AMATH 515 Problem Set 1

Tyler Chen

Problem 1

Explain why each of the following functions is convex.

(a) Indicator function to a convex set: $\delta_C(x) = \begin{cases} 0 & \text{if} \quad x \in C \\ \infty & \text{if} \quad x \notin C. \end{cases}$

(b) Support function to any set: $\sigma_C(x) = \sup_{c\in C} c^T x$.

(c) Perspective function: $f(x,t) = tg(x/t)$, where g is a convex function, $x \in \mathbb{R}^n$, and $t > 0$ is a positive scalar.

Solution

(a) Pick any x, y. Then either $x, y \in C$ or WLOG $x \notin C$. If $x, y \in C$ then by the definition of convexity, $\lambda x + (1-\lambda)y \in C$ for all $\lambda \in [0,1]$. Therefore,

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

Otherwise, $\delta_C(x) = \infty$ so that,

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

(b) Pick any x, y and $\lambda \in [0,1]$. Then,

$$\begin{aligned} \sigma_C(\lambda x + (1-\lambda)y) &= \sup_{c\in C}\left(\lambda c^T x + (1-\lambda)c^T y\right) \\ &\leq \lambda \sup_{c_1\in C} c_1^T x + (1-\lambda)\sup_{c_2 \in C} c_2^T y \\ &= \lambda\sigma_C(x) + (1-\lambda)\sigma_C(y) \end{aligned}$$

(c) Intuitively the graph of $g(x)$ expands linearly in time so clearly the epigraph is a convex set.

Given two points and a point between them, we find the two points on the rays from the origin through the original two points such that the time value is the same as the middle point. We then write the middle point as a convex combination of these new points, and use the fact that for a fixed time $tg(x/t)$ is trivially convex.

To this end, fix two points (x,s) and (y,t) and $\lambda \in (0,1)$. For convenience define,

$$T = \lambda s + (1-\lambda)t$$

Define the points,

$$M = (\lambda x + (1-\lambda)y, T), \qquad X = \left(\frac{x}{s}T, T\right), \qquad Y = \left(\frac{y}{t}T, T\right)$$

Note that all these points along with T depend on λ and that M is the convex combination of (x, s) and (y, t) with parameter λ while X and Y lie on the rays passing from the origin through (x, s) and (y, t) respectively.

Note further that with $\mu = \lambda s/T$ we have,

$$M = \mu X + (1-\mu)Y = \left(\mu\left(T\frac{x}{s}\right) + (1-\mu)\left(T\frac{y}{t}\right), T\right)$$

Using the above expression for M and the definition of f we have that,

$$\begin{aligned} f(M) &= f(\mu(Tx/s) + (1-\mu)(Ty/t), T) \\ &= Tg\left(\frac{\mu(Tx/s) + (1-\mu)(Ty/t)}{T}\right) \\ &= Tg\left(\mu(x/s) + (1-\mu)(y/t)\right) \end{aligned}$$

By the convexity of g we have that,

$$Tg\left(\mu(x/s) + (1-\mu)(y/t)\right) \leq \mu Tg(x/s) + (1-\mu)Tg(y/t)$$

Now note that,

$$\mu Tg(x/s) = \lambda sg(x/s) = \lambda f(x, s)$$

and

$$(1-\mu)Tg(y/t) = (1-\lambda)tg(y/t) = (1-\lambda)f(y, t)$$

Therefore, using these observations and the definition of M we have that,

$$f(\lambda x + (1-\lambda)y, \lambda s + (1-\lambda t)) \leq \lambda f(x, s) + (1-\lambda)f(s, t)$$

This proves that f is convex. □

Problem 2

Convexity and composition rules. Suppose that f and g are $\mathcal{C}^2$ functions from $\mathbb{R}$ to $\mathbb{R}$, with $h = f \circ g$ their composition, defined by $h(x) = f(g(x))$.

(a) If f and g are convex, is this enough to ensure h is convex? If not, give a counter example, along with any additional hypotheses.

