DSA5103 Exercises

Zhao Peiduo

AY2024/25 Sem2

Disclaimer: Answers could be based on ChatGPT results which may be inaccurate / incorrect.

Lecture 1: Introduction and PCA

1 The PCA Optimization Problem

We can decompose the covariance matrix Σ as:

$$\Sigma = Q_{\text{full}} \Lambda Q_{\text{full}}^T$$

where:

- Q_{full} is an orthogonal matrix whose columns $q_1, \ldots, q_p$ are the eigenvectors of Σ.
- Λ is a diagonal matrix containing the eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$.

Thus, we write:

$$\Sigma = \begin{bmatrix} q_1 & \cdots & q_p \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_p \end{bmatrix} \begin{bmatrix} q_1 & \cdots & q_p \end{bmatrix}^T$$

The optimization problem becomes:

$$\max_{Q^T Q = I} \text{trace}(Q^T \Sigma Q)$$

Substituting the eigenvalue decomposition of Σ:

$$\text{trace}(Q^T \Sigma Q) = \text{trace}\left(Q^T Q_{\text{full}} \Lambda Q_{\text{full}}^T Q\right)$$

Define $\tilde{Q} = Q_{\text{full}}^T Q$, which is an orthogonal transformation, leading to:

$$\text{trace}(\tilde{Q}^T \Lambda \tilde{Q})$$

Since Λ is a diagonal matrix, the maximum value of the trace is obtained when $\tilde{Q}$ selects the top k eigenvectors corresponding to the largest eigenvalues.

The optimal choice for Q that maximizes the trace is to select the eigenvectors corresponding to the largest k eigenvalues:

$$Q = \begin{bmatrix} q_1 & \cdots & q_k \end{bmatrix}$$

Thus, the optimization problem reduces to selecting the top k eigenvectors of Σ, and the optimal solution is:

$$\text{trace}(Q^T \Sigma Q) = \lambda_1 + \lambda_2 + \cdots + \lambda_k$$

The solution to the PCA optimization problem is obtained by selecting the top k eigenvectors of the covariance matrix Σ, forming the matrix Q as:

$$Q = \begin{bmatrix} q_1 & \cdots & q_k \end{bmatrix}$$

This choice ensures that the projection of the data preserves the maximum variance.

2 Standardization Does Not Change the Covariance Matrix

The sample covariance matrix of the original data matrix X (assuming it has zero mean) is given by:

$$\Sigma = \frac{1}{n} X^T X$$

where:

- X is an $n \times p$ data matrix (with rows as observations and columns as variables),
- Σ is the $p \times p$ covariance matrix.

Standardizing each column of X results in:

$$Z_{\cdot j} = \frac{X_{\cdot j}}{\sigma_j}$$

where σ_j is the standard deviation of the j-th column of X, and Z represents the standardized data matrix.
The standardized data matrix Z can be written in matrix form as:

$$Z = XD^{-1}$$

where D is a diagonal matrix containing the standard deviations of each column:

$$D = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_p)$$

The covariance matrix of the standardized data Z is:

$$\Sigma_Z = \frac{1}{n} Z^T Z$$

Substituting $Z = XD^{-1}$:

$$\Sigma_Z = \frac{1}{n} (XD^{-1})^T (XD^{-1})$$

Expanding the terms:

$$\Sigma_Z = D^{-1} \left(\frac{1}{n} X^T X \right) D^{-1}$$

Since $\Sigma = \frac{1}{n} X^T X$, we obtain:

$$\Sigma_Z = D^{-1} \Sigma D^{-1}$$

The diagonal scaling by D^{-1} standardizes the variance of each variable to 1, ensuring that the resulting covariance matrix of the standardized data is:

$$\Sigma_Z = I$$

This shows that standardization converts the covariance matrix into the identity matrix, preserving the correlation structure. However, the relative relationships between variables (correlations) remain unchanged.
Since standardization only rescales variables without affecting their relationships, the original covariance matrix structure is preserved. Thus, standardization does not change the fundamental covariance relationships among variables, only their magnitudes.

Lecture 2: Gradient (Descent) Methods and Linear Regression

3 Convergence of Gradient Descent

Question: Let A be a Symmetric Positive Definite (SPD) matrix.Then, the A-norms of the error vectors $d_k = x_\star - x_k = -A^{-1}r_k$. The error vectors in the gradient descent algorithm satisfy the relation:

$$\|d_{k+1}\|_A \leq \frac{\lambda_{\max} - \lambda_{\min}}{\lambda_{\max} + \lambda_{\min}} \|d_k\|_A.$$

This implies that as $k \to \infty$, we have $\|d_k\|_A \to 0$, which further implies that $d_k \to 0$, ensuring the convergence of the gradient descent method from any initial guess x_0.

Proof: First, we have the A-norm squared of d_k:

$$\|d_k\|_A^2 = (Ad_k, d_k) = (-r_k, d_k) = (r_k, A^{-1}r_k).$$

Exercise 1: Explain how we get $\|d_{k+1}\|_A^2$: substitute k with $k+1$ we get $(Ad_{k+1}, d_{k+1}) = (-r_{k+1}, d_{k+1})$.

Substituting $d_{k+1} = d_k + \alpha_k r_k$ and expanding:

$$\|d_{k+1}\|_A^2 = (-r_{k+1}, d_k + \alpha_k r_k).$$

Using the inner product linearity, we rewrite it as:

$$\|d_{k+1}\|_A^2 = (-r_{k+1}, d_k) - \alpha_k (r_{k+1}, r_k).$$

Exercise 2: Show that $(r_{k+1}, r_k) = 0$
Since $x_{k+1} = x_k - \alpha_k r_k$,

$$r_{k+1} = Ax_{k+1} - b = A(x_k - \alpha_k r_k) - b.$$
$$r_{k+1} = Ax_k - \alpha_k Ar_k - b = r_k - \alpha_k Ar_k.$$

Therefore:

$$(r_{k+1}, r_k) = (r_k - \alpha_k Ar_k, r_k).$$
$$(r_{k+1}, r_k) = (r_k, r_k) - \alpha_k (Ar_k, r_k).$$

Since the step size in gradient descent is chosen as:

$$\alpha_k = \frac{(r_k, r_k)}{(Ar_k, r_k)},$$

we substitute α_k:

$$(r_{k+1}, r_k) = (r_k, r_k) - \frac{(r_k, r_k)}{(Ar_k, r_k)}(Ar_k, r_k) = 0 \quad \text{(proven)}$$

We now want to show that

$$\|d_{k+1}\|_A^2 = \|d_k\|_A^2 \left(1 - \frac{(r_k, r_k)}{(Ar_k, r_k)} \times \frac{(r_k, r_k)}{(r_k, A^{-1}r_k)}\right).$$

Proof: Starting from the definition:

$$\|d_{k+1}\|_A^2 = (-r_{k+1}, d_k)$$

Expanding r_{k+1} using the update formula:

$$\|d_{k+1}\|_A^2 = (-r_k + \alpha_k Ar_k, d_k)$$
$$\|d_{k+1}\|_A^2 = (-r_k, d_k) + \alpha_k (Ar_k, d_k)$$

Using previously derived identities:

$$(-r_k, d_k) = (r_k, A^{-1}r_k), \quad (Ar_k, d_k) = (Ar_k, A^{-1}r_k) = (r_k, r_k)$$
$$\|d_{k+1}\|_A^2 = (r_k, A^{-1}r_k) - \alpha_k (r_k, r_k)$$
$$\|d_{k+1}\|_A^2 = (r_k, A^{-1}r_k) - \frac{(r_k, r_k)}{(Ar_k, r_k)}(r_k, r_k)$$

Exercise 3: Recognize that $\|d_k\|_A^2 = (r_k, A^{-1}r_k)$ and factor out $\|d_k\|_A^2$:

$$\|d_{k+1}\|_A^2 = \|d_k\|_A^2 \left(1 - \frac{(r_k, r_k)}{(Ar_k, r_k)} \times \frac{(r_k, r_k)}{(r_k, A^{-1}r_k)}\right)$$

