Optimization, MATH 164
E. K. Ryu
Winter 2025

Homework 1
Due on Friday, January 24, 2025.

Problem 1: *Least-squares derivatives.* Let $X_1, \dots, X_N \in \mathbb{R}^p$ and $Y_1, \dots, Y_N \in \mathbb{R}$. Define

$$X = \begin{bmatrix} X_1^\intercal \\ \vdots \\ X_N^\intercal \end{bmatrix} \in \mathbb{R}^{N \times p}, \qquad Y = \begin{bmatrix} Y_1 \\ \vdots \\ Y_N \end{bmatrix} \in \mathbb{R}^N.$$

Let

$$\mathcal{L}(\theta) = \frac{1}{2}\|X\theta - Y\|^2.$$

Show that $\nabla_\theta \mathcal{L}(\theta) = X^\intercal(X\theta - Y)$.

Hint. Use the fact that

$$Mv = \sum_{i=1}^N M_{:,i} v_i \in \mathbb{R}^p$$

for any $M \in \mathbb{R}^{p \times N}$, $v \in \mathbb{R}^N$, where $M_{:,i}$ is the ith column of M for $i = 1, \dots, N$.

Problem 2: *Diverging univariate GD.* Consider the univariate function $f(x) = x^2/2$. Show that

$$x_{k+1} = x_k - \alpha f'(x_k)$$

with $x_0 \neq 0$ diverges if $\alpha > 2$.

Problem 3: *Diverging multivariate GD.* Let $X \in \mathbb{R}^{N \times p}$ and $Y \in \mathbb{R}^N$, and consider the optimization problem

$$\underset{\theta \in \mathbb{R}^p}{\text{minimize}} \quad f(\theta)$$

with

$$f(\theta) = \frac{1}{2}\|X\theta - Y\|^2.$$

Show

$$\theta_{k+1} = \theta_k - \alpha \nabla f(\theta_k)$$

with $\alpha > 2/\rho(X^\intercal X)$ diverges for most starting points $\theta_0 \in \mathbb{R}^m$. Here, ρ denotes the spectral radius, i.e., $\rho(X^\intercal X)$ is the largest eigenvalue of the symmetric matrix $X^\intercal X$. For simplicity, you may assume $X^\intercal X$ is invertible.

Hint. Let $\theta_\star = (X^\intercal X)^{-1} X^\intercal Y$ and show that

$$\theta_{k+1} - \theta_\star = \text{Some function of } (\theta_k - \theta_\star).$$

Remark. "Most starting points" can be formalized as "almost everywhere with respect to the Lebesgue measure". If you are unfamiliar with measure theory, you can understand the statement as holding for all starting points except for a lower dimensional set.

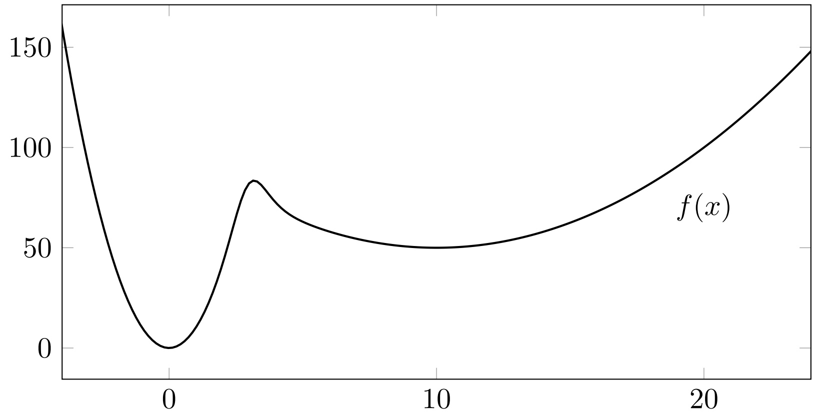
Problem 4: *GD converging to wide local minima.* Consider the optimization problem

$$\underset{x \in \mathbb{R}}{\text{minimize}} \quad f(x)$$

with

$$f(x) = \frac{10x^2 + e^{3(x-3)}((x-10)^2/2 + 50)}{1 + e^{3(x-3)}}.$$

Code for evaluating f and f' is implemented in the starter code `wideMinima.py`. We call the global minimum near $x = 0$ the *sharp* minimum and the local minimum near $x = 10$ the *wide* minimum.



Implement gradient descent and run it with random starting points within the range $[-5, 20]$. Experimentally demonstrate that gradient descent with step size $\alpha = 0.01$ converges to either of the two minima, with $\alpha = 0.3$ converges to the wide minimum, and with $\alpha = 4$ does not converge for most starting points.

Remark. The moral of this problem is that the step size of GD (and SGD) determines the sharpness of the minima the algorithm converges to. This has implications on the generalization performance in machine learning.[1]

[1] Y. Jiang, B. Neyshabur, H. Mobahi, D. Krishnan, S. Bengio, Fantastic Generalization Measures and Where to Find Them, *ICLR*, 2020.
P. Foret, A. Kleiner, H. Mobahi, B. Neyshabur, Sharpness-aware Minimization for Efficiently Improving Generalization, *ICLR*, 2020.

