Math Problem Set 6

Matthew Brown
OSM Boot Camp 2018

July 30, 2018

Problem 9.1. *Proof.* Let L be an unconstrained linear objective function. Suppose that L has a minimizer x^*. I'll show that L must be constant.

Suppose that L is not constant, i.e, there exists y such that $Ly \neq Lx^*$. If $Ly < Lx^*$, then x^* is not a minimizer and we have a contradiction. If $Ly > Lx^*$, then $L(x^* - y) < 0$ and we can consider the point $x^* + x^* - y$:

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

so x^* is not a minimizer and we have a contradiction □

Problem 9.2. *Proof.* Minimizing $||Ax - b||$ is equivalent to

$$\begin{aligned}(Ax-b)^T(Ax-b) &= (x^TA^T - b^T)(Ax-b)\\ &= x^TA^TAx - x^TA^Tb - b^TAx + 2b^Tb\\ &= x^TA^TAx - 2b^TAx + 2b^Tb\end{aligned}$$

Note that A^TA is positive semidefinite. Taking the FOC of this expression yields:

$$\begin{aligned}2x^TA^TA - 2b^TA &= 0\\ \iff x^TA^TA &= b^TA\\ \iff A^TAx &= A^Tb\end{aligned}$$

And because A^TA is positive definite, the second order-condition

$$2A^TA > 0$$

will always be satisfied. □

Problem 9.3.
Steepest Descent
This method computes the gradient at each point. We know that the negative of the gradient is the direction of steepest descent, so we then search for the distance α to travel along that gradient which makes the function as small as possible. Repeat until some stopping

criterion is fulfilled. Geometrically, the fact I prove in 9.5 shows that the path of steepest descent is orthogonal to the previous path of steepest descent. This method works only for differentiable functions.

Newton's Method
Start with original guess x_0, and minimize a degree-two taylor approximation of the function at the point x_0. The point at which the approximation attains its minimum is the new x, and we can iterate this process until a stopping criterion is reached. This method works only for functions with positive definite Hessians.

Conjugate Gradient
This is a modification of steepest descent so that subsequent iterations converge in a direction that is "Q - conjugate" to the previous descent directions. It only works for quadratic functions. Some geometric intuition that would come if we did a change of basis from the standard basis to the basis of "Q-conjugate" vectors that we use in constructing the problem - in this case, I *think* the descent vectors are orthogonal with the usual inner product!

Relative strengths and weaknesses:
The three methods are all good for different types of problem. The steepest descent method has the steps which take the smallest amount of computational power, but it takes a larger number of steps to converge on average. Conversely, Newton methods converge in fewer steps, but the steps are much more computationally expensive. The Conjugate gradient method provides a nice middle ground between the two - it takes fewer steps than the steepest descent, and the steps are less expensive than the newton method.

If the dimension is not too big, and especially if the function is differentiable we can use Newton's method. It's often a good idea to use steepest descent to get a better starting x_0 for Newton's method or when the function is not differentiable. If the dimensionality is very large, we are forced to resort to conjugate gradient.

Problem 9.4. *Proof.* I'll need to show both directions.

- $\Leftarrow$ Suppose x_0 is chosen such that $Df(x_0)^T = Qx_0 - b$ is an eigenvector for Q, i.e, $Q(Qx_0 - b) = \lambda(Qx_0 - b)$ for some $\lambda \in \mathbb{R}$. Recall the defintion of x_1 for quadratic forms:

  $$x_1 = x_0 - \alpha(Qx_0 - b)$$

  I choose α to minimize $f(x_1)$. If $\lambda = \frac{1}{\alpha}$, then

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

  as desired.

- $\Rightarrow$ Suppose the algorithm converges in one step. Then I know that $Qx_1 = b$, and thus that

$$Q(x_0 - \alpha(Qx_0 - b)) = Qx_0 - \alpha Q(Qx_0 - b) = b$$

  Rearranging, we get

$$Q(Qx_0 - b) = \frac{1}{\alpha}(Qx_0 - b)$$

  so $Qx_0 - b$ is an eigenvector with eigenvalue $\frac{1}{\alpha}$

□

Problem 9.5. *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 α 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

