MATH 6001. Nonlinear Optimization in Machine Learning.
Final Project.

There are 8 problems that are marked with underlines. Each problem is worth 5 points, and the total is 40 points.

Part 1. Consider the heavy ball method iteration at dimension 2

$$x^{k+1} = x^k - \alpha \nabla f(x^k) + \beta(x^k - x^{k-1}) \ , \tag{0.1}$$

where $x^{-1} = x^0 \in \mathbb{R}^2$ and $\alpha, \beta > 0$. Let the function f be quadratic of the form

$$f(x) = \frac{1}{2} x^T Q x - b^T x + c \ , \tag{0.2}$$

such that the Hessian matrix Q is a 2×2 positive definite matrix with the two eigenvalues

$$0 < m = \lambda_2 \leq \lambda_1 = L < \infty \ , \tag{0.3}$$

and $b \in \mathbb{R}^2$ and $c \in \mathbb{R}$. Set

$$\alpha = \frac{4}{(\sqrt{L} + \sqrt{m})^2} \ , \ \beta = \frac{\sqrt{L} - \sqrt{m}}{\sqrt{L} + \sqrt{m}} \ . \tag{0.4}$$

Following the "spectral method" proof of the convergence of Nesterov's scheme in §4.1 of Chapter 4 in the Lecture Notes, we can obtain a linear convergence rate on the convex quadratic $f(x)$ in (0.2). We split the proof into 5 steps.

Problem 1. Write the algorithm as a linear recursion $w^{k+1} = T w^k$ for appropriate choice of matrix T and state variables w^k.

Problem 2. Use a transformation to express T as a block–diagonal matrix, with 2×2 blocks T_i on the diagonals, where each T_i depends on a single eigenvalue λ_i of Q.

Problem 3. Find the eigenvalues $\mu_{i,1}$, $\mu_{i,2}$ of each T_i as a function of λ_i, α and β.

Problem 4. Show that for the given values of α and β, these eigenvalues are all complex.

Problem 5. Show that in fact $|\mu_{i,1}| = |\mu_{i,2}| = \sqrt{\beta}$ for all $i = 1, 2$, so that $\rho(T) \equiv \max_{i=1,2} \max(|\mu_{i,1}|, |\mu_{i,2}|) = \sqrt{\beta} \approx 1 - \kappa^{-1/2}$, where the condition number $\kappa = \dfrac{L}{m} \gg 1$.

Part 2. Consider the standard finite–sum objective function $f(x) = \frac{1}{n}\sum\limits_{i=1}^{n} f_i(x)$, which is commonly seen in machine learning. Let us further assume that the component functions f_i are further associated with Gaussian noise model, that is

$$[\nabla f_i(x)]_j = [\nabla f(x)]_j + \varepsilon_{ij} \ , \ \text{for all } i = 1, 2..., n \text{ and } j = 1, 2, ..., d \ , \tag{0.5}$$

where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)$ is a Gaussian distribution with mean 0 and variance σ^2.

Problem 6. Show that when we estimate the gradient using a randomly sampled minibatch $\mathcal{S} \subseteq \{1, 2, ..., n\}$, that is,

$$g = \frac{1}{|\mathcal{S}|}\sum_{i\in\mathcal{S}} \nabla f_i(x) \ , \tag{0.6}$$

then we have

$$\mathbf{E}\|g - \nabla f(x)\|^2 = \frac{d}{|\mathcal{S}|}\sigma^2 \ .$$

Problem 7. Following Problem 6, show that

$$\mathbf{E}(\|g\|^2) = \|\nabla f(x)\|^2 + \frac{d}{|\mathcal{S}|}\sigma^2 \ .$$

Problem 8. Consider a minibatch strategy for the additive Gaussian noise model, where the gradient estimate is given by

$$g(x; \xi_1, \xi_2, ..., \xi_s) = \nabla f(x) + \frac{1}{s}\sum_{j=1}^{s} \xi_j \ ,$$

where each ξ_j is i.i.d with distribution $\mathcal{N}(0, \sigma^2 I)$, that is a multivariate normal distribution with mean 0 and covariance matrix $\sigma^2 I$, and $s \geq 1$. Show that

$$\mathbf{E}_{\xi_1,\xi_2,...,\xi_s}(\|g(x; \xi_1, \xi_2, ..., \xi_s)\|^2) = \|\nabla f(x)\|^2 + \frac{d}{s}\sigma^2 \ .$$