4 Steepest Descent Method with Exact Line Search Example 5

Question: We are given the problem:

$$\min_{x=(x_1,x_2)\in\mathbb{R}^2} f(x) = x_1^2 + x_2^2 + 2x_1 + 4.$$

This is a convex optimization problem.
Applying the steepest descent method with exact line search, starting at $x^{(0)} = [2, 1]$, we proceed as follows.
Solution: The gradient is computed as:

$$\nabla f(x) = \begin{bmatrix} 2x_1 + 2 \\ 2x_2 \end{bmatrix}, \quad \nabla f(x^{(0)}) = \begin{bmatrix} 6 \\ 2 \end{bmatrix}.$$

The search direction is given by:

$$p^{(0)} = -\nabla f(x^{(0)}) = [-6, -2].$$

To find the step size α_0, we minimize:

$$\phi(\alpha) = f(x^{(0)} + \alpha p^{(0)}) = f(2 - 6\alpha, 1 - 2\alpha).$$

Expanding:

$$\phi(\alpha) = (2 - 6\alpha)^2 + (1 - 2\alpha)^2 + 2(2 - 6\alpha) + 4.$$

Since $\phi(\alpha)$ is quadratic (convex), setting $\phi'(\alpha) = 0$ gives $\alpha_0 = 0.5$.
Thus, the next iterate is:

$$x^{(1)} = x^{(0)} + \alpha_0 p^{(0)} = [-1, 0].$$

Since $\nabla f(x^{(1)}) = [0, 0]$, the method terminates.
Exercise: Verify that $[-1, 0]$ is a global minimizer by sufficient condition and convexity.
Solution: The function $f(x)$ is convex since its Hessian matrix:

$$H = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

is positive definite (all eigenvalues are positive). A stationary point where $\nabla f(x) = 0$ in a convex function is a global minimizer. Since $\nabla f(x^{(1)}) = 0$, we conclude that $x^{(1)} = [-1, 0]$ is the global minimum.

5 Steepest Descent Method with Exact Line Search Example 6

Question: We now minimize:

$$\min_{x=(x_1,x_2)\in\mathbb{R}^2} f(x) = x_1^2 + 2x_2^2 - 2x_1x_2 - 2x_2.$$

This function is convex.
Applying the steepest descent method with exact line search, starting at $x^{(0)} = [0, 0]$, we compute $x^{(1)}, x^{(2)}, x^{(3)}$.
Solution:

- Compute the gradient:
$$\nabla f(x) = \begin{bmatrix} 2x_1 - 2x_2 \\ 4x_2 - 2x_1 - 2 \end{bmatrix}.$$
Evaluating at $x^{(0)} = [0, 0]$:
$$\nabla f(x^{(0)}) = \begin{bmatrix} 0 \\ -2 \end{bmatrix}.$$
- Compute the search direction:
$$p^{(0)} = -\nabla f(x^{(0)}) = \begin{bmatrix} 0 \\ 2 \end{bmatrix}.$$
- Compute step size α_0:
$$\alpha_0 = \frac{1}{2}.$$
- Compute the next iterate:
$$x^{(1)} = x^{(0)} + \alpha_0 p^{(0)} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$
- Compute the next descent direction:
$$p^{(1)} = -\nabla f(x^{(1)}) = \begin{bmatrix} 1 \\ 0 \end{bmatrix}.$$
- Compute step size α_1:
$$\alpha_1 = \frac{1}{2}.$$
- Compute the next iterate:
$$x^{(2)} = x^{(1)} + \alpha_1 p^{(1)} = \begin{bmatrix} \frac{1}{2} \\ 1 \end{bmatrix}.$$
- Compute the next descent direction:
$$p^{(2)} = -\nabla f(x^{(2)}) = \begin{bmatrix} 0 \\ \frac{1}{2} \end{bmatrix}.$$
- Compute step size α_2:
$$\alpha_2 = \frac{1}{2}.$$
- Compute the next iterate:
$$x^{(3)} = x^{(2)} + \alpha_2 p^{(2)} = \begin{bmatrix} \frac{1}{2} \\ \frac{3}{4} \end{bmatrix}.$$

Since the gradient at $x^{(3)}$ is close to zero, the method converges. The global minimizer is $x^* = [1, 1]$.

Lecture 4: Proximal Operator and (Accelerated) Proximal Gradient Methods

6 Projection onto the Positive Orthant

Question: Let $C = \mathbb{R}^n_+ = \{x \in \mathbb{R}^n \mid x_i \geq 0,\ i = 1, \ldots, n\}$ (the positive orthant in $\mathbb{R}^n$). Given any $z \in \mathbb{R}^n$, determine $\Pi_C(z)$, the projection of z onto C, and prove that

$$[\Pi_C(z)]_i = \max\{\, z_i,\, 0\,\}, \quad i = 1, \ldots, n.$$

Relevant Concepts: The projection $\Pi_C(z)$ is by definition the unique vector $x^* \in C$ that minimizes the Euclidean distance $\|x - z\|_2$ over all $x \in C$. Because the objective $\|x - z\|_2^2 = \sum_{i=1}^n (x_i - z_i)^2$ separates across coordinates, this minimization can be done for each component independently. In other words, x^* satisfies, for each i,

$$x_i^* = \arg\min_{y \geq 0} (y - z_i)^2.$$

Proof: For a fixed index i, consider the one-dimensional problem $\min_{y \geq 0}(y - z_i)^2$. This is a convex quadratic function in y. If $z_i \geq 0$, the minimum occurs at $y = z_i$ (since the unconstrained minimizer is $y = z_i$, which is feasible). If $z_i < 0$, the unconstrained minimizer $y = z_i$ is not feasible under $y \geq 0$, so the constrained minimum occurs at the boundary $y = 0$. In summary, the minimizer is $y = \max\{z_i,\, 0\}$. Because the full projection problem splits into n independent scalar problems, the ith component of the projection is given by $x_i^* = \max\{z_i, 0\}$. Thus $\Pi_C(z) = \max\{z, 0\}$ (interpreted componentwise), which is exactly the claimed formula.

Finally, to verify that this x^* indeed is the projection of z onto C, we note two things: (1) by construction $x^* \in C$ and x^* achieves the smallest squared distance coordinatewise, hence globally minimal distance $\|x^* - z\|$; (2) the projection onto a closed convex set is unique (Projection Theorem), so this choice is the unique projection. Therefore $x^* = \Pi_C(z)$ and $x_i^* = \max\{z_i, 0\}$ for each i, as required.

7 Projection onto an ℓ_2-Norm Ball

Question: Let $C = \{x \in \mathbb{R}^n \mid \|x\|_2 \leq 1\}$ be the unit ℓ_2-norm ball in $\mathbb{R}^n$. Given any $z \in \mathbb{R}^n$, find the projection $\Pi_C(z)$ and prove that

$$\Pi_C(z) = \begin{cases} z, & \text{if } \|z\|_2 \leq 1, \\ \frac{z}{\|z\|_2}, & \text{if } \|z\|_2 > 1 \ , \end{cases}$$

i.e. if z lies outside the unit ball, the projection is the point on the ball in the direction of z.

Relevant Concepts: The projection onto a ball can be derived using symmetry and Lagrange multipliers. Intuitively, if $\|z\|_2 > 1$, the best way to minimize $\|x - z\|_2$ under the constraint $\|x\|_2 \leq 1$ is to choose x on the boundary ($\|x\|_2 = 1$) in the exact direction of z (since the Euclidean norm is radially symmetric). More formally, one can solve the constrained minimization

$$\min_x \frac{1}{2}\|x - z\|_2^2 \quad \text{s.t. } \|x\|_2 \leq 1$$

using the Karush–Kuhn–Tucker (KKT) conditions.

Proof: If $\|z\|_2 \leq 1$, then $z \in C$ already lies in the ball, and the closest point in C to z is trivially z itself. Now assume $\|z\|_2 > 1$. In this case, the optimal $x^* = \Pi_C(z)$ must lie on the boundary $\{x : \|x\|_2 = 1\}$, because if x^* were inside the ball, we could scale it outward to achieve a point on the boundary that is closer to z, contradicting the optimality of x^*. By symmetry of the Euclidean norm, it is clear that the minimizing x^* should be chosen in the direction of z. Let $x^* = \alpha \frac{z}{\|z\|_2}$ for some scalar $\alpha \geq 0$. The constraint $\|x^*\|_2 = 1$ forces $\alpha = 1$. Thus $x^* = \frac{z}{\|z\|_2}$, which is the intersection of the ray from the origin through z with the unit sphere. This x^* is feasible and clearly minimizes the distance to z (any other point on the sphere has a larger angular deviation from z or is further along the ray, hence further from z). To be rigorous, one can apply the KKT conditions: the Lagrangian for the constrained problem is $L(x, \lambda) = \frac{1}{2}\|x - z\|_2^2 + \lambda(\|x\|_2^2 - 1)$. At optimum, $\nabla_x L = x - z + 2\lambda x = 0$. For $\|z\|_2 > 1$, complementary slackness implies $\lambda > 0$ and $\|x^*\|_2 = 1$. Thus $(1 + 2\lambda)x^* = z$. Taking norms on both sides: $(1 + 2\lambda)\|x^*\|_2 = \|z\|_2$ which gives $1 + 2\lambda = \|z\|_2$ (since $\|x^*\|_2 = 1$). Therefore $x^* = \frac{z}{\|z\|_2}$, as claimed.

