AMATH 515 Homework 1

Cade Ballew

January 26, 2022

1 Problem 1

1.1 Part a

Consider $h(x) = g(a^T x)$ for vectors a, x. By the chain rule,

$$\nabla h(x) = a\nabla g(a^T x)$$

and

$$\nabla^2 h(x) = a\nabla^2 g(a^T x)a^T = \nabla^2 g(a^T x)aa^T$$

because g is a function from $\mathbb{R}$ to $\mathbb{R}$, so its gradient and Hessian are simply scalars.

1.2 Part b

Now, consider

$$r(x) = \left(\sum_{i=1}^{n} \log(1 + \exp(a_i^T x)) - b^T Ax\right) + \lambda\|x\|^2.$$

and take

$$g(x) = \sum_{i=1}^{n} \log(1 + \exp(x_i))$$

Then,

$$\frac{\partial g}{\partial x_i} = \frac{e^{x_i}}{1 + e^{x_i}}$$

so

$$\nabla g(x) = \begin{pmatrix} \frac{e^{x_1}}{1+e^{x_1}} \\ \vdots \\ \frac{e^{x_n}}{1+e^{x_n}} \end{pmatrix}.$$

Also,

$$\frac{\partial^2 g}{\partial x_i^2} = \frac{e^{x_i}}{(1 + e^{x_i})^2}$$

and

$$\frac{\partial^2 g}{\partial x_i \partial x_j} = 0,$$

so

$$\nabla^2 g(x) = \begin{pmatrix} \frac{e^{x_1}}{(1+e_1^x)^2} & & \\ & \ddots & \\ & & \frac{e^{x_n}}{(1+e_n^x)^2} \end{pmatrix}.$$

Now, we can apply the chain rule to r to conclude that

$$\nabla r(x) = A^T \nabla g(Ax) - A^T b + 2\lambda x = A^T \begin{pmatrix} \frac{e^{x_1}}{1+e_1^x} \\ \vdots \\ \frac{e^{x_n}}{1+e^{x_n}} \end{pmatrix} - A^T b + 2\lambda x$$

and

$$\nabla^2 r(x) = A^T \nabla^2 g(Ax) A + 2\lambda I = A^T \begin{pmatrix} \frac{e^{x_1}}{(1+e_1^x)^2} & & \\ & \ddots & \\ & & \frac{e^{x_n}}{(1+e_n^x)^2} \end{pmatrix} A + 2\lambda I.$$

1.3 Part c

Now, consider

$$r(x) = \left(\sum_{i=1}^{n} \exp(a_i^T x) - b^T A x \right) + \lambda \|x\|^2.$$

and take

$$g(x) = \sum_{i=1}^{n} e^{x_i}$$

Then,

$$\frac{\partial g}{\partial x_i} = e^{x_i}$$

so

$$\nabla g(x) = \begin{pmatrix} e^{x_1} \\ \vdots \\ e^{x_n} \end{pmatrix}.$$

Also,

$$\frac{\partial^2 g}{\partial x_i^2} = e^{x_i}$$

and

$$\frac{\partial^2 g}{\partial x_i \partial x_j} = 0,$$

so

$$\nabla^2 g(x) = \begin{pmatrix} e^{x_1} & & \\ & \ddots & \\ & & e^{x_n} \end{pmatrix}.$$

Now, we can apply the chain rule to p to conclude that

$$\nabla p(x) = A^T \nabla g(Ax) - A^T b + 2\lambda x = A^T \begin{pmatrix} e^{x_1} \\ \vdots \\ e^{x_n} \end{pmatrix} - A^T b + 2\lambda x$$

and

$$\nabla^2 p(x) = A^T \nabla^2 g(Ax) A + 2\lambda I = A^T \begin{pmatrix} e^{x_1} & & \\ & \ddots & \\ & & e^{x_n} \end{pmatrix} A + 2\lambda I.$$

1.4 Part d

Now, take

$$q(x) = \|Ax - b\| + \lambda \|x\|$$

and consider $m(x) = \|x\|$. Then,

$$\frac{\partial m}{\partial x_i} = \frac{x_i}{\sqrt{x_1^2 + \ldots + x_n^2}} = \frac{x_i}{\|x\|},$$

so

$$\nabla m(x) = \frac{x}{\|x\|}.$$

Also,

$$\frac{\partial^2 m}{\partial x_i^2} = \frac{1}{\|x\|} - \frac{x_i^2}{\|x\|^3}$$

and

$$\frac{\partial^2 m}{\partial x_i x_j} = -\frac{x_i x_j}{\|x\|^3},$$

so

$$\nabla^2 m(x) = \frac{1}{\|x\|}\left(I - \frac{xx^T}{\|x\|^2}\right).$$

Now, we note that the calculations are essentially the same for $\tilde{m}(x) = \|x - b\|$ due to the chain rule and conclude that

$$\nabla \tilde{m}(x) = \frac{x}{\|x - b\|}$$

and

$$\nabla^2 m(x) = \frac{1}{\|x - b\|}\left(I - \frac{xx^T}{\|x - b\|^2}\right).$$