Problem 5: *L-smoothness lemma.* Let $f\colon \mathbb{R}^n \to \mathbb{R}$ be L-smooth. Show that

$$f(x) + \langle \nabla f(x), \delta \rangle - \frac{L}{2}\|\delta\|^2 \le f(x+\delta), \qquad \forall\, x, \delta \in \mathbb{R}^n.$$

Problem 6: *Verifying L-smoothness.* Let

$$\ell(r) = \begin{cases} r^2/2 & \text{for } -1 \le r \le 1 \\ |r| - 1/2 & \text{otherwise} \end{cases}$$

be the so-called Huber loss function. Consider the optimization problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad f(x) = \sum_{i=1}^{N} \ell(a_i^\intercal x - b_i),$$

where $a_i \neq 0$ and $b_i \in \mathbb{R}$ for $i = 1, \ldots, N$. Show the following:

(a) ℓ is 1-smooth.

(b) $\ell(a_i^\intercal x - b)$ is $(\|a_i\|^2)$-smooth for $i = 1, \ldots, N$.

(c) f is $(\sum_{i=1}^N \|a_i\|^2)$-smooth.

Hint. For (a), note that ℓ is continuously differentiable with $|\ell'(\cdot)| \le 1$, and use the fundamental theorem of calculus. For (b), use the Cauchy–Schwartz inequality. For (c), use the triangle inequality.

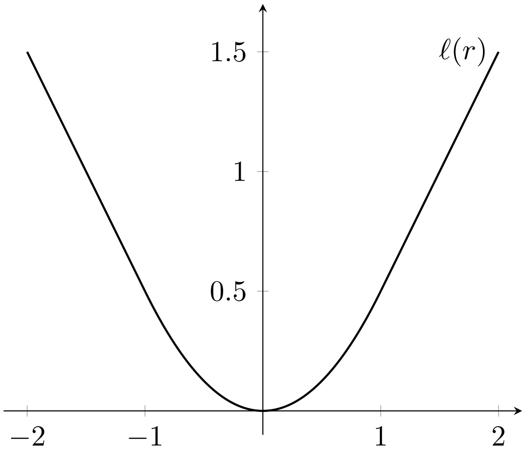


Figure 1: The Huber loss function of Problem 6.

Problem 7: *Counterexamples with gradient descent.* Consider the problem

$$\underset{x\in\mathbb{R}^d}{\text{minimize}} \quad f(x)$$

and the gradient descent algorithm

$$x_{k+1} = x_k - \alpha\nabla f(x_k).$$

We describe several problem instances in the following. For each instance, answer the following questions. (i) What is $p_\star = \inf_x f(x)$? (ii) Does $x_k \to x_\star$ for some global minimizer $x_\star \in \mathbb{R}^d$? (iii) Does $\nabla f(x_k) \to 0$? (iv) Does $f(x_k) \to p_\star$?

(a) Consider any $x_0 \in \mathbb{R}^d$ and $f(x) = c^\intercal x$ with some $c \neq 0$.

(b) Consider $x_0 = 0$ and f defined as

$$\begin{aligned} g(r) &= \begin{cases} 1-r & \text{for } r \le 0 \\ \frac{1}{r+1} & \text{for } r > 0 \end{cases} \\ f(x) &= g(c^\intercal x) \end{aligned}$$

with some nonzero $c \in \mathbb{R}^d$.

(c) Consider $x_0 = 0$ and f defined as

$$\begin{aligned} h(r) &= \begin{cases} 2-r & \text{for } r \le 1 \\ 1-\log(r) & \text{for } r > 1 \end{cases} \\ f(x) &= h(c^\intercal x + 1) \end{aligned}$$

with some nonzero $c \in \mathbb{R}^d$.

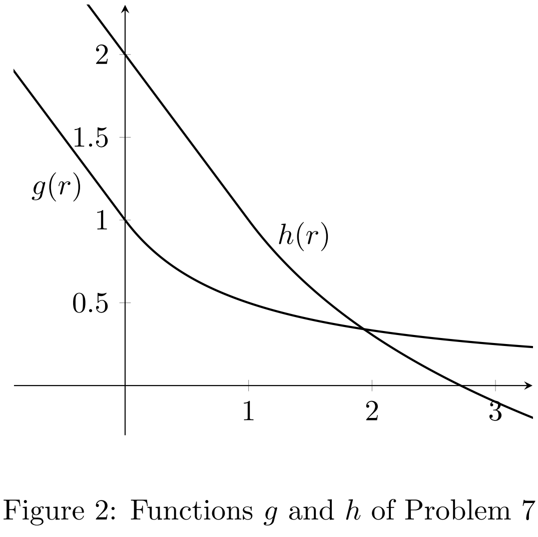


Figure 2: Functions g and h of Problem 7.