(b) If f is convex and g is concave, can h be convex? Is there an additional hypothesis that guarantees h is convex?

(c) Show that if $f : \mathbb{R}^m \to \mathbb{R}$ is convex and $g : \mathbb{R}^n \to \mathbb{R}^m$ affine, then h is convex.

(d) Show that the following functions are convex:

(i) Logistic regression objective: $\sum_{i=1}^n \log(1 + \exp(\langle a_i, x\rangle)) - b^T Ax$

(ii) Poisson regression objective: $\sum_{i=1}^n \exp(\langle a_i, x\rangle) - b^T Ax$.

Solution

(a) No. Let $f(x) = -x$ and $g(x) = x^2$. Then $f(g(x)) = -x^2$ which is clearly not convex.

Suppose f is non-decreasing. Then since g is convex we have $g(\lambda x + (1-\lambda)y) \leq \lambda g(x) + (1-\lambda)g(y)$. Therefore, since f is non-decreasing and convex,

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

This proves $f \circ g$ is convex. □

(b) Yes. Let $f(x) = g(x) = x$ so that $f(g(x)) = x$. Then f is convex, g is concave, and $f(g(x))$ is convex.

Suppose f is non-increasing. Then since g is concave we have $g(\lambda x + (1-\lambda)y) \geq \lambda g(x) + (1-\lambda)g(y)$. Therefore, since f is non-increasing and convex,

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

This proves $f \circ g$ is convex. □

(c) Since g is affine we can write $g(x) = Ax + b$ for some $A \in \mathbb{R}^{m\times n}$ and $b \in \mathbb{R}^n$. Observe then that,

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

Thus,

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

This proves $f \circ g$ is convex. □

(d) Suppose $f(x) = G(Ax)$ where $G(z) = \sum_i g(z_i)$.

I claim that if f is convex if g is convex ($g'' \geq 0$). To this end, observe,

$$\begin{aligned} G(\lambda u + (1-\lambda)v) &= \sum_i g(\lambda u_i + (1-\lambda)v_i) \\ &\leq \sum_i \lambda g(u_i) + (1-\lambda)g(v_i) \\ &= \lambda G(u) + (1-\lambda)G(v) \end{aligned}$$

Therefore $G : \mathbb{R}^m \to \mathbb{R}$ is convex so by (c) $f(x)$ is convex.

(i) Let $g(t) = \log(1+\exp(t)) - b_i t$. Then,

$$g'(t) = \frac{\exp(t)}{1+\exp(t)} - b_i, \qquad g''(t) = \frac{\exp(t)}{(1+\exp(t))^2}$$

Note that $g''(t) > 0$ for all t so $f(x) = G(Ax)$ is convex.

(ii) Let $g(t) = \exp(t) - b_i t$. Then,

$$g'(t) = \exp(t) - b_i, \qquad g''(t) = \exp(t)$$

Note that $g''(t) > 0$ for all t so $f(x) = G(Ax)$ is convex.

Problem 3

A function f is *strictly convex* if

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y), \quad \lambda \in (0,1).$$

(a) Give an example of a strictly convex function that does not have a minimizer.

(b) Show that a sum of a strictly convex function and a convex function is strictly convex.

(c) What conditions (if any) are necessary (sufficient???) to ensure that the following problems have a unique minimizer?

(i) Least squares: $\min_x \frac{1}{2}\|Ax-b\|^2$

(ii) Logistic: $\min_x \sum_{i=1}^n \log(1+\exp(\langle a_i, x\rangle)) - b^TAx$

(iii) Elastic net logistic:

$$\min_x \sum_{i=1}^n \log(1+\exp(\langle a_i, x\rangle)) + \lambda(\alpha\|x\|_1 + (1-\alpha)\|x\|^2), \quad \lambda > 0, \alpha \in (0,1)$$

Solution

(a) Let $f : (-1,0)$ be defined as $f(x) = x^2$. Since $\lambda, (1-\lambda) \in (0,1)$ then $\lambda^2 < \lambda$ and $(1-\lambda)^2 < (1-\lambda)$. Therefore,

$$\begin{aligned} f(\lambda x + (1-\lambda)y) &= \lambda^2 x^2 + (1-\lambda)^2 y^2 \\ &< \lambda x^2 + (1-\lambda)y^2 \\ &= \lambda f(x) + (1-\lambda)f(y) \end{aligned}$$

This proves f is strictly convex.

