Tutorial/Worksheet on "M0b & M0c. Calculus/Optimization & Linear Algebra Background"

CS5691 PRML Jul–Nov 2025

August 27, 2025

1. [Reconcile with high-school geometry]
   (a) In high-school, you would've seen that a line in 2D is defined by the equation $y = mx + c$. Show that this high-school definition is equivalent to the parametric definition of line seen in class (in Slide 9 of "M0b (v2). Background on Calculus & Optimization").
   (Hint: Express m and c in terms of parametric definition's $\mathbf{a}$ and $\mathbf{u}$, and vice versa).
   (b) If dimension $d = 2$, then a 1D line in $\mathbb{R}^d$ and a $(d-1)$-dimensional hyperplane in $\mathbb{R}^d$ refer to the same entity. Show this formally by proving that the definitions of a line and a hyperplane in $\mathbb{R}^2$ are indeed equivalent.
   (c) In high-school, you would've seen that a plane in 3D is defined by the equation $ax+by+cz = d$. Show that this high-school definition is equivalent to the definition of plane seen in class (in Slide 9 of "M0b (v2)").
   (d) Justify why the equation of a hyperplane that goes through a point $a \in \mathbb{R}^d$ and has normal vector w is given by: $w^T(x - a) = 0$.
   (Note: Above equation is equivalent to the high-school equation $a(x-x_0)+b(y-y_0)+c(z-z_0) = 0$ for a plane in 3D that goes through the point (x_0, y_0, z_0) and normal to the vector $[a, b, c]^T$.)

2. [Graph of functions]
   (a) For each of these functions below, plot the graph of the function, and also separately show the contour plots of the function (preferably using pen/paper or using plotting tools like `https://www.wolframalpha.com/` if pen/paper becomes unwieldy). Mention whether each of these functions are convex, concave, both, or neither. If a function is convex (concave), also indicate whether it is strictly convex (concave).
      i. $f(x) = x_1 + x_2$
      ii. $f(x) = x_1^2 + x_2^2$
      iii. $f(x) = x_1^2 - x_2^2$
      iv. $f(x) = -(x_1 - 2)^2 - (x_2 - 1)^2$
      v. $f(x) = x_1 x_2$
   (b) Show that the graph G_f of a linear function $f : \mathbb{R}^d \to \mathbb{R}$ is a d-dimensional hyperplane in $\mathbb{R}^{d+1}$ by showing that:
      i. G_f can be expressed in terms of $\{x \in \mathbb{R}^{d+1} : w^T x = b\}$ for an appropriate choice of w and b, and that
      ii. the dimension of the vector subspace $G_f \subseteq \mathbb{R}^{d+1}$ is d (by providing a basis of cardinality d for this subspace).
      iii. Also show whether the linear function f is convex, concave, both, or neither.
      iv. How will the answers to the above parts change if the function f is affine (i.e., if $f(x) = w^T x + b$ for some $b \neq 0$ instead of being just $f(x) = w^T x$)?
      (Note: For answer to part (ii), we need an extension of a vector space called affine space to show that the linear approximation hyperplane is still d-dimensional in the affine space; this is just for your information and affine space topic will not be tested in the quizzes/exams.)
   (c) Prove: $f(x) = x_1 x_2$ is not (jointly) convex in x_1 and x_2, but is separately convex in x_1 or x_2.

(d) Answer the exercise or example questions in the Appendix of the "M0b (v2)" slides, and also review facts about the existence and uniqueness of minimizers of a convex function in the last Appendix slide of "M0b (v2)".

3. [Linear approximation of a function and maximum rate of gradient descent]

   (a) Find the linear approximation $L_a(f)(x)$ to the function $f(x) = f(x_1, x_2) = 3 + \frac{x_1^2}{16} + \frac{x_2^2}{9}$ at $a = (-4, 3)$.

   (b) Show that the graph of this linear approximation is a hyperplane passing through the point $(a, f(a))$. Specifically express the hyperplane in the form $\{y \in \mathbb{R}^3 : w^T(y - v) = 0\}$ for appropriate choice of the normal vector w to this hyperplane and a point v on the hyperplane.

