Optimization for Data Science, Autumn 2026 - Homework 2
Due Sunday Nov/1, 11:59 pm Chicago time (Gradescope)

Your submitted solutions should be your own work. You may discuss the problems with peers and consult sources, including AI tools, but you must acknowledge the sources you use. **Write all submitted solutions from scratch, in your own words, and without assistance from an LLM during the writing process. Copying AI-generated solutions is not permitted.** You are responsible for understanding every part of your submission and may be asked to defend it in a one-on-one oral examination. If you cannot demonstrate that understanding, you may receive a zero on the assignment.

Problem 1 – Harder, Better, Faster, Stronger

For this problem you will analyze the convergence of a variant of Nesterov's accelerated gradient descent for strongly convex smooth functions. Let $f\colon \mathbf{R}^d \to \mathbf{R}$ be a continuous μ strongly convex function with L Lipschitz gradient. Write $f^\star = \min f = f(w^\star)$.

Algorithm 1 `AcceleratedGradientDescent`

Input: Start point w_0, number of iterations T

Initialize $y_0 \leftarrow w_0$, $\lambda_1 \leftarrow 1$

Iterate $(k = 0, \dots, T-1)$:

$$\lambda_{k+2} \leftarrow \frac{1+\sqrt{1+4\lambda_{k+1}^2}}{2}, \quad w_{k+1} \leftarrow y_k - \frac{1}{L}\nabla f(y_k), \quad y_{k+1} \leftarrow w_{k+1} + \left(\frac{\lambda_{k+1}-1}{\lambda_{k+2}}\right)(w_{k+1}-w_k)$$

Output w_T

(a) In class we saw that Algorithm 1 achieves a rate of convergence of $f(w_k) - f^\star \leq \frac{2L\|w_0 - w^\star\|^2}{k^2}$ for $k \geq 1$. Find an integer value $0 < T^\star$ (in terms of μ and L) such that if we set $k \geq T^\star$ we have

$$f(w_k) - f^\star \leq (f(w_0) - f^\star)/2.$$

(b) Consider the method that at each iteration updates:

$$z_{i+1} \leftarrow \texttt{AcceleratedGradientDescent}(z_i, T^\star)$$

starting with $z_0 = w_0$. Here $T^\star$ is the value you found in the previous question. That is, at each iteration of the algorithm we run an instance of Accelerated Gradient Descent for $T^\star$ iterations and then restart the method. This is known as "Restarted accelerated gradient descent." The total number of iterations of this algorithm is the sum of all the iterations of Accelerated Gradient Descent that the algorithm performs. Show that for any $0 < \varepsilon < f(w_0) - f^\star$, this algorithm finds a point $z \in \mathbf{R}^d$ such that $f(z) - f^\star \leq \varepsilon$ after a total number of iterations of at most

$$\lceil 4\sqrt{L/\mu} \rceil \cdot \left\lceil \log_2\left(\frac{f(w_0) - f^\star}{\varepsilon}\right)\right\rceil \quad \text{iterations.}$$

Problem 2 – Getting over the span

The lower bound in Lecture 4 was first proved under the span assumption. Here you will remove that assumption for deterministic first-order methods. Let $f\colon \mathbf{R}^{2T} \to \mathbf{R}$ be the strongly convex chain constructed in the handwritten Lecture 4 notes. Use the notation κ, ρ, a, and E_j from those notes, and assume $\kappa > 1$. You may use the properties of this function proved in class without proving them again.

Consider a deterministic algorithm that queries $w_0 = 0, w_1, \dots, w_{T-1}$, receiving a function value and gradient at each query, and then returns w_T. Each point may depend on all previous oracle answers. There is no span restriction, and the output need not be one of the queried points. We will choose a rotation of f that is iteratively chosen as the queries arrive.

(a) Rotating the function. Let $U = [u_1 \ \cdots \ u_{2T}] \in \mathbf{R}^{2T\times 2T}$ be orthogonal, i.e., $UU^\top = U^\top U = I$, and define

$$F_U(w) = f(U^\top w).$$

Compute its gradient and Hessian. Show that this transformation preserves the smoothness and strong convexity constants, the minimum value, and the initial function gap Δ.

