CS590 Homework 2

Qilin Ye

March 18, 2025

Solution to problem 1. (a) Let us consider the simplest case with $n = d = 1$, and $y_1 = 0, x_1 = 1$. The loss function reduces to just $L(w) = \sigma(w)^2$ where σ is the sigmoid function. We know that the sigmoid function is bounded in $[0, 1]$; thus, so is $L(w)$. A nonconstant convex function defined on $\mathbb{R}$ cannot be bounded from both below and above, so $L(w)$ must be nonconvex.

(b) Consider a distribution/antiderivative defined as

$$\Phi(z) = \int_0^z \sigma(u)\,\mathrm{d}u$$

where we choose the lower limit as 0 to conveniently have $\Phi(0) = 0$. The antiderivative of $x_i\sigma(w^T x_i) - y_i x_i$ w.r.t. w is $\Phi(w^T x_i) - y_i w^T x_i$. That is, we have found

$$\hat{L}(w) := 2\sum_{i=1}^{n}\left(\Phi(w^T x_i) - y_i w^T x_i\right).$$

Since ϕ is monotonically increasing, Φ is convex. For a fixed x, $w \mapsto \Phi(w^T x_i) - y_i w^T x_i$ is also convex, as the first term is convex and the second term only linear/affine. Therefore, as the sum of n convex functions, $\hat{L}$ itself is convex.

Finally, recall $w*$ is the ground truth to $y = \sigma((w^*)^T x)$. Plugging this into the gradient $\nabla\hat{L}(w)$ we see that each term $\sigma(w^T x_i) - y_i$ becomes zero, and so $\nabla\hat{L}(w) = 0$. By convexity, this implies w^* is a global minimum of $\hat{L}$.

Solution to problem 2. (a) Observe $L_{(i,j)} = n_{i,j}(M_{i,j} - M^*_{i,j})^2$ depends only on the rows U_i, U_j, and V_i, V_j. This implies that the gradient $\nabla L_{(i,j)}$ depends only on the blocks corresponding to indices i and j.

As we assumed that $U^T U = mI$, the rows are orthogonal. Consequently, for non-identical (i, j) and (i', j'), the gradients are either supported on disjoint parameter blocks, or on blocks that are orthogonal due to our assumption, leading to $\left\langle \nabla L_{(i,j)}, \nabla L_{(i',j')}\right\rangle = 0$. This completes the proof.

(b) Recall

$$f((i,j)) = \sum_{(i',j')\in\Omega} K_{(i,j),(i',j')}\alpha_{(i',j')}$$

where the α's are learnable. For any $(i, j) \notin \Omega$, we immediately know that all corresponding K-entries (i.e. $K_{(i,j),(i',j')}$ for $(i', j') \in \Omega$ and this fixed (i, j)) are zero. Consequently the prediction must be 0.

(c) At initialization, M vanishes. Therefore, at initialization, each loss term $(M_{i,j} - z_i z_j)^2$ contributes a factor $z_i z_j \cdot U_j$ to the gradient $\nabla_{U_i} L_{i,j}$. As the rows are isotropic by the assumption that $U^T U = mI$, and each pair is observed with probability $1/2$, we are essentially deterministically summing over a lower triangle double array of i, j. Some algebra shows that $\mathbb{E}[\nabla_U L(U,V)]$ can be written as $\gamma z w^T$ for some $\gamma \in \mathbb{R}$ and vector w. As $\mathbb{E}[\nabla_U L(U,V)]$ is a rank one matrix whose range is spanned by z, its orthogonal projection down to z must be 0.

Now we orthogonally decompose $\nabla_U L(U,V)$:

$$\nabla_U L(U,V) = \underbrace{\mathbb{E}[\nabla_U L(U,V)]}_{\nabla_1} + \underbrace{\nabla_U L(U,V) - \mathbb{E}[\nabla_U L(U,V)]}_{\nabla_2}$$

where $\mathrm{Proj}_{z^\perp} \nabla_U L(U,V) = \mathrm{Proj}_{z^\perp} \nabla_2$ by what we have shown. As ∇_2 is a sum of zero-mean random matrices with bounded spectral norm, one for each $(i,j) \in \Omega$, then matrix bound yields

$$\|\nabla_2\| \leqslant \mathcal{O}(\log d/\sqrt{d}) \qquad \text{w.h.p.}$$

Finally, since ∇_1 has no component in $z^\perp$, the entire $z^\perp$-component of $\nabla_U L(U,V)$ arises from ∇_2. Weyl's perturbation theorem implies that a shift of $\mathcal{O}(\log d/\sqrt{d})$ cannot shift much of the gradient outside the z-direction. This completes our proof (fortunately we are not asked to provide exact probability bounds here...) that

$$\|\mathrm{Proj}_{z^\perp} \nabla_U L(U,V)\| \leqslant \mathcal{O}(\log d/\sqrt{d}) \|\nabla_U L(U,V)\|.$$