   (c) We claimed in class that the Gradient $\nabla f(x)$ gives the direction of maximum rate of change of $f(.)$ in the neighborhood of x, and hence Gradient Descent algorithm (taking small steps in the negative direction of the gradient) is an optimal strategy to find the minima of a function.

   Let's formalize and prove the above statement here. Recall the shorthand notation for the linear approximation of a function f as $\Delta f \simeq \nabla f \cdot \Delta x$. Let $\Delta x^* = \arg\min_{\Delta x : \|\Delta x\| = \epsilon} \nabla f \cdot \Delta x$ for some fixed small $\epsilon > 0$. Then prove using Cauchy-Schwarz inequality that $\Delta x^* = -\frac{\epsilon}{\|\nabla f\|} \nabla f$ (i.e., setting $\eta = \frac{\epsilon}{\|\nabla f\|}$ in Gradient Descent update equation $\Delta x = -\eta \nabla f$ gives the most decrease in the function's value for a small change of fixed length ϵ in the function's input).

4. [Links between inverses, symmetricity and positive definiteness]
   Appendix C of CMB book on "Properties of Matrices" may be a useful refresher of certain basic matrix identities. Use appropriate basic matrix identities, definitions of eigenvalues/vectors, and facts about determinant (trace) of matrix being the product (sum) of eigenvalues (with multiplicity) in that Appendix to provide simple concise proofs of the following about a real square matrix $A \in \mathbb{R}^{n \times n}$:

   (a) If A is symmetric, then A^{-1} is symmetric.

   (b) If λ is an eigenvalue of an invertible matrix A with eigenvector x, then $1/\lambda$ is an eigenvalue of A^{-1} with the same eigenvector x.

   (c) Show that the two definitions of positive definite (pd) matrix given below for a symmetric matrix are equivalent (by showing that Defn. 1 implies Defn. 2 and vice versa).
   (Note 1: Pd is naturally defined for symmetric matrices, so pd discussions in this course will be restricted to symmetric matrices.
   Note 2: Spectral Theorem for symmetric matrices can be used for one direction of this proof.)

   Defn. 1: A symmetric matrix A is pd if all its eigenvalues are positive.

   Defn. 2: A symmetric matrix A is pd if for all non-zero vectors x, the quadratic form $x^T A x > 0$.

   (d) If A is pd, then A^{-1} exists and this A^{-1} is also pd.

   (e) If A is pd and B is pd, then $A + B$ is pd (where B is another conformable matrix).

   (f) Write down at least four distinct examples of a 2×2 pd matrix.

5. [Matrix/Vector Derivatives]
   Verify the following identities regarding matrix or vector derivatives. Let $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times n}$. Consider a multivariate function $f(x, A) : \mathbb{R}^n \times \mathbb{R}^{n \times n} \to \mathbb{R}$. Then, the notation $\frac{\partial f}{\partial x}$ is simply the gradient of the function with respect to its input x alone, i.e., $\left(\frac{\partial f}{\partial x}\right)_i = \frac{\partial f}{\partial x_i}$. A similar notation applies when we take derivatives with respect to each matrix entry, i.e., $\left(\frac{\partial f}{\partial A}\right)_{ij} = \frac{\partial f}{\partial A_{ij}}$.

   (a) $\frac{\partial}{\partial x} x^T A x = A^T x + Ax$ (or $2Ax$ if A is symmetric)

   (b) $\frac{\partial}{\partial A} x^T A x = x x^T$ (outer product)

   (c) (Optional Ungraded) $\frac{\partial}{\partial A} \log(\det(A)) = (A^{-1})^T$ (or equivalently $(A^T)^{-1} = A^{-T}$)

6. Let X be a Bivariate Gaussian (BVG) random vector. That is, $X \sim \mathcal{N}(\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}, \begin{bmatrix} a & b \\ c & d \end{bmatrix})$. What are the permissible/legal values of the parameters a, b, c, d so that the BVG distribution of X has a well-defined probability density function (and is non-degenerate)?

7. Answer the exercises or homework questions in the different slides of "M0c" slide deck on Linear Algebra background.