(b) Choosing directions as queries arrive. Our goal is to iteratively pick a rotation U so that for all $k \in \{1, \dots, T-1\}$ we have

$$w_k \in \operatorname{span}\{u_1, \dots, u_{2k}\} \quad \text{and} \quad \nabla F_U(w_k) \in \operatorname{span}\{u_1, \dots, u_{2k+1}\}. \tag{1}$$

(1) Choose any unit vector u_1, and show that it determines $F_U(0)$ and $\nabla F_U(0)$.

(2) For $k \in \{1, \dots, T-1\}$, suppose that the first $2k-1$ columns have been chosen. Show that, once the algorithm chooses w_k, you can extend these columns to an orthonormal family $u_1, \dots, u_{2k+1}$ such that (1) holds.

(3) Show that the values of $F_U(w_k)$ and $\nabla F_U(w_k)$ do not depend on the last $2T-2k-1$ columns of U. Conclude that this iterative procedure to construct U preserves all previous oracle answers.

(c) Accounting for the final output. After all T answers, the algorithm determines w_T. At this point, $u_1, \dots, u_{2T-1}$ have been chosen. Show that you can choose the sign of the final column u_{2T} so that

$$\langle u_{2T}, w_T \rangle \le 0.$$

Explain why this choice does not change any oracle answer. Prove that running the algorithm on the resulting single, fixed function F_U reproduces the same queries, answers, and output.

(d) The lower bound. Apply the Cauchy–Schwarz argument from Step 4 of the Lecture 4 proof to show that any $w \in \mathbf{R}^{2T}$ with $(w)_{2T} \le 0$ satisfies $\frac{f(w)-f^\star}{\Delta} \ge \rho^{4T-2}$. **Hint:** The last coordinate of $w^\star - w$ is at least $a\rho^{2T-1}$. Deduce that

$$\frac{F_U(w_T) - F_U^\star}{\Delta} \ge \rho^{4T-2}.$$

Compare this with the ρ^{2T} bound under the span assumption. Explain why the dependence of the iteration lower bound on κ and the desired relative accuracy is unchanged up to a constant factor, provided the dimension is sufficiently large. Does the construction give one function that is hard for every algorithm, or a hard function for each algorithm?

Problem 3 – Pitfalls and hidden beauty of Newton

This question concerns the divergent behavior and interesting properties of the Newton method. In parts (a)–(c), we are finding a root of a scalar function F, using

$$w_{k+1} = w_k - \frac{F(w_k)}{F'(w_k)}.$$

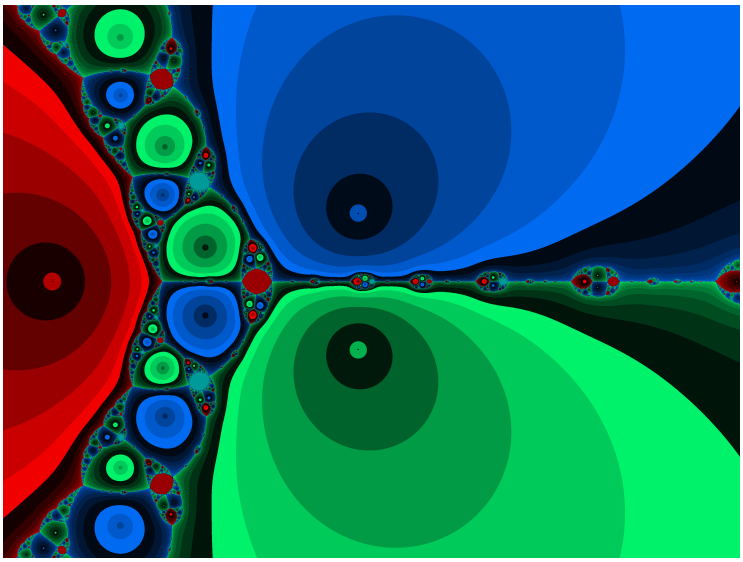
In class, we considered the special case $F = \nabla f$.

(a) Divergence. Find a continuously differentiable function $F\colon \mathbf{R} \to \mathbf{R}$ with exactly one root $w^\star$, such that $F'(w^\star) \neq 0$, and an initial point $w_0 \in \mathbf{R}$ for which the Newton iterates diverge to infinity. Check that every iterate is well defined.

(b) Cycles. Consider $F(w) = w^3 - 2w + 2$. Find two distinct points $w_0, w_1 \in \mathbf{R}$ such that the Newton iterates cycle between them forever:

$$w_0,\ w_1,\ w_0,\ w_1,\ w_0, \ldots.$$

(c) Fractals. Over the complex numbers, the boundaries between the basins of attraction of Newton can be fractals. The picture below comes from the polynomial in part (b).