8 Projection onto the Positive Semidefinite Cone

Question: Let $C = S^n_+ = \{A \in \mathbb{R}^{n\times n} \mid A = A^T,\ A \succeq 0\}$ be the cone of $n \times n$ symmetric positive semidefinite matrices. Given any symmetric matrix $A \in \mathbb{R}^{n\times n}$ with eigen-decomposition $A = Q\,\mathrm{diag}(\lambda_1, \ldots, \lambda_n)Q^T$ (where Q is orthonormal and $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of A), prove that

$$\Pi_{S^n_+}(A) = Q\ \mathrm{diag}(\max\{\lambda_1, 0\},\ \max\{\lambda_2, 0\},\ \ldots,\ \max\{\lambda_n, 0\})\, Q^T\ .$$

Relevant Concepts: The Frobenius norm $\|X\|_F = \sqrt{\langle X, X\rangle}$ (where $\langle X, Y\rangle = \mathrm{Tr}(X^TY)$) is the natural extension of the Euclidean norm to matrices. It is unitarily invariant, meaning $\|A - B\|_F = \|Q^T(A - B)Q\|_F$ for any orthonormal Q. A key fact from linear algebra is that any symmetric matrix A can be orthogonally diagonalized. Intuitively, to project A onto the cone S^n_+, one can "clamp" all negative eigenvalues of A to 0, since negative directions are not allowed in S^n_+ and adjusting any diagonal element outside of the eigenbasis would only increase the distance. Another useful observation is that for any fixed orthonormal basis, the closest diagonal matrix to a given diagonal matrix is obtained by matching each diagonal entry as closely as possible – here that means keeping each λ_i if it is nonnegative, or replacing it with 0 if it is negative.

Proof: Let $A = Q\Lambda Q^T$ be the eigendecomposition of A, where $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$. We seek a matrix $X \succeq 0$ minimizing $\|X - A\|_F^2 = \|X - Q\Lambda Q^T\|_F^2$. Because the Frobenius norm is invariant under orthonormal change of basis, we have $\|X - A\|_F = \|Q^TXQ - \Lambda\|_F$. Define $B = Q^TXQ$; since X is constrained to be symmetric PSD, B must be symmetric PSD as well (because $B = Q^TXQ$ and $X \succeq 0$ implies $B \succeq 0$). Now the problem becomes: minimize $\|B - \Lambda\|_F^2$ subject to B is PSD. Among all PSD choices of B, the quantity $\|B - \Lambda\|_F^2$ is minimized when B shares the same eigenbasis as Λ (this follows from the fact that any deviation of B's eigenvectors from those of Λ will introduce off-diagonal differences that strictly increase the Frobenius norm). Therefore, we can restrict attention to B that is diagonal in the same basis as Λ, i.e. take $B = \mathrm{diag}(\mu_1, \ldots, \mu_n)$ (with respect to the fixed basis that diagonalizes A). Now the problem further reduces to minimizing $\sum_{i=1}^n (\mu_i - \lambda_i)^2$ subject to $\mu_i \geq 0$ for each i. This is n independent scalar problems, exactly analogous to the projection onto the nonnegative ray considered earlier. By the same reasoning as in the projection onto $\mathbb{R}_+$ case, the minimizer is $\mu_i^* = \max\{\lambda_i,\ 0\}$ for each i. Thus the optimal $B^* = \mathrm{diag}(\max\{\lambda_1, 0\}, \ldots, \max\{\lambda_n, 0\})$. Transforming back to the original coordinates, $X^* = QB^*Q^T$. This X^* is symmetric and PSD (since B^* has no negative entries on the diagonal) and is exactly the stated formula.
Finally, we argue that this X^* indeed is the projection. By construction, X^* is PSD and achieves the minimum Frobenius distance to A. Moreover, if Y is any other PSD matrix, we have

$$\|A - Y\|_F^2\ =\ \|Q^TAQ - Q^TYQ\|_F^2\ =\ \|\Lambda - (Q^TYQ)\|_F^2\ .$$

Since Q^TYQ is PSD, it has some eigen-decomposition $U\Gamma U^T$ with $\Gamma_{ii} \geq 0$. But if $U \neq I$, then $U\Gamma U^T$ has off-diagonal entries in the Λ basis. Those off-diagonals contribute positively to the Frobenius norm difference (because Λ is diagonal). Thus any deviation from $U = I$ can only increase the distance. Therefore, without loss of generality we may assume Y shares the eigenvectors with A, i.e. $Y = Q\tilde{\Gamma}Q^T$ for some diagonal $\tilde{\Gamma}$ with nonnegative entries. Now

$$\|A - Y\|_F^2 = \sum_{i=1}^n (\lambda_i - \tilde{\Gamma}_{ii})^2 \geq \sum_{i=1}^n (\lambda_i - \mu_i^*)^2 = \|A - X^*\|_F^2,$$

since for each i, $\mu_i^* = \max\{\lambda_i, 0\}$ minimizes $(\lambda_i - \mu_i)^2$ over $\mu_i \geq 0$. This shows no PSD matrix can be closer to A than X^*. Hence $X^* = \Pi_{S^n_+}(A)$, as required.

9 Equivalence of Convexity Definitions

Question: Let $f : X \to (-\infty, +\infty]$ be a function, where X is a vector space (e.g. $\mathbb{R}^n$). Assume f is proper (not identically $+\infty$). We consider two definitions of convexity for f:

(1) epigraph of f is convex: $\mathrm{epi}(f) = \{(x, \alpha) \mid f(x) \leq \alpha\}$ is a convex set in $X \times \mathbb{R}$.

(2) f is convex: for all $x, y \in \mathrm{dom}(f)$ and all $\lambda \in [0, 1]$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda)f(y) \ .$$

Prove that (1) and (2) are equivalent statements, i.e. f is convex if and only if its epigraph is convex.

Relevant Concepts: The epigraph $\mathrm{epi}(f)$ being convex means: if (x_1, α_1) and (x_2, α_2) satisfy $\alpha_i \geq f(x_i)$, then for any $0 \leq \lambda \leq 1$, the point $(\lambda x_1 + (1-\lambda)x_2,\ \lambda\alpha_1 + (1-\lambda)\alpha_2)$ also lies in $\mathrm{epi}(f)$, which means

$$\lambda\alpha_1 + (1-\lambda)\alpha_2 \geq f(\lambda x_1 + (1-\lambda)x_2) \ .$$

Meanwhile, the usual definition of convexity of f (inequality (2) above) can be rewritten as

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda)f(x_2) \ .$$

To prove equivalence, one typically shows (1) ⇒ (2) and (2) ⇒ (1) by starting from one inequality and manipulating it into the form of the other.

Proof:

(1) $\implies$ (2): Assume $\mathrm{epi}(f)$ is convex. Take any $x_1, x_2 \in \mathrm{dom}(f)$ (so $f(x_i) < +\infty$) and $0 \leq \lambda \leq 1$. Let $\alpha_1 = f(x_1)$ and $\alpha_2 = f(x_2)$. Then (x_1, α_1) and (x_2, α_2) are in $\mathrm{epi}(f)$. By convexity of the epigraph, the point

$$(\bar{x}, \bar{\alpha}) := (\lambda x_1 + (1-\lambda)x_2,\ \lambda\alpha_1 + (1-\lambda)\alpha_2)$$

must also lie in $\mathrm{epi}(f)$. By the definition of epigraph membership, this means $\bar{\alpha} \geq f(\bar{x})$, i.e.

$$\lambda f(x_1) + (1-\lambda)f(x_2) \ = \ \bar{\alpha} \ \geq \ f(\bar{x}) \ = \ f(\lambda x_1 + (1-\lambda)x_2) \ .$$

This is exactly the inequality required in (2), so f satisfies the Jensen convexity condition.