Finally, we can apply the chain rule to q to conclude that

$$\nabla n(x) = A^T \nabla \tilde{m}(Ax) + \lambda \nabla m(x) = A^T \frac{Ax}{\|Ax - b\|} + \lambda \frac{x}{\|x\|} = \frac{A^T Ax}{\|Ax - b\|} + \lambda \frac{x}{\|x\|}$$

and

$$\begin{aligned}\nabla^2 n(x) &= A^T \nabla^2 \tilde{m}(Ax) A + \lambda \nabla^2 m(x) = A^T \frac{1}{\|Ax - b\|} \left(I - \frac{(Ax)(Ax)^T}{\|Ax - b\|^2} \right) A + \lambda \frac{1}{\|x\|} \left(I - \frac{xx^T}{\|x\|^2} \right) \\ &= \frac{A^T A}{\|Ax - b\|} - \frac{A^T Axx^T A^T A}{\|Ax - b\|^3} + \frac{\lambda}{\|x\|} \left(I - \frac{xx^T}{\|x\|^2} \right).\end{aligned}$$

In our definitions of the gradient and Hessian, we see both $\|Ax - b\|$ and $\|x\|$ present in the denominators, so we require that $Ax - b \neq 0$ and $x \neq 0$ to ensure that the gradient and Hessian exist at x.

2 Problem 2

2.1 Part a

To see that the function $\delta_C(x) = \begin{cases} 0 & \text{if} \quad x \in C \\ \infty & \text{if} \quad x \notin C. \end{cases}$ is convex, first consider the case where $x, y \in C$. Then, for any $\lambda \in [0, 1]$, $\lambda x + (1 - \lambda) y \in C$, so

$$\delta_C(\lambda x + (1 - \lambda) y) = 0 = \lambda \delta_C(x) + (1 - \lambda) \delta_C(y).$$

Now consider the case where at least one of x, y is not in C. Then,

$$\lambda \delta_C(x) + (1 - \lambda) \delta_C(y) = \infty,$$

so

$$\delta_C(\lambda x + (1 - \lambda) y) \leq \lambda \delta_C(x) + (1 - \lambda) \delta_C(y).$$

automatically since the maximum value $\delta_C(\lambda x + (1 - \lambda) y)$ can take is ∞. Thus, in all cases,

$$\delta_C(\lambda x + (1 - \lambda) y) \leq \lambda \delta_C(x) + (1 - \lambda) \delta_C(y),$$

so δ_C is convex.

2.2 Part b

If we define

$$\sigma_C(x) = \sup_{c \in C} c^T x,$$

then for any x, y and $\lambda \in [0, 1]$,

$$\begin{aligned}\sigma_C(\lambda x + (1 - \lambda) y) &= \sup_{c \in C} c^T (\lambda x + (1 - \lambda) y) \leq \sup_{c \in C} c^T (\lambda x) + \sup_{c \in C} c^T ((1 - \lambda) y) \\ &= \lambda \sup_{c \in C} c^T x + (1 - \lambda) \sup_{c \in C} c^T y = \lambda \sigma_C(x) + (1 - \lambda) \sigma_C(y),\end{aligned}$$

so σ_C is convex.

2.3 Part c

For some arbitrary norm, consider x, y to be arbitrary arguments that the norm can take and let $\lambda \in [0,1]$. Then, by the definition of a norm (first using the triangle inequality then absolute homogeneity),

$$\|\lambda x + (1-\lambda)y\| \leq \|\lambda x\| + \|(1-\lambda)y\| = |\lambda|\|x\| + |1-\lambda|\|y\| = \lambda\|x\| + (1-\lambda)\|y\|.$$

Thus, an arbitrary norm is convex.