However, for every $x \in (-1,0)$, $x/2 \in (-1,0)$ and $f(x) > f(x/2)$. This proves f has no minimizer. □

(b) Suppose f is strictly convex and g is convex. Then, for all $\lambda \in (0,1)$,

$$f(\lambda x + (1-\lambda)y) < \lambda f(x) + (1-\lambda)f(y)$$

and

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

Thus,

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

This proves $f+g$ is strictly convex. □

(c) We assume all minimization is done over $\mathbb{R}^m$. We also note that it's not really clear if the problem is asking for necessary or sufficient conditions so we kind of just discuss the problems in general.

(i) A minimizer always exists since the function is continuous and goes to infinity as $\|x\| \to \infty$, but is bounded below by zero.

Since for any $z \in \ker(A)$ we have $A(x+z) = Ax$ we need that A is injective for the minimizer to be unique.

(ii) Note that the scalar function $t \mapsto \log(1+\exp(t))$ has no minimizer on $\mathbb{R}$. Intuitively logistic regression can then fail to have a minimizer if we can find x so that we have negative values in the exponentials and the b^TAx term does not become negative.

In particular, suppose that we can find x so that each $a_i^Tx \leq 0$ and that $b^TAx \leq 0$. In this case we can pick larger and larger positive scalar multiples of x and drive down both the sum and the b^TAx terms.

More rigorously, note that we have gradient $A^T\nabla G(Ax) - A^Tb$ where $\nabla G(Ax)$ has entries $\exp(a_i^Tx)/(1+\exp(a_i^Tx))$.

A necessary condition for a minimizer is that the gradient is zero. That is,

$$A^T\left(\nabla G(Ax) - b\right) = 0$$

(iii) We note that this function is strictly convex and so if there is a minimizer it is unique. Moreover, since the function is continuous and goes to infinity as $\|x\|$ goes to infinity, but is bounded below by zero, a minimizer must exist.

To see the function is strictly convex, first note that $\lambda(\alpha\|x\|_1 + (1-\alpha)\|x\|^2)$ is convex since all norms are convex. Next, recall that we have previously shown that the scalar function $t \mapsto \log(1+\exp(t))$ is strictly convex by observing that the second derivative is strictly positive and that this implied that the sum term is strictly convex.

Finally, by (b) we have that the sum of a strictly convex function and a convex function is strictly convex.

Problem 4

Lipschitz constants.

(a) Find a global bound for β of the least-squares objective $\frac{1}{2}\|Ax-b\|^2$.

(b) Find a global bound for β of the regularized logistic objective

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

(c) Do the gradients for Poisson regression admit a global Lipschitz constant?

Solution

(a) Recall that,

$$\nabla\left(\frac{1}{2}\|Ax-b\|^2\right)=A^TAx-A^Tb$$

and observe that A^TAx-A^Tb is continuous.

Moreover, for any x,y,

$$\left\|(A^TAx-A^Tb)-(A^TAy-A^Tb)\right\|=\left\|A^TA(x-y)\right\|\leq\left\|A^TA\right\|\|x-y\|$$

Therefore the least-squares objective $\frac{1}{2}\|Ax-b\|^2$ is β-smooth provided

$$\beta\geq\left\|A^TA\right\|=\lambda_{\max}(A)=\sigma_{\max}(A)^2$$

(b) Let $f(x)=G(Ax)+\frac{\lambda}{2}\|x\|^2$ where $G(z)=\sum_i g(z_i)$ and $g(z_i)=\log(1+\exp(z_i))$.