(2) $\implies$ (1): Now assume f satisfies (2). We want to show $\mathrm{epi}(f)$ is convex. Take any two points $(x_1, \alpha_1), (x_2, \alpha_2) \in \mathrm{epi}(f)$. By definition of epigraph, this means $\alpha_1 \geq f(x_1)$ and $\alpha_2 \geq f(x_2)$. Consider any $0 \leq \lambda \leq 1$ and form the point $(\bar{x}, \bar{\alpha})$ as above. We need to show $(\bar{x}, \bar{\alpha}) \in \mathrm{epi}(f)$, i.e. $\bar{\alpha} \geq f(\bar{x})$.

Since $\alpha_1 \geq f(x_1)$ and $\alpha_2 \geq f(x_2)$, we have

$$\bar{\alpha} \ = \ \lambda\alpha_1 + (1-\lambda)\alpha_2 \ \geq \ \lambda f(x_1) + (1-\lambda)f(x_2) \ .$$

Now apply the convexity condition (2) to the points x_1, x_2. Using (2) we get

$$f(\bar{x}) = f(\lambda x_1 + (1-\lambda)x_2) \ \leq \ \lambda f(x_1) + (1-\lambda)f(x_2) \ \leq \ \bar{\alpha} \ .$$

Thus $\bar{\alpha} \geq f(\bar{x})$. This proves $(\bar{x}, \bar{\alpha}) \in \mathrm{epi}(f)$, establishing that the epigraph is convex.
Having shown both directions, we conclude that (1) and (2) are indeed equivalent: a proper function f is convex if and only if its epigraph is a convex set.

10 Normal Cone Computations for Specific Sets

Question: The normal cone of a set C at a point x is defined as

$$N_C(x) := \{z \in X \mid \langle z, y - x\rangle \le 0, \ \forall\, y \in C\},$$

i.e. the set of all vectors z that form an acute angle (or are orthogonal) with every direction from x into C. Consider the following specific convex sets and points:

(1) $C = \{x \in \mathbb{R}^2 \mid x_1 + x_2 \le 1, \ x_1 \ge 0, \ x_2 \ge 0\}$ (a right triangle in $\mathbb{R}^2$ with vertices at $(0,0), (1,0), (0,1)$). For x at the points $x^{(a)} = (0,1)$, $x^{(b)} = (0.5, 0.5)$, and $x^{(c)} = (0.1, 0.2)$, find $N_C(x)$.

(2) $C = \mathbb{R}^n_+$ (the nonnegative orthant in $\mathbb{R}^n$). For a point $x \in C$ (specifically $x = (1, 1, 0, 0, \ldots, 0)$, which has first two coordinates positive and the rest zero), describe $N_C(x)$.

Relevant Concepts: The normal cone $N_C(x)$ describes the set of outward normal vectors to C at x. If x lies in the interior of C, $N_C(x) = \{0\}$ since no nonzero vector can be a normal to C at an interior point (every direction is allowable inside C). If x lies on a boundary face of C, then $N_C(x)$ consists of all combinations of normals to each of the constraints active at x. For a polyhedral set described by linear inequalities, the normal cone at a boundary point is generated by the gradients of the active constraint inequalities (with nonnegative coefficients). In $\mathbb{R}^2$, this geometric picture can be visualized: at a corner of a polygon, the normal cone is the cone spanned by the outward normals to the edges meeting at that corner; at a point on a flat edge, the normal cone is just the line (ray) perpendicular to that edge; at an interior point, the normal cone is trivial $\{0\}$.

Proof:

(1) For the triangle C: the three defining inequalities are $x_1 + x_2 \le 1$, $x_1 \ge 0$, and $x_2 \ge 0$. The outward normals (pointing outside C) to these constraint boundaries are respectively $n^{(1)} = (1,1)$ (for $x_1 + x_2 = 1$), $n^{(2)} = (-1, 0)$ (for $x_1 = 0$), and $n^{(3)} = (0, -1)$ (for $x_2 = 0$). Now:

- At $x^{(a)} = (0,1)$: This is the vertex where $x_1 = 0$ and $x_1 + x_2 = 1$ are active (both constraints hold with equality at this point). The $x_2 \ge 0$ constraint is not active (since $x_2 = 1 > 0$). So the normal cone $N_C(0,1)$ is the cone generated by $n^{(1)}$ and $n^{(2)}$. Therefore:

  $$N_C(0,1) = \{\alpha(-1,0) + \beta(1,1) \mid \alpha \ge 0, \beta \ge 0\}.$$

  In simpler terms, $N_C(0,1)$ consists of all vectors whose second component is nonnegative (coming from $\beta(1,1)$) and whose first component is $\beta - \alpha$. We can just describe it as the cone spanned by $(-1,0)$ and $(1,1)$.

- At $x^{(b)} = (0.5, 0.5)$: Here $x_1 + x_2 = 1$ is active (since $0.5 + 0.5 = 1$), but $x_1 = 0.5 > 0$ and $x_2 = 0.5 > 0$ so the constraints $x_1 \ge 0$ and $x_2 \ge 0$ are not active. Thus $x^{(b)}$ lies in the relative interior of the hypotenuse of the triangle. The only active boundary is the line $x_1 + x_2 = 1$, whose outward normal is $(1,1)$. Therefore $N_C(0.5, 0.5)$ is the ray in direction $(1,1)$:

  $$N_C(0.5, 0.5) = \{\gamma(1,1) \mid \gamma \ge 0\}.$$

- At $x^{(c)} = (0.1, 0.2)$: This point lies strictly inside C (since $0.1 + 0.2 = 0.3 < 1$ and $0.1 > 0$, $0.2 > 0$). Therefore $x^{(c)}$ is an interior point of C. The normal cone at an interior point is $\{0\}$, because the conditions $\langle z, y - x^{(c)}\rangle \le 0$ for all $y \in C$ force $z = 0$ when $x^{(c)}$ can move in any direction within C. In other words:

  $$N_C(0.1, 0.2) = \{0\}.$$

(2) For $C = \mathbb{R}^n_+$ and $x = (1, 1, 0, 0, \ldots, 0)$: This x lies on some of the boundary faces of C: specifically, for coordinates $i = 3, 4, \ldots, n$, we have $x_i = 0$ (so the constraints $x_i \ge 0$ are active at equality), whereas for $i = 1, 2$, we have $x_i > 0$ (so $x_1 \ge 0$ and $x_2 \ge 0$ are not active as equalities). Thus x is on the boundary faces corresponding to the hyperplanes $x_3 = 0, x_4 = 0, \ldots, x_n = 0$, but not on the faces $x_1 = 0$ or $x_2 = 0$. The outward normal for the face $x_i = 0$ is the negative unit vector in the ith coordinate direction, $-e_i$ (where e_i is the standard basis vector). Therefore, by the polyhedral normal cone rule,

$$N_C(x) = \Big\{ \sum_{i=3}^{n} \lambda_i(-e_i) \,\Big|\, \lambda_i \ge 0 \Big\},$$

with the condition that there are no components in the e_1 or e_2 directions because those constraints are not active. Equivalently,

$$N_C(1,1,0,0,\ldots,0) = \{(0,0,\nu_3,\nu_4,\ldots,\nu_n) \mid \nu_3,\nu_4,\ldots,\nu_n \leq 0\},$$

the set of vectors whose first two components are zero and whose remaining components are non-positive. Indeed, for any $z = (z_1,\ldots,z_n)$ in this set and any $y = (y_1,\ldots,y_n) \in C$, we have:

$$\langle z,\, y - x\rangle = z_1(y_1 - 1) + z_2(y_2 - 1) + \sum_{i=3}^{n} z_i(y_i - 0).$$

Since $z_1 = 0, z_2 = 0$ and each $z_i \leq 0$ for $i \geq 3$ while $y_i \geq 0$, this sum is $\sum_{i=3}^{n} z_i y_i \leq 0$. Thus z satisfies the normal cone inequality. Conversely, if a vector has a positive component in an index where $x_i = 0$, then taking y moving in that coordinate will violate the inequality, and if it has a nonzero component in coordinates 1 or 2, one can move y slightly in that direction (which is allowable since $x_1, x_2 > 0$) to make the inner product positive, forcing that component to be zero. So the description above indeed captures all of $N_C(x)$.

