Machine Learning for Economic Analysis
Problem Set 2

Jonas Lieber*

Due: 11:59pm Wed, Feb 7, 2023

Problem 1. *Gradient Descent*

1. *Write a function that has the folowing inputs*
   (a) *a function* $f : \mathbb{R}^p \to \mathbb{R}$ *to minimize,*
   (b) *a starting point* $x^0 \in \mathbb{R}^p$,
   (c) *the gradient* $\nabla f : \mathbb{R}^p \to \mathbb{R}^p$ *of* f, *defined as* $(\nabla f(x))_j = \frac{\partial}{\partial x_j} f(x)$,
   (d) *a function* s *that depends on* f, *the gradient evaluated at the current data point* $\nabla f(x^k)$, *the iteration index* k, *the current point* x^k *and a tuning parameter* $\gamma \in (0, 1)$ *and returns a stepsize between 0 and 1,*
   (e) *a* $\varepsilon > 0$ *that will be used for the stopping criterion and does the following iteration*
   while($\left\|\nabla f(x^k)\right\|_2 > \varepsilon$)
   $sz = s(f(\cdot), \nabla f(x^k), k, x^k, \gamma)$
   $x^{k+1} = x^k - sz\nabla f(x^k)$

2. *write the following functions for computing the stepsize*
   (a) $s(f(\cdot), \nabla f(x^k), k, x^k, \gamma) = \gamma$, *i.e. a constant step size (use* $\gamma = 0.01$ *as default),*
   (b) $s(f(\cdot), \nabla f(x^k), k, x^k, \gamma) = 1/k$, *i.e. a stepsize that decreases with the number of iterations,*
   (c) *backtracking line search as seen in class with parameter* γ *(use* $\gamma = 0.9$ *as default).*

3. *Consider the function*

$$f(x) = 2x^4 - 9x^3 + 4x^2 + 11x + 3.$$

   (a) *Plot* f *in the interval [-2,4].*
   (b) *Is the function convex in that interval? (no proof is needed, just look at the plot)*
   (c) *For a starting point* $x^0 = -2$, $\varepsilon = 0.001$, *use gradient descent to minimize this function using all three methods. Compare the speed at which they converge (in terms of number of iterations and time).*

*Department of Economics, Yale University. jonas.lieber@yale.edu

(d) *For a starting point $x^0 = 1$, $\varepsilon = 0.001$, use gradient descent to minimize this function using all three methods. Compare the speed at which they converge (in terms of number of iterations and time).*

(e) *Does the starting point matter?*

Problem 2. *Convexity of OLS*[1]

1. *For any fixed $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$, we consider the function $g_{x,y} : \mathbb{R}^p \to \mathbb{R}$ defined by*

$$g_{x,y}(\beta) = y - \sum_{j=1}^{p} x_i \beta_j.$$

Show that for any $\lambda \in [0,1]$ and any $\beta^1, \beta^2 \in \mathbb{R}^p$,

$$g_{x,y}(\lambda\beta^1 + (1-\lambda)\beta^2) = \lambda g_{x,y}(\beta^1) + (1-\lambda)g(x,y)(\beta^2).$$

Does this imply that $g_{x,y}$ is convex according to the definition from class?

2. *Show that the function $h : \mathbb{R} \to \mathbb{R}$ defined by*

$$h(z) = z^2$$

is convex using the definition from class.

3. *For any fixed $x \in \mathbb{R}^p$ and $y \in \mathbb{R}$, we consider the function $g_{x,y} : \mathbb{R}^p \to \mathbb{R}$ defined by*

$$q_{x,y}(\beta) = h(g_{x,y}(\beta))$$

is convex. Hint: Use the linearity of $g_{x,y}$ and the convexity of h.

4. *For any two convex functions $r : \mathbb{R}^p \to \mathbb{R}$ and $s : \mathbb{R}^p \to \mathbb{R}$, show that the sum, i.e. the function $t : \mathbb{R}^p \to \mathbb{R}$ defined by*

$$t(w) = r(w) + s(w)$$

is convex.

5. *Use the previous result in an induction to show that for any finite number of convex functions $(v_i : \mathbb{R}^p \to \mathbb{R})_{i=1,\ldots,n}$, the sum, defined by*

$$v(z) = \sum_{i=1}^{n} v_i(z)$$

is convex.

[1] This problem is long, but only because I have split it up into small steps.

6. *Conclude that for any given* $((x_i, y_i))_{i=1,\ldots,n}$ *where* $x_i \in \mathbb{R}^p$ *and* $y_i \in \mathbb{R}$ *for all* i*, the OLS objective function* $f : \mathbb{R}^p \to \mathbb{R}$ *defined by*

$$f(\beta) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{i,j}\beta_j \right)^2$$

   is convex.

7. *For the case of solving the OLS problem, please discuss why it is useful for optimization when an objective is convex. Hint: Consider part 3 of problem 1.*

Problem 3. *Quickfire*

1. *Why is optimization relevant for Machine Learning?*

2. *Consider the model*

$$Y = f(X) + U$$

   where U *is independent of* X*. Suppose the variance of* U *is* $\sigma > 0$*. For a regression problem where we care about the Mean Squared Error, would you prefer the dataset* $((X_i, f(X_i)))_{i=1,\ldots,n}$ *over* $((X_i, Y_i))_{i=1,\ldots,n}$*? Why (not)?*[2]

3. *A norm is a function* $f : \mathbb{R}^p \to \mathbb{R}$ *such that*

   (a) $f(x) \geq 0$ *for all* $x \in \mathbb{R}^d$,

   (b) $f(x) = 0$ *if and only if* $x = 0 \in \mathbb{R}^p$,

   (c) *for all* $x \in \mathbb{R}^p$ *and* $a \in \mathbb{R}$, $f(ax) = af(x)$,

   (d) *for all* $x, y \in \mathbb{R}^p$, $f(x + y) \leq f(x) + f(y)$.

   Show that any norm is convex using the definition from class. Which properties of a norm did you use?

4. *In class, we have looked at minimization problems. Show that maximization problem can be written as a minimization problem.*

Problem 4. ***graduate students only***

1. *Show that the feasible set in a convex problem is convex.*

2. *Show that any local optimum in a convex problem is a global optimum. Hint: Suppose it wasn't the global optimum and then use the previous result.*

[2]Hints: 1) What would the bias-variance trade-off be for the dataset $((X_i, f(X_i)))_{i=1,\ldots,n}$? 2) What is the optimal solution to a regression problem?