Then,

$$\nabla f(x)=A^T\nabla G(Ax)+\lambda x$$

Now observe that,

$$(\nabla G(z))_i=g'(z_i)=\frac{\exp(z_i)}{1+\exp(z_i)}\in(0,1)$$

Let $h(t)=e^t/(1+e^t)$ so that $h'(t)=e^t/(1+e^t)^2$. This is maximal at $t=0$ with value $h'(0)=1/4$. Therefore $h(t)$ is Lipshitz with constant $1/4$.

Therefore,

$$\|\nabla G(u)-\nabla G(v)\|^2=\sum_{i=1}^{n}((\nabla G(u))_i+(\nabla G(v))_i)^2\leq\sum_{i=1}^{n}\frac{1}{4}(u_i-v_i)^2=\frac{1}{4}\|u-v\|^2$$

Therefore, $\nabla G : \mathbb{R}^n \to \mathbb{R}^n$ is Lipschitz with constant $1/2$. Therefore,

$$\|\nabla G(Ax) - \nabla G(Ay)\| \leq \sqrt{n}\,\|Ax - Ay\| \leq \frac{1}{2}\,\|A\|\,\|x-y\|$$

Then, by the triangle inequality

$$\begin{aligned}
\|\nabla f(x) - \nabla f(y)\| &= \left\|A^T\nabla G(Ax) + \lambda x - (A^T\nabla G(Ay) + \lambda y)\right\| \\
&= \left\|A^T(\nabla G(Ax) - \nabla G(Ay)) + \lambda(x-y)\right\| \\
&\leq \frac{1}{2}\left\|A^T\right\|\|\nabla G(Ax) - \nabla G(Ay)\| + \lambda\,\|x-y\| \\
&\leq \left(\frac{1}{2}\left\|A^T\right\|\|A\| + \lambda\right)\|x-y\|
\end{aligned}$$

This proves f is β-smooth for any,

$$\beta \geq \frac{1}{2}\left\|A^T\right\|\|A\| + \lambda = \frac{1}{2}\sigma_{\max}(A)^2 + \lambda$$

(c) We note that the Hessian is of the form $A^T H_f A$ where H_f has $\exp(a_i^T x)$ on the diagonal.

The Hessian is not bounded in norm as $\exp(a_i^T x)$ is unbounded over $\mathbb{R}^m$ (provided $H_f \neq 0$). As a result, Poisson regression will not admit a global quadratic upper bound (see this by Taylor expansion). Therefore the function is not β-smooth.

Problem 5

Behavior of steepest descent for logistic vs. poisson regression.

(a) Given the sample (logistic) data set and starter code, implement gradient descent for ℓ_2-regularized logistic regression. Plot (a) the objective value and (b) the norm of the gradient (as a measure of optimality) on two separate figures. For the figure in (b), make sure the y-axis is on a logarithmic scale.

(b) Implement Newton's method for the same problem. Does the method converge? If necessary, use the line search routine provided to scale your updated directly to ensure descent. Add the plots for Newton's method (a) and (b) to your Figures 1 and 2. What do you notice?

(c) Using the sample (Poisson) data and starter code provided, implement gradient descent and Newton's method for ℓ_2-regularized Poisson regression. You may need to use the line search routine for both algorithms. Make the same plots as you did for the logistic regression examples.

(d) What do you notice qualitatively about steepest descent vs. Newton?

Solution

Suppose $f(x) = G(Ax) - b^T Ax + \frac{\lambda}{2}\left\|x\right\|^2$ where $G(z) = \sum_i g(z_i)$.

Then,

$$\nabla f(x) = A^T \nabla G(Ax) - A^T b + \lambda x$$

where $\nabla G(Ax) = [g'(a_1^T x), \ldots, g'(a_n^T x)]$.

Therefore,

$$H_f = A^T H_G(Ax) A - \lambda I$$

where $H_G(Ax) = \operatorname{diag}([g''(a_1^T x), \ldots, g''(a_n^T x)])$.