11 Normal Cone is Closed (and Convex)

Question: Let $C \subseteq X$ be a nonempty convex set in a Euclidean space X, and $x \in C$. Prove that the normal cone $N_C(x)$ is a closed (and convex) set.

Relevant Concepts: The normal cone was defined as

$$N_C(x) = \{z \mid \langle z, y - x \rangle \leq 0, \ \forall y \in C\}.$$

This can be viewed as the intersection of an (often infinite) family of closed halfspaces: for each fixed $y \in C$, consider the set $H_y = \{z \mid \langle z, y - x \rangle \leq 0\}$. Each H_y is a closed halfspace in the dual space (because the inequality describes a closed condition). Then $N_C(x) = \bigcap_{y \in C} H_y$. The intersection of an arbitrary collection of closed sets is closed (this is a basic topological fact). Also, note that each H_y is convex, and an intersection of convex sets is convex, so $N_C(x)$ is convex as well.

Proof: For any fixed $y \in C$, the set $H_y = \{z \mid \langle z, y - x \rangle \leq 0\}$ is closed in X (the function $z \mapsto \langle z, y - x \rangle$ is continuous and H_y is the preimage of the closed ray $(-\infty, 0]$). Now $N_C(x)$ can be written as

$$N_C(x) = \bigcap_{y \in C} H_y \ .$$

An arbitrary intersection of closed sets remains closed (because if a sequence $z^{(k)}$ lies in every H_y and converges to some z, then for each y, the limit $\langle z, y - x \rangle = \lim_{k \to \infty} \langle z^{(k)}, y - x \rangle \leq 0$, so $z \in H_y$ for all y, hence $z \in \bigcap_y H_y$). Therefore $N_C(x)$ is closed.
Moreover, each H_y is clearly a convex set (since the defining inequality is linear in z). The intersection of any collection of convex sets is convex. Hence $N_C(x)$ is convex as well. In summary, $N_C(x)$ is a closed convex cone.

12 Subgradient of an Indicator Function equals Normal Cone

Question: Let $C \subseteq X$ be a convex set and consider its indicator function

$$\delta_C(x) = \begin{cases} 0, & x \in C, \\ +\infty, & x \notin C \ . \end{cases}$$

Show that the subdifferential of δ_C at a point x is exactly the normal cone to C at x. In notation:

$$\partial \delta_C(x) = N_C(x) \ ,$$

for every $x \in X$ (and note if $x \notin C$, then $\partial \delta_C(x) = \emptyset$ by definition since x is outside the domain of δ_C).

Relevant Concepts: The subgradient of a convex function f at x is defined as

$$\partial f(x) = \{v \mid f(z) \geq f(x) + \langle v, z - x\rangle, \ \forall z \in X\} \ .$$

For the indicator function δ_C, if $x \in C$, then $\delta_C(x) = 0$. For any $z \in X$, $\delta_C(z) \geq \delta_C(x) + \langle v, z - x\rangle$ translates to: - If $z \in C$, then $\delta_C(z) = 0$, so the inequality becomes $0 \geq 0 + \langle v, z - x\rangle$, i.e. $\langle v, z - x\rangle \leq 0$ for all $z \in C$. - If $z \notin C$, then $\delta_C(z) = +\infty$ and the inequality holds trivially (since the left side is $+\infty$ and the right side is finite). Thus $\partial \delta_C(x)$ consists of all v such that $\langle v, z - x\rangle \leq 0$ for all $z \in C$ (when $x \in C$). But this is exactly the definition of $N_C(x)$. If $x \notin C$, then $\partial \delta_C(x)$ is empty by the usual convention that we only consider subgradients at points in the domain (where $f(x)$ is finite).

Proof: First consider $x \in C$. By definition:

$$\begin{aligned} v \in \partial \delta_C(x) &\iff \delta_C(z) \geq \delta_C(x) + \langle v, \, z - x\rangle \quad \forall z \in X \ . \\ &\iff \text{For all } z \in X : \begin{cases} z \in C & \implies 0 \geq 0 + \langle v, z - x\rangle, \\ z \notin C & \implies +\infty \geq 0 + \langle v, z - x\rangle \ , \end{cases} \end{aligned}$$

where we used $\delta_C(x) = 0$ since $x \in C$. The second condition (when $z \notin C$) is automatically true for any v, because $+\infty \geq$ (finite number) is always satisfied. The first condition simplifies to $\langle v, z - x\rangle \leq 0$ for all $z \in C$. But this is exactly the defining condition for $v \in N_C(x)$. Therefore $\partial \delta_C(x) = N_C(x)$ for $x \in C$.

If $x \notin C$, then $\delta_C(x) = +\infty$ and by convention $\partial \delta_C(x) = \emptyset$ (since x is not in the effective domain of δ_C where the subgradient is defined). This is consistent with $N_C(x)$ as well, since if $x \notin C$ one might say there is no normal cone at an exterior point (or one can define $N_C(x) = \emptyset$). In any case, the equality holds trivially (both sides are empty sets for $x \notin C$).

13 Subdifferential of a Nonsmooth Function

Question: Consider the function

$$f(x) = \max\{x^2 - 1, 0\}, \quad x \in \mathbb{R}.$$

Find and justify the subdifferential $\partial f(x)$ at different points of x.

Relevant Concepts: The subdifferential of a convex function f at x is given by:

$$\partial f(x) = \{v \mid f(y) \geq f(x) + v(y - x), \quad \forall y \in \mathbb{R}\}.$$

If f is differentiable at x, then $\partial f(x)$ consists of the single element $\{f'(x)\}$. However, at nondifferentiable points, the subdifferential is a set of all possible supporting slopes.

Proof:
First, observe that $f(x)$ can be rewritten in piecewise form:

$$f(x) = \begin{cases} x^2 - 1, & \text{if } x < -1 \text{ or } x > 1, \\ 0, & \text{if } -1 \leq x \leq 1. \end{cases}$$

Now we analyze the subdifferential at different regions:
1. For $x < -1$ or $x > 1$: Here, $f(x) = x^2 - 1$ is differentiable, and its derivative is $f'(x) = 2x$. Since f is differentiable at these points, the subdifferential contains only one element:

$$\partial f(x) = \{2x\}.$$

2. For $-1 < x < 1$: In this interval, $f(x) = 0$, which is a constant function. The function is differentiable, and its derivative is $f'(x) = 0$. Therefore, the subdifferential is:

$$\partial f(x) = \{0\}.$$

3. For $x = -1$: The left-hand derivative is $\lim_{x \to -1^-} 2x = -2$. The right-hand derivative is $\lim_{x \to -1^+} 0 = 0$. Since the function is not differentiable at $x = -1$, the subdifferential is given by the convex hull of these values:

$$\partial f(-1) = [-2, 0].$$

4. For $x = 1$: The left-hand derivative is $\lim_{x \to 1^-} 0 = 0$. The right-hand derivative is $\lim_{x \to 1^+} 2x = 2$. Similarly, the subdifferential at $x = 1$ is:

$$\partial f(1) = [0, 2].$$

Thus, the complete subdifferential description is:

$$\partial f(x) = \begin{cases} \{2x\}, & \text{if } x < -1 \text{ or } x > 1, \\ \{0\}, & \text{if } -1 < x < 1, \\ [-2, 0], & \text{if } x = -1, \\ [0, 2], & \text{if } x = 1. \end{cases}$$

This matches the computed subgradients, and the result is proven.

14 Equivalence of Subgradients for Primal and Conjugate (Fenchel's Theorem)

Question: Let $f : X \to (-\infty, +\infty]$ be a closed, proper, convex function on a Euclidean space X, and let $f^* : X^* \to (-\infty, +\infty]$ denote its conjugate function (Legendre–Fenchel transform), defined by

$$f^*(y) = \sup_{x \in X} \{\langle y, x\rangle - f(x)\} \ .$$

Prove that the following conditions are equivalent for a pair $(x, y) \in X \times X^*$:

(1) $f(x) + f^*(y) = \langle y, x\rangle$.

(2) $y \in \partial f(x)$.

(3) $x \in \partial f^*(y)$.

