - b^T x$ where $Q \in \mathcal{M}_n(\mathbb{R})$ such that $Q > 0$ and $b \in \mathbb{R}^n$.

Suppose the algorithm converges in one step.
Then we have $Qx_1 = b$ and $Q(x_0 - \alpha Q(Qx_0 - b)) = b$
Consider the kernel of $I - \alpha Q$.

$$\begin{aligned}(I - \alpha Q)(Qx_0 - b) &= Qx_0 - b - \alpha Q(Qx_0 - b) \\ &= Q(x_0 - \alpha Q(x_0 - b)) - b \\ &= Qx_1 - b \\ &= 0\end{aligned}$$

$\Rightarrow Qx_1 - b \in \text{Ker}(I - \alpha Q)$, and therefore it is an eigenvector of $Q$ with eigenvalue $\alpha$.

Suppose $x_0$ is chosen such that $Df(x_0)^T = Qx_0 - b$ is an eigenvector for $Q$.
So $Q(Qx_0 - b) = \lambda(Qx_0 - b)$ for some $\lambda \in \mathbb{R}$.
Define $x_1 = x_0 - \alpha Q(Qx_0 - b) = x_0 - \alpha\lambda(Qx_0 - b)$ and choose $\alpha$ such that $\alpha$ minimizes $f(x_1)$.
$\Rightarrow$ we are using the Method of Steepest Descent.
Consider $\alpha = \frac{1}{\lambda^2}$. Then:

$$\begin{aligned}Qx_1 &= Q(x_0 - \alpha\lambda(Qx_0 - b)) \\ &= Qx_0 - \alpha\lambda^2(Qx_0 - b) \\ &= Qx_0 - Qx_0 - b \\ &= b\end{aligned}$$

With $Qx_1 = b$, then $x_1$ is the minimum of the function, so $\alpha = \frac{1}{\lambda^2}$ is the correct $\alpha$, and the algorithm converges in one step.

**9.5** Thanks to my dear friend Matt for this crafty one.

*Proof.* I will begin by stating without proof a result of vector calculus.

**Fact:** The gradient of a function at a point $Df^T(x)$ is orthogonal to the level set of the function at the point $x$.

This fact gives some idea about where I'm going with this proof: first I'll show that I can reduce the proposition to the statement that the two gradients $Df^T(x_k)$ and $Df^T(x_{k+1})$ are orthogonal, and then I'll use the fact to show that this is indeed the case.

Consider $\langle x_{k+1} - x_k, x_{k+2} - x_{k+1}\rangle$.

$$\begin{aligned}\langle x_{k+1} - x_k, x_{k+2} - x_{k+1}\rangle &= \langle x_k - \alpha_{k+1}Df^T(x_k) - x_k, x_{k+1} - \alpha_{k+2}Df^T(x_{k+1}) - x_{k+1}\rangle \\ &= \langle -\alpha_{k+1}Df^T(x_k), -\alpha_{k+2}Df^T(x_k + 1)\rangle\end{aligned}$$

And if I want to set this equal to zero, I can pull out the scalars $-\alpha_{k+1}, -\alpha_{k+2}$ and set $\langle Df^T(x_k), Df^T(x_k+1)\rangle = 0$. So we see that

$$\langle x_{k+1} - x_k, x_{k+2} - x_{k+1}\rangle = 0 \iff \langle Df^T(x_k), Df^T(x_k+1)\rangle = 0$$

I'll now show that the gradients are orthogonal.

Consider the gradient $Df^T(x_k)$. $-Df^T(x_k)$ is the direction of steepest descent, and $x_{k+1} = x_k - \alpha Df^T(x_k)$. We choose $\alpha$ to minimize $f(x_{k+1})$. Consider the evaluation of the gradient $Df(x_k)$ at the point $x_{k+1}$.

**Claim:** $Df(x_k)(x_{k+1}) = 0$

*Proof of Claim:* This will be an intuitive argument which follows from the continuity of the derivative ($f$ is $C^1$). Suppose $-Df(x_k)(x_{k+1}) < 0$. Then, I can go a bit further along the descent to

$$x^* = x_k - (\alpha + \varepsilon)Df^T(x_k), \varepsilon > 0$$

such that $f(x^*) < f(x_{k+1})$. Similarly, if $-Df(x_k)(x_{k+1}) > 0$, then I can go a bit less far along the descent to

$$x^* = x_k - (\alpha - \varepsilon)Df^T(x_k), \varepsilon > 0$$

such that $f(x^*) < f(x_{k+1})$. So we see that $Df(x_k)(x_{k+1}) = 0$, which proves the claim.

Excellent. Now, $Df(x_k)(x_{k+1}) = 0$, so the gradient $Df^T(x_k)$ is tangent to the level set of $f$ at the point $x_{k+1}$. We know from our fact that $Df^T(x_{k+1})$ is orthogonal to the level set of $f$ at $x_{k+1}$, so it is orthogonal to $Df^T(x_k)$ as well, which concludes the proof.

□

**9.6-9.9** see my jupyter notebook

**9.10**

Suppose $f(x) = \frac{1}{2}x^TQx - b^Tx$ with $Q \in \mathcal{M}_n(\mathbb{R})$ symmetric and positive definite and $b \in \mathbb{R}^n$.
Let $x_0$ be an initial guess of the minimizer of $f$. Then:

$$\begin{aligned} x_1 &= x_0 - D^2f(x_0)^{-1}Df(x_0) \\ &= x_0 - Q^{-1}(Qx_0 - b) \\ &= x_0 - x_0 + Q^{-1}b = Q^{-1}b \end{aligned}$$

Further by the FOC, for a minimizer of $f$, $x^*$ we have: $f'(x) = Qx^* - b = 0$ so $Qx^* = b$ and $x^* = Q^{-1}b$.
$\Rightarrow x_1$ is a minimizer of $f$.