3Blue1Brown has an absolutely fantastic video about it. For this question, you just need to watch the video :).

(d) Affine invariance. Now consider Newton's method for minimizing a function. Let $f\colon \mathbf{R}^d \to \mathbf{R}$ be C^2 with invertible Hessians everywhere. Let $A \in \mathbf{R}^{d\times d}$ be invertible, and define $\psi(z) = f(Az)$. Starting from $w_0 = Az_0$, run

$$w_{k+1} = w_k - [\nabla^2 f(w_k)]^{-1}\nabla f(w_k), \text{ and}$$
$$z_{k+1} = z_k - [\nabla^2 \psi(z_k)]^{-1}\nabla \psi(z_k).$$

Show that $w_k = Az_k$ for every k.

Problem 4 – Filling in the details of BFGS

Write $g_k = \nabla f(w_k)$, and let $B_k \succ 0$ be an approximation to the Hessian, with $H_k = B_k^{-1}$. A quasi-Newton step has the form

$$p_k = -H_k g_k \qquad \text{and} \qquad w_{k+1} = w_k + \eta_k p_k.$$

The measured change in position and gradient is

$$s_k = w_{k+1} - w_k \qquad \text{and} \qquad y_k = g_{k+1} - g_k.$$

Here y_k denotes a gradient difference, as in Lecture 7; it is not the extrapolated point from Problem 1.

(a) Lemma D2.0. Prove the lemma from the handwritten notes for Lecture 7: if f is L-smooth and $B_k \succ 0$, then

$$f(w_{k+1}) \le f(w_k) - \left(\eta_k - \frac{L\eta_k^2}{2\lambda_{\min}(B_k)}\right) \|g_k\|_{B_k^{-1}}^2,$$

where $\|v\|_M^2 = v^\top M v$. Deduce a range of positive step sizes that guarantees strict decrease whenever $g_k \neq 0$.

(b) The BFGS update. Suppose $y_k^\top s_k > 0$, and set $\rho_k = (y_k^\top s_k)^{-1}$. The inverse BFGS update is

$$H_{k+1} = (I - \rho_k s_k y_k^\top) H_k (I - \rho_k y_k s_k^\top) + \rho_k s_k s_k^\top.$$

Show that H_{k+1} is symmetric positive definite and satisfies the secant equation $H_{k+1} y_k = s_k$. Explain why the next direction is a descent direction unless the gradient is zero. Expand the formula into a rank-two update that can be evaluated in $O(d^2)$ operations.

(c) The Wolfe conditions have a solution. Let f be continuously differentiable, let p be a descent direction at w, and suppose f is bounded below on the ray $\{w + \eta p : \eta \ge 0\}$. For any $0 < c_1 < c_2 < 1$, prove that there exists $\eta > 0$ satisfying

$$f(w + \eta p) \le f(w) + c_1 \eta \nabla f(w)^\top p, \text{ and}$$
$$\nabla f(w + \eta p)^\top p \ge c_2 \nabla f(w)^\top p.$$

Hint: Compare $\phi(\eta) = f(w + \eta p)$ with the line $\phi(0) + c_1 \eta \phi'(0)$.

(d) Finding a Wolfe step. Under the assumptions of part (c), consider the following procedure, where $\phi(\eta) = f(w + \eta p)$. Prove that, in exact arithmetic, it terminates after finitely many iterations and returns a positive step size satisfying both Wolfe conditions.

Algorithm 2 `WolfeLineSearch`

Input: Function ϕ, constants $0 < c_1 < c_2 < 1$

Initialize $\ell \leftarrow 0$, $u \leftarrow +\infty$, $\eta \leftarrow 1$

Iterate until a step is accepted:

1. If $\phi(\eta) > \phi(0) + c_1 \eta \phi'(0)$, set $u \leftarrow \eta$.
   Else if $\phi'(\eta) < c_2 \phi'(0)$, set $\ell \leftarrow \eta$.
   Otherwise, accept η and stop.
2. Set
$$\eta \leftarrow \begin{cases} (\ell + u)/2, & \text{if } u < +\infty, \\ 2\eta, & \text{otherwise.} \end{cases}$$

Output The accepted step size η.
