Homework 4: Optimization geometry and shared representations

Due Friday, November 6, 2026, at 11:59 p.m. This assignment is worth 100 points and 10% of the course grade. Late homework is not accepted; the two-lowest-grade replacement described in the syllabus covers ordinary illness, overload, and missed work. Complete every exercise and part assigned points. Submit one ZIP archive named hw04-ccid.zip, containing a single top-level directory named hw04-ccid with your PDF, LaTeX source, and any supporting files. Replace ccid with your University of Alberta CCID.

General instructions. Justify every answer, including yes/no answers, even when the question does not explicitly ask for a justification.

Exercise profiles. Filled mountains show difficulty on a four-mountain scale: one is a direct application, two uses a familiar method, three combines ideas, and four is a substantial challenge. Filled cogs show technical load on a three-cog scale: one is light, two is sustained, and three is dense. Filled hourglasses show estimated time: one is under 20 minutes, two is 20–45 minutes, three is 45–90 minutes, and four is more than 90 minutes. Pale pictograms show the unused part of each scale. The work row uses the canonical order proof, calculation, construction, computation, and interpretation; pale icons mark work types not used. ◎ marks the intended learning outcome. Time estimates cover the non-optional parts. Ratings are not prerequisites or grading weights. A superscript asterisk marks an exercise based on supplementary lecture material; it is not a difficulty rating and does not make the exercise optional when assigned.

Name: YOUR NAME CCID: YOUR CCID

Points: (a): 7, (b): 7, (c): 6; total 20.

Exercise 1: Equal margins preserve a signed-gradient ray.[*]

◎ Recognize when coordinate normalization confines an adaptive method to one parameter ray.

Let $\mathbf{X} \in \mathbb{R}^{n \times d}$ have row i equal to $x_i^\top$, let $y \in \{-1, 1\}^n$, and define

$$s = \operatorname{sign}(\mathbf{X}^\top y).$$

Assume $\mathbf{X}s = cy$ for some $c > 0$.

(a) For a differentiable margin loss $L(w) = \sum_{i=1}^n \ell(y_i \langle w, x_i \rangle)$, prove that

$$\nabla L(\alpha s) = \ell'(c\alpha)\mathbf{X}^\top y.$$

When $\ell' < 0$, prove by induction that signed gradient descent initialized at zero remains on the ray $\{\alpha s : \alpha \geq 0\}$.

(b) For the squared loss $Q(w) = \frac{1}{2}\|\mathbf{X}w - y\|_2^2$, prove

$$\nabla Q(\alpha s) = (c\alpha - 1)\mathbf{X}^\top y.$$

Explain why an adaptive update that divides each nonzero gradient coordinate by a matching coordinatewise magnitude also remains on the ray.

(c) Explain why replacing the denominator $|g_j|$ by $|g_j| + \epsilon$ for fixed $\epsilon > 0$ generally breaks the exact ray calculation as the gradient vanishes. Contrast the consequences for a finite least-squares interpolant and for logistic loss on separable data.

Solution

Write your solution here.

Stochastic and incremental updates

Points: (a): 4, (b): 12, (c): 4; total 20.

Exercise 2: Removing the stochastic noise floor.

Turn the stochastic PL recursion into an explicit $O(1/t)$ bound using a decreasing stepsize.

Let $\delta_t \geq 0$ satisfy

$$\delta_{t+1} \leq (1 - \mu\eta_t)\delta_t + \frac{\beta v}{2B}\eta_t^2.$$

Choose

$$\eta_t = \frac{2}{\mu(t + t_0)}, \qquad t_0 \geq \frac{2\beta}{\mu},$$

and define

$$C_0 = \max\left\{t_0\delta_0, \frac{2\beta v}{B\mu^2}\right\}.$$

(a) Verify that $\eta_t \leq 1/\beta$ and rewrite the recursion as

$$\delta_{t+1} \leq \left(1 - \frac{2}{t + t_0}\right)\delta_t + \frac{2\beta v}{B\mu^2(t + t_0)^2}.$$

(b) Prove by induction that

$$\delta_t \leq \frac{C_0}{t + t_0} \qquad \text{for every } t \geq 0.$$

(c) State why this bound has no fixed noise floor, in contrast to the constant-stepsize bound, whose nondecaying term is $\beta\eta v/(2\mu B)$.