In particular, show that $y \in \partial f(x)$ if and only if $x \in \partial f^*(y)$, and in this case the equality (1) (often called the Fenchel-Young equality) holds. This result shows that the subdifferential of f^* is the "inverse" relation of the subdifferential of f.

Relevant Concepts: Fenchel's inequality states that for any x and any y, we have $f(x)+f^*(y) \geq \langle y, x\rangle$ (by definition of f^* as a supremum). The three conditions above are different ways of characterizing when this inequality holds at equality and the suprema in the definitions are attained. Specifically: - $y \in \partial f(x)$ means $f(z) \geq f(x) + \langle y, z - x\rangle$ for all z. - $x \in \partial f^*(y)$ means $f^*(w) \geq f^*(y) + \langle x, w - y\rangle$ for all w. One approach is to start by showing (1) $\iff$ (2): use Fenchel's inequality and the definition of subgradient. Then deduce (1) $\implies$ (3) by symmetry (or by applying the previous result to f^* in place of f). The equivalence of (2) and (3) then follows.

Proof: First, recall Fenchel's inequality: by the definition of $f^*(y)$ as a supremum, for any $x \in X$ and $y \in X^*$,

$$f^*(y) \geq \langle y, x\rangle - f(x) \ ,$$

which we can rearrange as $f(x)+f^*(y) \geq \langle y, x\rangle$. This inequality holds for all x, y, and (1) is the condition that this holds with equality for the particular pair (x, y).

Now we prove the equivalences:

(1) $\iff$ (2): Assume (1) holds, i.e. $f(x)+f^*(y) = \langle y, x\rangle$. By the definition of $f^*(y)$ as $\sup_u\{\langle y, u\rangle - f(u)\}$, the equality $f(x) + f^*(y) = \langle y, x\rangle$ implies two things: (i) $f^*(y) = \langle y, x\rangle - f(x)$ (so the supremum is at least achieved by $u = x$), and (ii) for all $u \in X$, $\langle y, u\rangle - f(u) \leq \langle y, x\rangle - f(x)$ (since otherwise the supremum defining $f^*(y)$ would be larger than $\langle y, x\rangle - f(x)$). The latter condition can be rearranged for any u as:

$$f(u) - f(x) \geq \langle y, u - x\rangle, \qquad \forall u \in X.$$

This is exactly the subgradient inequality that says $y \in \partial f(x)$. Conversely, if $y \in \partial f(x)$, then the inequality above holds. In particular, taking u that (approximately) maximizes $\langle y, u\rangle - f(u)$ (the supremum), we get $\langle y, u\rangle - f(u) \leq \langle y, x\rangle - f(x)$. Hence $\langle y, x\rangle - f(x) = f^*(y)$. Rearranging gives $f(x)+f^*(y) = \langle y, x\rangle$. So (2) implies (1).

Thus (1) and (2) are equivalent.

(1) $\iff$ (3): This is analogous by symmetry, or by applying the above result to the conjugate function. Since f is closed convex, $f^{**} = f$. We can repeat the argument with f^* in place of f and x and y swapped. Specifically, Fenchel's inequality for f^* yields: $f^*(y) + f^{**}(x) \geq \langle y, x\rangle$. But $f^{**}(x) = f(x)$, so again we have $f(x) + f^*(y) \geq \langle y, x\rangle$ with equality if and only if $x \in \partial f^*(y)$ (by the same reasoning as before). Thus (1) holds if and only if $x \in \partial f^*(y)$.

Combining the two parts, we have shown:

$$f(x) + f^*(y) = \langle y, x\rangle \iff y \in \partial f(x) \iff x \in \partial f^*(y) \ .$$

In other words, (1), (2), and (3) are all equivalent. This means exactly that y is a subgradient of f at x if and only if x is a subgradient of f^* at y, and that happens precisely when the Fenchel-Young equality $f(x) + f^*(y) = \langle y, x\rangle$ holds.

15 Computation of the Conjugate Function f^*, Exam Question

Question: Let $f(x) = \|x\|_1$, where $x \in \mathbb{R}^n$. Compute the conjugate function $f^*(y)$.

Relevant Concepts:

- The **convex conjugate** (Fenchel conjugate) of a function $f(x)$ is given by:

$$f^*(y) = \sup_x \{\langle y, x\rangle - f(x)\}.$$

- The ℓ_1-norm is defined as $\|x\|_1 = \sum_{i=1}^n |x_i|$.
- The ℓ_∞-norm is given by $\|y\|_\infty = \max_i |y_i|$, which bounds the inner product $\langle y, x\rangle$.
- **Hölder's inequality**—Let (S, Σ, μ) be a **measure space** and let $p, q \in [1, \infty]$ with

$$\frac{1}{p} + \frac{1}{q} = 1.$$

  Then for all measurable real or complex valued functions f and g on S,

$$\|fg\|_1 \leq \|f\|_p \|g\|_q.$$

Solution: We compute $f^*(y)$ using the definition:

$$f^*(y) = \sup_x \{\langle y, x\rangle - \|x\|_1\}.$$

Case 1: $\|y\|_\infty \leq 1$

- By Hölder's inequality:

  $\langle y, x\rangle - \|x\|_1 \leq \|x\|_1(\text{flip sign for negatives})\|y\|_\infty(\text{replace all y with ymax}) - \|x\|_1 = \|x\|_1(\|y\|_\infty - 1).$

- Since $\|y\|_\infty - 1 \leq 0$, the supremum is maximized at $x = 0$, giving:

$$f^*(y) = 0.$$

Case 2: $\|y\|_\infty > 1$

- There exists some coordinate k such that $|y_k| > 1$.
- Consider the sequence $x^{(m)}$ defined as:

$$x^{(m)} = (0, \ldots, 0, m \cdot \text{sign}(y_k), 0, \ldots, 0).$$

- Then, we compute:

$$f^*(y) \geq \langle y, x^{(m)}\rangle - \|x^{(m)}\|_1 = m|y_k| - m = m(|y_k| - 1).$$

- As $m \to +\infty$, the term $m(|y_k| - 1) \to +\infty$, so:

$$f^*(y) = +\infty.$$

Conclusion: The conjugate function is the indicator function of the ℓ_∞-ball:

$$f^*(y) = \delta_C(y), \quad C = \{y \in \mathbb{R}^n \mid \|y\|_\infty \leq 1\}.$$

16 Conjugate of a Norm Function (Dual Norm Ball Indicator)

Question: Let $1 < p < \infty$ and let q be the conjugate exponent satisfying $\frac{1}{p} + \frac{1}{q} = 1$. Consider the function $f(x) = \lambda\|x\|_p$ on $\mathbb{R}^n$, where $\lambda > 0$ is a constant and $\|\cdot\|_p$ is the ℓ_p norm. Show that the Fenchel conjugate $f^*(y)$ is the indicator function of the closed ℓ_q ball of radius λ. In formula form, prove that

$$f^*(y) = \delta_C(y),$$

where $C = \{\, y \in \mathbb{R}^n \mid \|y\|_q \leq \lambda \,\}$.

Relevant Concepts: The definition of the conjugate is

$$f^*(y) = \sup_{x\in\mathbb{R}^n} \left\{\langle y, x\rangle - \lambda\|x\|_p\right\}\,.$$

This is a maximization problem. We expect that if $\|y\|_q > \lambda$, the supremum will be $+\infty$ (meaning no finite conjugate, hence $f^*(y) = +\infty$ which matches $\delta_C(y)$ since y is outside the set C). If $\|y\|_q \leq \lambda$, we expect the supremum to be attained at some finite x giving a finite value (indeed, in such case one could expect the supremum to be 0 at $x = 0$). We can guess that the optimal x is in the direction of y (because $\langle y, x\rangle$ for fixed y is maximized for a given $\|x\|_p$ when x is aligned with y). Using Hölder's inequality: $\langle y, x\rangle \leq \|y\|_q\|x\|_p$. This suggests: - If $\|y\|_q < \lambda$, then $\langle y, x\rangle - \lambda\|x\|_p \leq (\|y\|_q - \lambda)\|x\|_p \leq 0$ for all x, with supremum 0 attained in the limit $x \to 0$. - If $\|y\|_q = \lambda$, then $\langle y, x\rangle - \lambda\|x\|_p \leq 0$ always, and for x parallel to y it equals 0 for any magnitude, so the supremum is 0 (achieved arbitrarily close with large x in that direction, but not giving $+\infty$ because the coefficient is zero). - If $\|y\|_q > \lambda$, then for x parallel to y, $\langle y, x\rangle - \lambda\|x\|_p = (\|y\|_q - \lambda)\|x\|_p$, which can be made arbitrarily large as $\|x\|_p \to \infty$, hence supremum $+\infty$.