3 Problem 3

3.1 Part a

To see that the composition of two convex functions is not necessarily convex, consider $f(x) = -x$ and $g(x) = x^2$. Clearly, both of these functions are convex, because their second derivatives (0 and 2, respectively) are nonnegative everywhere. However, $h(x) = f(g(x)) = -x^2$ is not convex. One can see this by taking $x = 0$, $y = 2$, $\lambda = 1/2$. Then,

$$h(\lambda x + (1-\lambda)y) = h(1) = -1 > -2 = \lambda h(x) + (1-\lambda)h(y).$$

A sufficient condition for this composition h to be convex is for f to also be nondecreasing. To see this, consider any $x, y \in \mathbb{R}$ and $\lambda \in [0,1]$. Then, by the convexity of g,

$$g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y).$$

Now, the fact that g is nondecreasing combined with this gives that

$$f(g(\lambda x + (1-\lambda)y)) \leq f(\lambda g(x) + (1-\lambda)g(y)).$$

Finally, we note that $g(x), g(y) \in \mathbb{R}$ and use the convexity of f to find that

$$f(\lambda g(x) + (1-\lambda)g(y)) \leq \lambda f(g(x)) + (1-\lambda)f(g(y))$$

and conclude that

$$h(\lambda x+(1-\lambda)y) = f(g(\lambda x+(1-\lambda)y)) \leq \lambda f(g(x))+(1-\lambda)f(g(y)) = \lambda h(x)+(1-\lambda)h(y),$$

so h must be convex.

3.2 Part b

Now, if f is convex and g is concave, the assumption that f is nonincreasing gives that $h = f \circ g$ is convex. The proof proceeds in largely the same way, except now we have that

$$g(\lambda x + (1-\lambda)y) \geq \lambda g(x) + (1-\lambda)g(y),$$

but the fact that f is now nonincreasing still gives that

$$f(g(\lambda x + (1-\lambda)y)) \leq f(\lambda g(x) + (1-\lambda)g(y)).$$

The rest of the proof proceeds unchanged (as f is still convex), and we conclude that

$$h(\lambda x+(1-\lambda)y) = f(g(\lambda x+(1-\lambda)y)) \leq \lambda f(g(x))+(1-\lambda)f(g(y)) = \lambda h(x)+(1-\lambda)h(y),$$

so h must be convex.

3.3 Part c

If $f : \mathbb{R}^m \to \mathbb{R}$ is convex and $g : \mathbb{R}^n \to \mathbb{R}^m$ affine, then g can be represented by $g(x) = Ax + b$, so for $x, y \in \mathbb{R}^n$ and $\lambda \in [0, 1]$,

$$\begin{aligned} h(\lambda x + (1-\lambda)y) &= f(g(\lambda x + (1-\lambda)y)) = f(A(\lambda x + (1-\lambda)y) + b) = f(\lambda Ax + (1-\lambda)Ay + b) \\ &= f(\lambda(Ax + b) + (1-\lambda)(Ax + b)). \end{aligned}$$

Now, we simply apply the convexity of f to conclude that

$$\begin{aligned} h(\lambda + (1-\lambda)y) &\leq \lambda f(Ax + b) + (1-\lambda)f(Ax + b) = \lambda f(g(x)) + (1-\lambda)f(g(y)) \\ &= \lambda h(x) + (1-\lambda)h(y), \end{aligned}$$

so h must be convex.

3.4 Part d

3.4.1 Part i

To see that the function $\sum_{i=1}^{n} \log(1+\exp(a_i^T x)) - b^T Ax$ is convex, we note that if we define

$$g(x) = \sum_{i=1}^{n} \log(1 + \exp(x_i)) + b^T x,$$

then this is simply $g(Ax)$ (Note that this definition does require that A be square, but this definition is primarily for ease of argument; if A is not square then we can drop the $b^T x$ term from the definition of g and simply observe that it will not impact convexity because it is a linear term). However, because the second term is linear, we have already computed the Hessian of g in problem 1 part b as

$$\nabla^2 g(x) = \begin{pmatrix} \frac{e^{x1}}{(1+e_1^x)^2} & & \\ & \ddots & \\ & & \frac{e^{xn}}{(1+e_n^x)^2} \end{pmatrix}.$$

It is easy to see that this matrix is positive definite; because our matrix is diagonal, we can look at the diagonal elements which are the eigenvalues and simply need to confirm that they are positive. This is clearly true, because the exponential is always positive if its arguments are real, so both the numerator and denominator of each diagonal entry must be positive. Thus, we can conclude that g is convex since its Hessian is positive (semi)-definite. Now, we note that $x \mapsto Ax$ is an affine transform, so we simply apply part c to conclude that $g(Ax)$, our original function, is convex.

3.4.2 Part ii

Similarly, we can see that the function $\sum_{i=1}^{n} \exp(a_i^T x) - b^T Ax$ is convex by defining

$$g(x) = \sum_{i=1}^{n} \exp(x_i) + b^T x$$

and noting that since the linear term makes no contribution, we have the Hessian from problem 1 part c as

$$\nabla^2 g(x) = \begin{pmatrix} e^{x_1} & & \\ & \ddots & \\ & & e^{x_n} \end{pmatrix}.$$

Again, we can see that this matrix is positive definite, as its diagonal elements are exponential and therefore positive. Thus, we know that $g(x)$ is convex, meaning that we can again apply part c to conclude that $g(Ax)$, our original function, is convex.