(a) Figure 1 shows the objective value (left) and norm of the gradient (right) vs iteration for Gradient Descent applied to ℓ_2-regularized logistic regression.

(b) Figure 2 shows the objective value (left) and norm of the gradient (right) vs iteration for Newton's method applied to ℓ_2-regularized logistic regression.

(c) Figures 3 and 4 respectively show the objective value (left) and norm of the gradient (right) vs iteration for Gradient Descent and Newton's method applied to ℓ_2-regularized Poisson regression.

(d) The Newton solvers converge in many fewer steps. On the semilog plot of the gradient it is also clear that Newton's method begins to converge more quickly once the norm of the gradient is sufficiently small.

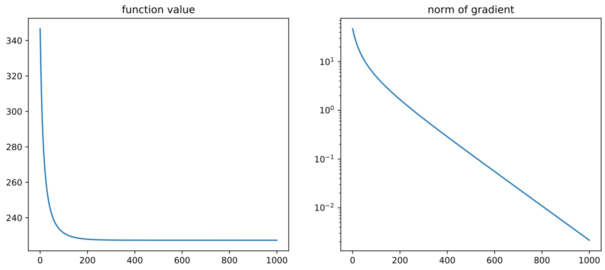

Figure 1: Gradient Descent on Logistic Regression

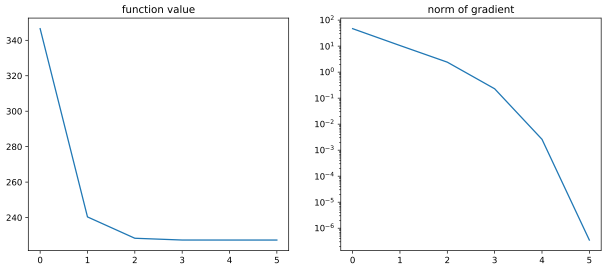

Figure 2: Newton's Method on Logistic Regression

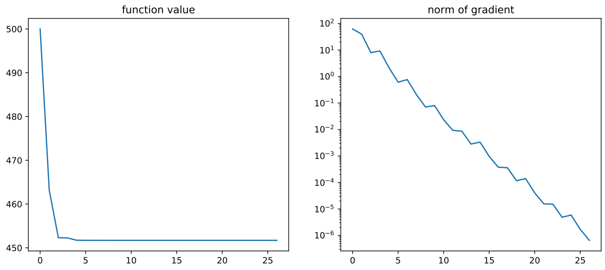

Figure 3: Gradient Descent on Poisson Regression

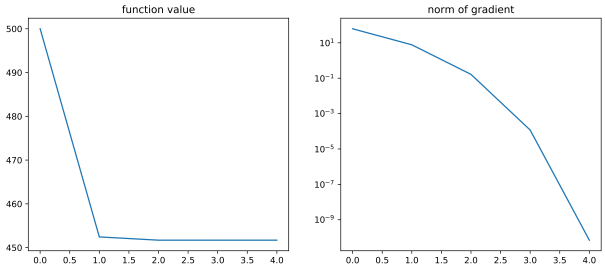

Figure 4: Newton's Method on Poisson Regression

Bonus

Consider the poisson model

$$P(b_i|\lambda_i) = \frac{1}{b_i!}\exp(-\lambda_i + b_i\ln(\lambda_i)).$$

Develop an inference model for x by using the assumption $\lambda_i = a_i^T x$. Write down an optimization problem, and solve it using a method of your choice. The ln is only defined for positive inputs, so you want to make sure you start and stay feasible, i.e. that you never let x be such that any $a_i^T x$ fall below 0.

Test your ability to predict outcomes using the standard Poisson model vs. your new model. One way to do this is to divide your dataset into 'training' and 'testing', solve for x using the training dataset, and evaluate how far you are from the observed b_i on the test set. Include a table or plot that meaningfully compares the standard Poisson model vs. the one you developed here. Describe any implementation issues you encountered and what you learned. Do the patterns you observed hold over different realizations of sampled data?

Solution