Proof: We compute

$$f^*(y) = \sup_{x\in\mathbb{R}^n} \{\langle y, x\rangle - \lambda\|x\|_p\}.$$

If $\|y\|_q > \lambda$, consider $x = t\,u$ where $u = \mathrm{sgn}(y)$ is a unit vector chosen such that equality holds in Hölder's inequality for y and u (specifically, u is a vector with $|u_i|^p = |y_i|^q/\|y\|_q^q$). Then $\|u\|_p = 1$ and $\langle y, u\rangle = \|y\|_q$. So

$$\langle y, x\rangle - \lambda\|x\|_p = t\langle y, u\rangle - \lambda t = t(\|y\|_q - \lambda).$$

If $\|y\|_q > \lambda$, this linear function of t can be made arbitrarily large by taking $t \to +\infty$. Hence $f^*(y) = +\infty$ for $\|y\|_q > \lambda$. This matches $\delta_C(y) = +\infty$ since $y \notin C$.
If $\|y\|_q \leq \lambda$, then for any x, Hölder's inequality gives $\langle y, x\rangle \leq \|y\|_q\|x\|_p$. Thus

$$\langle y, x\rangle - \lambda\|x\|_p \leq (\|y\|_q - \lambda)\|x\|_p.$$

If $\|y\|_q < \lambda$, the right-hand side is non-positive for all x, and approaches 0 from below as $\|x\|_p \to 0$. If $\|y\|_q = \lambda$, the right-hand side is $0 \cdot \|x\|_p = 0$ for all x, so $\langle y, x\rangle - \lambda\|x\|_p \leq 0$ with equality achieved when x is aligned with y (in fact, for x parallel to y the expression is exactly zero regardless of norm, hence the supremum is 0). In either case, the supremum is 0. Therefore

$$f^*(y) = 0 \quad \text{for all } y \text{ with } \|y\|_q \leq \lambda.$$

Combining these results, we have

$$f^*(y) = \begin{cases} 0, & \|y\|_q \leq \lambda, \\ +\infty, & \|y\|_q > \lambda, \end{cases}$$

which is exactly $\delta_{\{y:\|y\|_q\leq\lambda\}}(y)$. This shows $f^*(y) = \delta_C(y)$ for $C = \{y : \|y\|_q \leq \lambda\}$.

17 Moreau Envelope and Proximal Mapping of $f(x) = \lambda|x|$

Let $f(x) = \lambda|x|$, where $x \in \mathbb{R}$ and $\lambda > 0$. This function is convex, proper, and lower semicontinuous. We aim to compute the **Moreau envelope** and the **proximal mapping** of f.

Question

1. Show that the **Moreau envelope** of $f(x)$ is given by:

$$M_f(x) = \begin{cases} \frac{1}{2}x^2, & |x| \leq \lambda \\ \lambda|x| - \frac{\lambda^2}{2}, & |x| > \lambda \end{cases}$$

2. Show that the **proximal mapping** of $f(x)$ is:

$$P_f(x) = \operatorname{sign}(x) \cdot \max\{|x| - \lambda, 0\}$$

Relevant Concepts

- The **Moreau envelope** of a convex function f is:

$$M_f(x) = \min_{y \in \mathbb{R}} \left\{ f(y) + \frac{1}{2}(y - x)^2 \right\}$$

- The **proximal mapping** $P_f(x)$ is the minimizer of the above:

$$P_f(x) = \arg\min_{y \in \mathbb{R}} \left\{ f(y) + \frac{1}{2}(y - x)^2 \right\}$$

Solving the Minimization Problem

To find $P_f(x)$, define the objective:

$$g(y) := \lambda|y| + \frac{1}{2}(y - x)^2$$

We solve $\min_y g(y)$. The function is convex and piecewise differentiable. Consider the cases based on the sign of y:

Case 1: $y > 0$ Then $|y| = y$, and

$$g(y) = \lambda y + \frac{1}{2}(y - x)^2$$

Take derivative:

$$g'(y) = \lambda + (y - x) = \lambda + y - x$$

Set $g'(y) = 0$ to minimize:

$$y = x - \lambda$$

This is valid only if $x - \lambda > 0$, i.e., $x > \lambda$

Case 2: $y < 0$ Then $|y| = -y$, and

$$g(y) = -\lambda y + \frac{1}{2}(y - x)^2$$

Take derivative:

$$g'(u) = -\lambda + (y - x) = -\lambda + y - x$$

Set $g'(y) = 0$:

$$y = x + \lambda$$

Valid if $x + \lambda < 0$, i.e., $x < -\lambda$

Case 3: $y = 0$ If $|x| \leq \lambda$, then the minimum is achieved at $y = 0$. The derivative is not defined at $y = 0$, so we check subdifferential:

$$g(y) = \lambda|| \cdot || + \frac{1}{2}(y - x)^2$$

$$\partial g(y) = \lambda[-1, 1] - x$$

$$0 \in \partial g(y) \Rightarrow \lambda[-1, 1] = x \Rightarrow x \leq |\lambda|$$

Final Expression: Proximal Mapping

Combining all cases:

$$P_f(x) = \begin{cases} x - \lambda, & x > \lambda \\ 0, & |x| \leq \lambda \\ x + \lambda, & x < -\lambda \end{cases} = \text{sign}(x) \cdot \max\{|x| - \lambda, 0\}$$

Moreau Envelope

We now substitute $u = P_f(x)$ into the expression:

$$M_f(x) = f(P_f(x)) + \frac{1}{2}(x - P_f(x))^2$$

This gives:

$$M_f(x) = \begin{cases} \frac{1}{2}x^2, & |x| \leq \lambda \\ \lambda|x| - \frac{\lambda^2}{2}, & |x| > \lambda \end{cases}$$

Since when $x > \lambda$,

$$M_f(x) = f(x - \lambda) + \frac{\lambda^2}{2} = \lambda(x - \lambda) + \frac{\lambda^2}{2} = \lambda x - \frac{\lambda^2}{2}$$

when $x < \lambda$,

$$M_f(x) = f(x + \lambda) + \frac{\lambda^2}{2} = -\lambda(x + \lambda) + \frac{\lambda^2}{2} = -\lambda x - \frac{\lambda^2}{2}$$

Lecture 5: Support Vector Machine(SVM), Duality, KKT

The Normal Cone of a Hyperplane is 1-Dimensional

Question: Prove that the normal cone of a hyperplane $H = H_{\beta,\beta_0} = \{x \in \mathbb{R}^p \mid \beta^T x + \beta_0 = 0\}$ is 1-dimensional.

Relevant Concepts:

- **Hyperplane Definition:** A hyperplane H in $\mathbb{R}^p$ is a $(p-1)$-dimensional affine subspace defined by a linear equation $\beta^T x + \beta_0 = 0$.
- **Normal Cone:** The normal cone of H at a point $\bar{x} \in H$ is defined as

$$N_H(\bar{x}) = \{v \in \mathbb{R}^p \mid \langle v, z - \bar{x} \rangle \leq 0, \quad \forall z \in H\}.$$

- **Linear Subspace Properties:** For a hyperplane, the normal direction is given by β, which is perpendicular to every vector lying in H.

Proof:

- The hyperplane H is defined by the equation $\beta^T x + \beta_0 = 0$, meaning every vector in H satisfies this condition.
- The normal cone at any $\bar{x} \in H$ is given by:

$$N_H(\bar{x}) = \{\lambda \beta \mid \lambda \in \mathbb{R}\}.$$

  This means any normal vector to H must be a scalar multiple of β.
- Since $N_H(\bar{x})$ consists only of multiples of β, it forms a 1-dimensional subspace (a line through the origin in the direction of β).
- The normal cone cannot be higher-dimensional because any additional independent direction would contradict the definition of a hyperplane, which has exactly one normal direction.

Conclusion: Since the normal cone consists of only scalar multiples of β, it forms a one-dimensional subspace, proving that $N_H(\bar{x})$ is always 1-dimensional.