4 Problem 4

4.1 Part a

Consider the function $f(x) = e^x$ where $x \in \mathbb{R}$. To see that this function is strictly convex, note that $f''(x) = e^x > 0$ for all $x \in \mathbb{R}$ (note that we look at the 2nd derivative instead of the Hessian since we have a single-variable function). However, e^x is strictly increasing and $e^x \to 0$ as $x \to -\infty$, but $e^x \neq 0$ for any $x \in \mathbb{R}$, so f does not have a minimizer.

4.2 Part b

To see that a sum of a strictly convex function and a convex function is strictly convex, let $h = f + g$ where f, g are convex functions on some domain. Let x, y be in the domain of g and let $\lambda \in (0, 1)$. Then,

$$\begin{aligned} h(\lambda x + (1-\lambda)y) &= f(\lambda x + (1-\lambda)y) + g(\lambda x + (1-\lambda)y) \\ &< (\lambda f(x) + (1-\lambda)f(y)) + (\lambda g(x) + (1-\lambda)g(y)) \\ &= \lambda(f(x) + g(x)) + (1-\lambda)(f(x) + g(x)) = \lambda h(x) + (1-\lambda)h(y), \end{aligned}$$

so h is strictly convex.

4.3 Part c

We know from class that the solution(s) of

$$\min_x \frac{1}{2}\|Ax - b\|^2$$

are given by the solutions of

$$(A^T A)x = A^T b.$$

Because the range of $A^T A$ is always equal to the range of A^T, this system is always consistent, so we cannot have no solutions. However, this system has one solution if $A^T A$ is nonsingular and infinitely many solutions if $A^T A$ is singular. If the columns of A are linearly independent, then $A^T A$ is nonsingular, and the minimization problem must have a unique solution.

5 Problem 5

5.1 Part a

The function

$$f(x) = \frac{1}{2}\|Ax - b\|^2$$

has gradient

$$\nabla f(x) = A^T(Ax - b),$$

so we can use the properties of an operator norm to write

$$\|\nabla(x) - \nabla(y)\| = \|A^T(Ax-b) - A^T(Ay-b)\| = \|A^T A(x-y)\| \leq \|A^T A\|_2 \|x-y\|$$

for x, y in the domain of f. This means that if we take $\beta \geq \|A^T A\|_2$, then our function is β-smooth.

5.2 Part b

Now, we use the theorem from class that a function f is β-smooth iff

$$\lambda_{\max}(\nabla^2 f(x)) \leq \beta$$

for all x. From problem 1 part b, we know that the Hessian of

$$r(x) = \sum_{i=1}^{n} \log(1 + \exp(\langle a_i, x\rangle)) + \frac{\lambda}{2}\|x\|^2$$

is given by

$$\nabla^2 r(x) = A^T \begin{pmatrix} \frac{e^{x_1}}{(1+e^{x_1})^2} & & \\ & \ddots & \\ & & \frac{e^{x_n}}{(1+e^{x_n})^2} \end{pmatrix} A + \lambda I.$$

Since the center matrix is diagonal and $A^T A$ is positive semi-definite, we can bound the eigenvalues of $\nabla^2 r(x)$ by maximizing the diagonal entries (eigenvalues) of this matrix. The derivative of an entry is given by

$$\frac{e_i^x(1 - e^{x_i})}{(1 + e^{x_i})^3}$$

which we set equal to zero and find $x_i = 0$. We can plot this function and find that this does indeed maximize it with a value of 1/4. Writing out this bound,

$$\lambda_{\max}(\nabla^2 r(x)) \leq \lambda_{\max}(A^T((1/4)I)A + \lambda I) = \frac{1}{4}\lambda_{\max}(A^T A) + \lambda = \frac{1}{4}\|A^T A\|_2^2 + \lambda.$$

Thus, if we take $\beta \geq \frac{1}{4}\|A^T A\|_2^2 + \lambda$, then our function is β-smooth.

5.3 Part c

For the Poisson regression, have from problem 1 part c that its Hessian is given by

$$\nabla^2 p(x) = A^T \nabla^2 g(Ax)A + 2\lambda I = A^T \begin{pmatrix} e^{x_1} & & \\ & \ddots & \\ & & e^{x_n} \end{pmatrix} A + \lambda I.$$

However, we can immediately see that bounding the eigenvalues here will not be possible due to the diagonal matrix present. Because its entries are exponential, they are unbounded as $x_i \to \infty$. Thus, we can not place a bound on them and cannot place a bound on the eigenvalues of the Hessian as a result, meaning that the gradients for Poisson regression do not admit a global Lipschitz constant.