Vector Calculus Review

1 Proving Derivative Identities

In the notes, we highlight a few important gradients without providing proofs. In this question, we explore why a few of these identities are true. Prove each of the equalities below.

Hint: It is usually easiest to prove these component-by-component; show each component of left-hand-side equals that of the right-hand side. Apply derivative rules when possible to save work.

(a)

$$\frac{\partial}{\partial x} Ax = A$$

(b)

$$\nabla_x \, w^\top x = w$$

(c)

$$\nabla_x \, x^\top A x = (A + A^\top)x$$

(d)

$$\nabla_x \, a^\top x x^\top b A x = \left(ab^\top + ba^\top\right) x$$

2 More Gradient Practice

Next, we consider a few more interesting gradients. Try and take advantage of common derivatives and derivative rules (especially the chain rule) to avoid having to these compute gradients component-by-component when possible.

(a)

$$\nabla_x \, ||Ax - b||_2 + ||x||_2^4$$

(b)

$$\nabla_x \; \mathrm{tr}\left(Axx^\top\right)$$

(c)

$$\nabla_x \; - y^\top \ln x$$

(d)

$$\nabla_w \, y \ln g(x) + (1 - y)\ln(1 - g(x)) \text{ where } g(x) = \frac{1}{1 + e^{-w^\top x}}$$

3 Matrix Derivatives

Now, we extend the definition of the gradient to include derivatives of scalar functions of matrices. For a function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, define the gradient of f with respect to X as the $m \times n$ matrix whose entries correspond to the partials of f with respect to components of X.

$$[\nabla_X f(X)]_{ij} = \frac{\partial f}{\partial X_{ij}},$$

Hint: Compute each component if you have to. The cyclic property of the trace will also be really useful here; whenever you have a scalar function, you can add a trace in front for free and shuffle around the matrices or vectors inside. (This is affectionately called the trace trick.)

(a)

$$\nabla_X \ \mathrm{tr}\left(A^\top X\right)$$

(b)

$$\nabla_X \ a^\top X b$$

(c)

$$\nabla_X \ \|X\|_F^2$$

(d)

$$\nabla_X \ \|AX\|_F^2$$

4 Application: Generalized Tikhonov Regularization

Let $x_1, \ldots, x_n \in \mathbb{R}^d$ be sample points packaged into a design matrix $X \in \mathbb{R}^{n\times d}$. Recall in traditional regularized least squares, we find the weight vector w which minimizes the ℓ^2-distance between the predictions Xw and labels y. In generalized Tikhonov regularization, we instead find w to minimize:

$$f(w; P, Q, W, w_0) = (Xw - y)^\top P(Xw - y) + (w - w_0)^\top Q(w - w_0)$$

where P, Q are positive definite matrices and w_0 is a fixed vector. One interpretation of this objective is that we are interested in considering weighted ℓ_2-distances—the matrices P and Q amplify and suppress certain directions. Given this objective, find a closed-formed solution for the optimal weights

$$w^* = \arg\min_w f(w; P, Q, W, w_0).$$