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

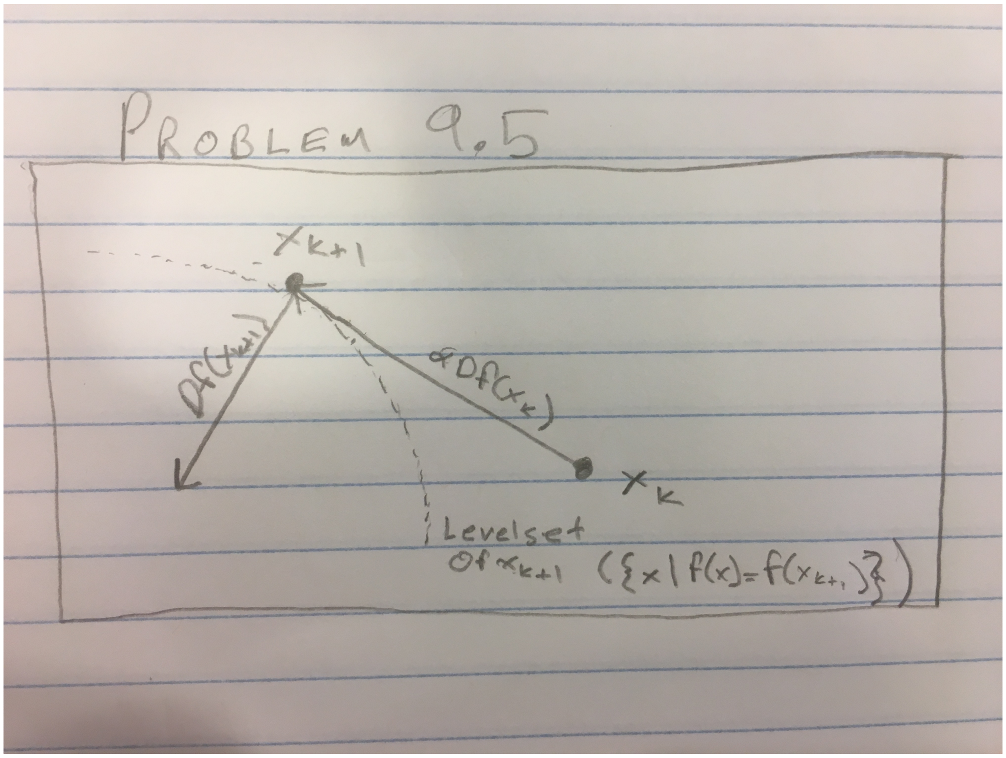

Figure 1: A Nice Picture for Problem 9.5

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.

See Figure 1 for some geometric intuition.

□

Problem 9.6. Jupyter

Problem 9.7. Jupyter

Problem 9.8. Jupyter (though incomplete - apologies)

Problem 9.9. Jupyter

Problem 9.10. *Proof.* We know that x^* is the unique minimizer of f iff

$$f'(x) = 0 \iff Qx^* - b = 0 \iff x^* = Q^{-1}b$$

Now let us start Newton's method from an arbitrary initial guess x_0. Calculate x_1:

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

which is what was desired. □

Problem 9.12. *Proof.* Choose λ_i arbitrarily, and let v_i be its eigenvector. Then

$$Bv_i = (A + \mu I)v_i = Av_i + \mu I v_i = \lambda_i v_i + \mu v_i = (\lambda_i + \mu)v_i$$

□

Problem 9.15. I'll multiply the left side by the right:

Proof.

$$\begin{aligned} &(A + BCD)(A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1}) \\ &= AA^{-1} - AA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} + BCDA^{-1} - BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \\ &= I - B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} + BCDA^{-1} - BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \\ &= I - B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} + BCDA^{-1} - BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \\ &= I + BCDA^{-1} - (B(C^{-1} + DA^{-1}B)^{-1} + BCDA^{-1}B(C^{-1} + DA^{-1}B)^{-1})DA^{-1} \\ &= I + BCDA^{-1} - ((B + BCDA^{-1}B)(C^{-1} + DA^{-1}B)^{-}1))DA^{-1} \\ &= I + BCDA^{-1} - (BC(C^{-1} + DA^{-1}B)(C^{-1} + DA^{-1}B)^{-}1))DA^{-1} \\ &= I + BCDA^{-1} - BCDA^{-1} = I \end{aligned}$$

□

Problem 9.16. Thanks Rebekah for the help here.
For Sherman-Morrison-Woodbury, let $A = A_k^{-1}, B = y - A_k s_k, C = ||s_k||^2, D = s_k^T$. Then

$$\begin{aligned} A_{k+1} = A + (BCD)^{-1} &= A^{-1} - A^{-1}B(C^{-1} + DA^{-1}B)^{-1}DA^{-1} \\ &= A_k^{-1} - A_k^{-1} \frac{(y_k - A_k s_k)s_k^T A_k^{-1}}{(||s_k||^2 + s_k^T A_k^{-1}(y_k - A_k s_k))} \\ &= A_k^{-1} + \frac{(s_k - A_k^{-1}y_k)s_k^T A_k^{-1}}{s_k^T A_k^{-1} y_k} \end{aligned}$$

Problem 9.18. *Proof.* I choose α_k to minimize the function $\phi_k(\alpha) = f(x_k + \alpha_k d_k)$, so I need $\phi_k'(\alpha_k) = 0$. Because f is a quadratic, I know that

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

And from this last line we see:

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

□

Problem 9.20. *Proof.* I will prove that $r_i^T r_k = 0$ for all $i < 0$ by induction on k.

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_is.

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. □

Note: There has to be a simpler way to do this proof. Apologies if this was very roundabout, but it's all I could think of!