HOMEWORK 5 (MODERN OPTIMIZATION ALGORITHMS, MATH 485)

Due date: See class webpage on Canvas.

Unless otherwise stated, $\mathcal{H}$ is a real Hilbert space with inner product $\langle \cdot \mid \cdot \rangle$ and induced norm $\| \cdot \|$.

Instructions for Homework 5: For this assignment, you are welcome to use an LLM for assistance, **and** you are welcome to use any of the course textbooks as a reference. If you use an LLM, you must disclose on your homework:

1. which section(s) of the HW were used for LLM assistance/hints,
2. which LLM(s) were used, and
3. what steps you took to ensure your solution is valid. Keep in mind that your submission is your own, and you have full responsibility for its correctness.

If any course textbooks are used, their specific theorem/results must be cited, and the homework must explain how that result is used to solve your problem.

Let $\mathcal{H} = \mathbb{R}^n$ be equipped with the standard Euclidean inner product (dot product), let $A \in \mathbb{R}^{m \times n}$, $b \in \mathbb{R}^m$, $\lambda > 0$, and recall, for every $x \in \mathbb{R}^n$, $\|x\|_1 = \sum_{i=1}^n |x_i|$. Consider the LASSO problem,

$$\operatorname*{minimize}_{x \in \mathcal{H}} \frac{1}{2}\|Ax - b\|^2 + \lambda \|x\|_1. \tag{1}$$

Definition A function is *coercive* if $\lim\limits_{\|x\| \to \infty} f(x) = \infty$, i.e., for every sequence $(x_n)_{n \in \mathbb{N}}$ of points in $\mathcal{H}$ satisfying $\|x_n\| \to \infty$, we have $f(x_n) \to +\infty$.

QUESTIONS

1. This question pertains to existence of optimal solutions.
   (a) Show that if $f \in \Gamma_0(\mathcal{H})$ is coercive, then $\operatorname{argmin} f \neq \varnothing$.
   (b) Show that if $f, g \in \Gamma_0(\mathcal{H})$, and f is coercive and g is bounded below, i.e., there exists $g_0 \in \mathbb{R}$ such that $(\forall x \in \mathcal{H})\ \ g(x) \geq g_0$, then $f + g$ is coercive.
   (c) Use 1(a) and 1(b) to show that there is at least one optimal solution to (1).
2. This problem uses the Forward-Backward algorithm to solve (1). You may assume 1(c) is true.
   (a) Use the fact that $\nabla(\frac{1}{2}\|\cdot\|^2) = \mathrm{Id}$ to provide a short derivation for $\nabla(\frac{1}{2}\|A \cdot - b\|^2)$.
   (b) Provide pseudocode for an iterative algorithm based on the Forward-Backward algorithm for specifically solving (1).
   (c) Use the theorem provided in class to characterize convergence of the algorithm in Part 2(b).
3. This problem uses the Douglas-Rachford algorithm to solve (1). You may assume 1(c) is true, that $n = m$ and that A is orthogonal, so $A^* = A^{-1}$.

(a) Find an expression for $\text{Prox}_{\gamma f}$ where $f = \frac{1}{2}\|A \cdot -b\|^2$. For this part of this problem, finding a reference citation may be easier than a direct proof.

(b) Provide pseudocode for an iterative algorithm based on the Douglas-Rachford algorithm for specifically solving (1).

(c) Use the theorem provided in class to characterize convergence of the algorithm in Part 3(b).

4. Use the product-space Douglas-Rachford algorithm from class to provide the pseudocode for an algorithm for solving the sparse linear binary classifier training problem from earlier this semester. Provide a theorem that characterizes convergence of this algorithm for this problem.

5. Even if we have a function that is differentiable, we may still get better numerical performance by using its proximity operator instead of its gradient, as shown in the following reference.

   P. L. Combettes and L. E. Glaudin, Proximal activation of smooth functions in splitting algorithms for convex image recovery, *SIAM J. Imaging Sci.*, vol. 12 (4), pp. 1905–1935, 2019.
   Preprint: https://arxiv.org/pdf/1803.02919

Conduct a numerical investigation of both algorithms in problems 2 and 3 to investigate which algorithm does better. You may plot function value versus iteration, distance to a solution, or both. You can compare this to iteration count, wallclock time, or some other metric you deem reasonable. You may choose A, b, and λ based on your interest, subject to the assumptions of problem 3, using operators from class or a properly cited reference. The solution to this problem may be typed and printed, and the code should also be emailed to the instructor.

6. As seen in class, the following reference has proven that the iterates of the Frank-Wolfe algorithm may not converge, even for convex objective functions:

   J. Bolte, C. W. Combettes, E. Pauwels, The iterates of the Frank–Wolfe algorithm may not converge, *Math. Oper. Res.*, vol. 49 (4), pp. 2565–2578, 2024.
   Preprint: https://arxiv.org/pdf/2202.08711

   However, one can prove (via a continuity argument) that every convergent subsequence of the iterates converges to a solution of the underlying minimization problem. This question pertains to the setting when the objective function is nonconvex.

Suppose that f is L-smooth (possibly nonconvex), $L > 0$, and that $C \subset \mathcal{H}$ is nonempty, convex, and compact. Suppose that $(x_t)_{t\in\mathbb{N}}$ is the sequence of iterates generated by the Frank-Wolfe algorithm using the short-step rule[1] for the stepsize sequence $(\gamma_t)_{t\in\mathbb{N}}$. Let $(x_{t_k})_{k\in\mathbb{N}}$ be any *convergent* subsequence of the Frank-Wolfe iterates, i.e., there exists $x^* \in C$ such that $\lim_{k\to\infty} \|x_{t_k} - x^*\| = 0$. Prove or disprove that x^* is a stationary point for minimizing f over C, i.e.,

$$0 \in \partial_{\mathfrak{C}}(f + \iota_C)(x^*). \tag{2}$$

Hint: Use the relationship between the Frank-Wolfe gap and Clarke stationarity proven in class.

[1] Let $v_t = \text{LMO}_C(\nabla f(x_t))$. If $x_t = v_t$ we have $\gamma_t = 0$; otherwise, $\gamma_t = \min\{1, \langle \nabla f(x_t) \mid x_t - v_t\rangle / (L\|v_t - x_t\|^2)\}$.