Since inverses are unique and any minimizer of $f$ has the form $Q^{-1}b$, the minimizer must be unique.
Thus for any initial guess $x_0$, one iteration of Newton gives us the unique minimizer of $f$.

**9.12**
Suppose $A \in \mathcal{M}_n(\mathbb{F})$ has eigenvalues $\lambda_1, ..., \lambda_n$ and $B = A + \mu I$.
Let $v_i$ be an eigenvector corresponding to $\lambda_i$. Then:
$Bv_i = (A + \mu I)v_i = Av_i + \mu I v_i = \lambda_i v_i + \mu v_i = (\lambda_i + \mu)v_i$
$\Rightarrow$ $v_i$ is an eigenvector of $B$ with eigenvalue $\lambda_i + \mu$.

**9.15**
$(A + BCD)(A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1})$
$= AA^{-1} - AA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} + BCDA^{-1} - BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$
$= I - B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} + BCDA^{-1} - BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$
$= I - B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} + BCDA^{-1} - BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$
$= I + BCDA^{-1} - (B(C^{-1} + DA^{-1}B)^{-1} + BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1})DA^{-1}$
$= I + BCDA^{-1} - ((B + BCDA^{-1}B)(C^{-1} + DA^{-1}B)^{-}1))DA^{-1}$
$= I + BCDA^{-1} - (BC(C^{-1} + DA^{-1}B)(C^{-1} + DA^{-1}B)^{-}1))DA^{-1}$
$= I + BCDA^{-1} - BCDA^{-1} = I$
$\Rightarrow$ since inverses are unique, $(A + BCD)^{-1} = A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}$

**9.18**
Let $Q \in \mathcal{M}_n(\mathbb{R})$ such that $Q > 0$ and define $f(x) = \frac{1}{2}x^TQx - b^Tx + c$.
To choose $\alpha_k$ to minimize $\phi_k(\alpha) = f(x_k + \alpha_k d_k)$, we need $\phi_k'(\alpha_k) = 0$.
Since $f$ is a quadratic we have:

$$\begin{aligned}\phi_k'(\alpha) &= -Df(x_k + \alpha_k d_k) \cdot d_k \\ &= [(x_k - \alpha_k d_k)^T Q - b^T]d_k \\ &= [x_k^T Q - b^T]d_k - (\alpha_k d_k)^T Q d_k = r_k^T d_k - \alpha_k(d_k^T Q d_k)\end{aligned}$$

$\Rightarrow \alpha_k = \frac{r_k^T d_k}{d_k^T Q d_k}$

**9.20** Thanks to Matt for this one

*Base Case:* $k = 1$. Recall that in my proof from Problem 9.5, I showed that $Df^T(x_k)$ was orthogonal to $Df^T(x_{k+1})$, where $x_{k+1} = x_k - \alpha_k Df^T(x_k)$. In general the conjugate gradient method constructs $x_{k+1}$ differently, so this theorem does not always apply. But in the first step, $r_0 = d_0 = -Df(x_0)^T$, so we see that:

$$\begin{aligned} x_1 &= x_0 + \alpha_0 d_0 = x_0 - \alpha_0 Df^T(x_0) \\ \implies r_1 &= Df(x_1)^T \perp Df(x_0)^T = r_0 \end{aligned}$$

which shows the base case.

*Inductive Case:* Assume that

$$r_i^T r_{k'} = 0 \text{ for all } i < k'$$

is true for any $k' < k$. I will show that the statement is also true for $k$. There is a bit of preliminary work necessary for my argument. Define the sets $D_{k-1} = \text{span}\{d_0, ..., d_{k-1}\}$ and $R_{k-1} = \text{span}\{r_0, ...r_{k-1}\}$. I'll state and justify a few facts about these sets.

**Fact 1**: $D_{k-1}$ and $R_{k-1}$ are both bases for subspaces of dimension $k-1$.
*Justification*: $R_{k-1}$ and $D_{k-1}$ are both orthogonal over some inner product space on $R^n$: $R_{k-1}$ by the inductive assumption, and $D_{k-1}$ by the property that it is $Q$-conjugate (and so orthogonal over the inner product space $\langle \cdot, \cdot \rangle_Q$. It is a theorem somewhere that orthogonal vectors are linearly independent, which shows the fact.

**Fact 2**: $D_{k-1} \subset R_{k-1}$
*Justification*: If $d \in D_{k+1}$, then it is a linear combination of elements $d_i, i \in \{0, 1, ..., k-1\}$. Therefore this fact will follow if I show that any element $d_i \in R_{k-1}$. And indeed,

$$\begin{aligned} d_i &= r_i - \beta_{i-1} d_{i-1} \\ &= r_i - \beta_{i-1}(r_{i-1} - \beta_{i-2} d_{i-2}) = r_i - \beta_{i-1} r_{i-1} + \beta_{i-1}\beta_{i-2} d_{i-2} \\ &= ... \\ &= \sum_{j=0}^{i} \left( \prod_{k=j}^{i-1} -\beta_k \right) r_j \end{aligned}$$

The actual final expression doesn't matter - what matters is that $d_i$ is expressed as a linear combination of the $r_i$s.

**Fact 3**: $D_{k+1} = R_{k+1}$
*Justification*: This follows from facts 1 and 2: Since the two spaces have the same dimension, one inclusion implies equality.

Now, we'll put this new knowledge to work and prove the inductive step. By Lemma 9.5.3, $d_i^T r_k = 0$ for any $i < j$. This means that $r_k \in D_{k-1}^{\perp} = R_{k-1}^{\perp}$ with the usual inner product, and therefore $r_i^T r_k = 0$ for all $i < k$, which was what we wanted to show.