Hints For (b), put $s = t + t_0$ and use $2C_0 - 2\beta v/(B\mu^2) \geq C_0$.

Solution

Write your solution here.

Adaptive geometry

Points: (a): 6, (b): 8, (c): 6; total 20.

Exercise 3: A rotated quadratic defeats diagonal preconditioning.

Distinguish coordinatewise rescaling from correction of rotated curvature.

Let $\kappa > 1$ and

$$Q = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}, \qquad H = Q\begin{pmatrix} 1 & 0 \\ 0 & \kappa \end{pmatrix}Q^\top, \qquad L(\theta) = \frac{1}{2}\theta^\top H\theta.$$

For a positive-definite preconditioner P, the transformed Hessian is $P^{1/2}HP^{1/2}$.

(a) Compute H and prove that its condition number is κ.

(b) Let $P_{\text{diag}} = \operatorname{diag}(H)^{-1}$. Compute $P_{\text{diag}}^{1/2} H P_{\text{diag}}^{1/2}$ and prove that its condition number remains κ.

(c) Let $P_{\text{full}} = H^{-1}$. Compute the transformed Hessian and its condition number. Explain precisely which geometric information is absent from P_{diag}.

HINTS The two diagonal entries of H are equal. Orthogonal similarity preserves eigenvalues.

Solution

Write your solution here.

Joint learning and transfer

Points: (a): 8, (b): 8, (c): 6; total 22.



Exercise 4: The approximation cost of sharing.

Compute a sharing cost when every task is individually realizable, and compare its effect with the estimation cost.

Let U, V be independent fair bits and let $X = (U, V)$ on every task. A fraction p of the T source tasks have label U, and the remaining fraction have label V, where pT is an integer. Take $\mathcal{H} = \{h_U, h_V\}$ with $h_U(U, V) = U$ and $h_V(U, V) = V$, $\mathcal{G} = \{z \mapsto z,\ z \mapsto 1 - z\}$, and 0–1 loss. Each task supplies n independent examples, and the task samples are independent.

(a) Compute $R_t^\star$, $A_t(h_U)$, and $A_t(h_V)$ for both kinds of task. Deduce the exact value of Δ_{share}.

(b) Specialize the agnostic joint-ERM bound to these classes, assuming exact empirical optimization.

(c) Suppose $0 < p < 1$. Explain why increasing n cannot make the average risk of a shared predictor tend to zero. Compare with consistent independent learning. Identify the assumption of the realizable joint zero-loss theorem that fails.

Solution

Write your solution here.

Linear representations

Points: (a): 6, (b): 5, (c): 7; total 18.



Exercise 5: Many repeated tasks can hide a shared feature.

Distinguish the number of source tasks from coverage of the shared feature directions.

Let $d = 3$, $k = 2$, and let every source-task head have the form $a_i = (c_i, 0)^\top$, where $c_i \in \mathbb{R}$. Consider the two feature matrices

$$B = [e_1\ e_2] \qquad \text{and} \qquad \widetilde{B} = [e_1\ e_3].$$

For either feature matrix, source task i generates $Y = X^\top B a_i + \varepsilon$, where $X \sim \mathcal{N}(0, I_3)$ and the same independent noise law is used in both models.

(a) Prove that the joint law of (X, Y) for every source task is identical under B and $\widetilde{B}$.

(b) For uniform sampling from M source tasks, compute the task-diversity matrix

$$D = \frac{1}{M} \sum_{i=1}^{M} a_i a_i^\top$$

and its eigenvalues.

(c) Let the new-task head be $a_{\text{new}} = (0, 1)^\top$. Compute its ambient parameter under each feature matrix. Explain why no procedure using only the source data can determine which of these two new-task parameters is correct.

Solution

Write your solution here.