CSCI 145 Problem 12: Spend a Diverse Batch

August 27, 2026

Submission Instructions

Please upload your work to Gradescope by **11:59pm on DATE**.

- Turn in a single PDF of *handwritten* solutions. Typed (including LaTeX) solutions will not be accepted.
- Problems are grouped into three-week **cycles**. Once per cycle, you will meet with the grader and present the solution to *one* problem from that cycle, selected at random when you arrive; bring your work and be ready to explain any of them.

Problem 12: Spend a Diverse Batch 2 points

A training set contains C source examples, each with many stochastic augmentations. The gradient from augmentation j of source c is modeled as

$$\mathbf{g}_{c,j} = \mathbf{g} + \mathbf{z}_c + \boldsymbol{\varepsilon}_{c,j},$$

where the vector noises have mean zero, the $\mathbf{z}_c$ are independent across sources, the $\boldsymbol{\varepsilon}_{c,j}$ are independent across all augmentations, and the two families are independent. Assume

$$\mathbb{E}\|\mathbf{z}_c\|_2^2 = \tau^2, \qquad \mathbb{E}\|\boldsymbol{\varepsilon}_{c,j}\|_2^2 = \sigma^2.$$

You have a budget of B gradient evaluations. If b_c augmentations are drawn from source c, then b_c are nonnegative integers satisfying $\sum_{c=1}^{C} b_c = B$. Design the allocation $(b_1, \ldots, b_C)$ that minimizes the mean-squared error of the averaged gradient. Prove that your allocation is optimal, give the minimum error as an explicit function of B, C, τ, and σ, and compare it with spending all B evaluations on augmentations of one source.

Use the result to assess this rule of thumb: "At fixed batch size, additional data augmentation is as valuable as an additional independent example." State precisely when the rule is approximately right and when it can be badly wrong.

Results you may use. For independent mean-zero random vectors, the expected squared norm of a sum is the sum of the expected squared norms. Among integers with a fixed sum, you may quote or prove the relevant balancing inequality.