Lecture 6: Block Coordinate Descent (BCD) Method

18 Verification of the Moreau Envelope and Proximal Mapping

Question: Verify the expressions for the Moreau envelope $M_f(x)$ and proximal mapping $P_f(x)$ of the function $f(x) = \lambda|x|$.
Relevant Concepts: The Moreau envelope is defined as:

$$M_f(x) = \min_{y\in\mathbb{R}} \left(\lambda|y| + \frac{1}{2}(y-x)^2 \right).$$

The proximal mapping is given by:

$$P_f(x) = \arg\min_{y\in\mathbb{R}} \left(\lambda|y| + \frac{1}{2}(y-x)^2 \right).$$

Case 1: $|x| \leq \lambda$
We minimize:

$$F(y) = \lambda|y| + \frac{1}{2}(y-x)^2.$$

To find the optimal y, we differentiate separately for $y > 0$ and $y < 0$:
For $y > 0$, $|y| = y$, and the derivative is:

$$\lambda + (y - x).$$

Setting this to zero gives:

$$y - x = -\lambda \quad \Rightarrow \quad y = x - \lambda.$$

For $y < 0$, $|y| = -y$, and the derivative is:

$$-\lambda + (y - x).$$

Setting this to zero gives:

$$y - x = \lambda \quad \Rightarrow \quad y = x + \lambda.$$

To determine the correct minimizer, note that if $|x| \leq \lambda$, then:

$$x - \lambda \leq 0 \quad \text{and} \quad x + \lambda \geq 0.$$

Thus, $y = 0$ is in the feasible region, and evaluating at $y = 0$:

$$F(0) = \lambda|0| + \frac{1}{2}x^2 = \frac{1}{2}x^2.$$

This confirms that the minimizer is $y = 0$, and hence:

$$P_f(x) = 0, \quad M_f(x) = \frac{1}{2}x^2.$$

Case 2: $|x| > \lambda$
For this case, we expect the minimizer to be nonzero. Consider the optimality conditions derived before:

- If $x > \lambda$, then $y = x - \lambda$ is positive.
- If $x < -\lambda$, then $y = x + \lambda$ is negative.

Checking feasibility:

- If $y = x - \lambda$ and $x > \lambda$, then $y > 0$.
- If $y = x + \lambda$ and $x < -\lambda$, then $y < 0$.

Evaluating $F(y)$ at these points:

$$F(y) = \lambda|y| + \frac{1}{2}(y-x)^2.$$

Since $y = x - \lambda$ (or $y = x + \lambda$ for $x < -\lambda$) is the global minimizer, we compute:

$$M_f(x) = \lambda|x| - \frac{\lambda^2}{2}.$$

Thus, the proximal mapping is:

$$P_f(x) = \operatorname{sign}(x)\max\{|x| - \lambda, 0\}.$$

Conclusion: We have verified that:

$$M_f(x) = \begin{cases} \frac{1}{2}x^2, & |x| \leq \lambda, \\ \lambda|x| - \frac{\lambda^2}{2}, & |x| > \lambda. \end{cases}$$

and

$$P_f(x) = \operatorname{sign}(x)\max\{|x| - \lambda, 0\}.$$

This confirms the correctness of the given formulas.

Lecture 7: Nonnegative Matrix Factorization(NMF)

Proof of Lemma 1

Given a matrix $B \in \mathbb{R}^{M \times N}$ and a nonzero vector $v \in \mathbb{R}^N$, then:

$$\frac{[Bv]_+}{v^T v} = \arg\min_{u \geq 0} ||B - uv^T||^2 \tag{1}$$

Similarly, given a matrix $B \in \mathbb{R}^{M \times N}$ and a nonzero vector $u \in \mathbb{R}^M$, then:

$$\frac{[B^T u]_+}{u^T u} = \arg\min_{v \geq 0} ||B - uv^T||^2 \tag{2}$$

The unique optimal solution is guaranteed due to $v \neq 0$ or $u \neq 0$ respectively

Proof:

$$B = [b_1, b_2, \ldots, b_N], \quad \mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_M \end{bmatrix}, \quad \mathbf{v} = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_N \end{bmatrix}$$

$$\mathbf{u}\mathbf{v}^\top = \begin{bmatrix} u_1 v_1 & u_1 v_2 & \cdots & u_1 v_N \\ u_2 v_1 & u_2 v_2 & \cdots & u_2 v_N \\ \vdots & \vdots & \ddots & \vdots \\ u_M v_1 & u_M v_2 & \cdots & u_M v_N \end{bmatrix}$$

$$B - \mathbf{u}\mathbf{v}^\top = \begin{bmatrix} b_{11} - u_1 v_1 & b_{12} - u_1 v_2 & \cdots & b_{1N} - u_1 v_N & (u_1 \text{ unknown}) \\ b_{21} - u_2 v_1 & b_{22} - u_2 v_2 & \cdots & b_{2N} - u_2 v_N & \\ \vdots & \vdots & \ddots & \vdots & \\ b_{M1} - u_M v_1 & b_{M2} - u_M v_2 & \cdots & b_{MN} - u_M v_N & (u_M \text{ unknown}) \end{bmatrix}$$

$$\begin{aligned}
\arg\min_{u \geq 0} \|B - uv^T\|_F^2 &= \min_{u_i \geq 0} \left\|(b_{i1} - u_i v_1, \ldots, b_{iN} - u_i v_N)\right\|^2, \quad i = 1, \ldots, M \\
&= \min_{u_i \geq 0} \sum_{j=1}^{N} (b_{ij} - u_i v_j)^2 \\
&= \min_{u_i \geq 0} (b_{i1}^2 + \ldots b_{iN}^2) + u_i(v_1^2 + \ldots + v_N^2) - 2u_i(b_{i1}v_i + \ldots + b_{iN}v_N) \\
&= \min_{u_i \geq 0} (\text{constant w.r.t. } u_i) + u_i(v_1^2 + \ldots + v_N^2) - 2(B(i,:)v)u_i \quad (B(i,:)v \text{ is } i^{th} \text{ row of B times v}) \\
&= \min_{u_i \geq 0} ||v||^2 u_i^2 - 2(B(i,:)v)u_i \\
&= \min_{u_i \geq 0} u_i^2 - \frac{2(B(i,:)v)u_i}{||v||^2} \\
&= \min_{u_i \geq 0} (u_i - \frac{B(i,:)v}{||v||^2})^2 \quad (\text{Complete the square by adding a constant w.r.t. } u_i \text{ over min})
\end{aligned}$$

Case 1: If $B(i,:)v \geq 0$, take $u_i = \frac{B(i,:)v}{||v||^2}$ for the minimum of 0:

$$\arg\min_{u_i \geq 0} (u_i - \frac{B(i,:)v}{||v||^2})^2 = \frac{B(i,:)v}{||v||^2}$$

Case 2: If $B(i,:)v < 0$, take $u_i = 0$ for minimum:

$$\arg\min_{u_i \geq 0} (u_i - \frac{B(i,:)v}{||v||^2})^2 = 0$$

Combining, we have

$$\arg\min_{u_i \geq 0} (u_i - \frac{B(i,:)v}{||v||^2})^2 = \frac{[B(i,:)v]_+}{||v||^2}$$

where:

$$[B(i,:)\,\mathbf{v}]_+ = \begin{cases} B(i,:)\,\mathbf{v}, & \text{if } B(i,:)\,\mathbf{v} > 0 \\ 0, & \text{otherwise} \end{cases}$$

Stacking over all is we have

$$\arg\min_{u \geq 0} ||B - uv^T||^2 = \frac{[Bv]_+}{v^T v}$$

For the second part of the proof, we take the transpose of $B - uv^T$ and get $B^T - vu^T$ while retaining the minimization relationship. Then it reduces to the proof for the first part:

$$\begin{aligned} \arg\min_{v \geq 0} \|B - uv^T\|_F^2 &= \arg\min_{v \geq 0} \|B^T - vu^T\|_F^2 \\ &= \frac{[B^T u]_+}{||u||^2} \end{aligned}$$

Proof of Lemma 5

Given a continuous and convex function $f(z)$, and two nonempty closed convex sets T and C satisfying $T \cap C \neq \emptyset$, assume:

$$\tilde{z} = \arg\min_{z \in T} f(z),$$

and $\tilde{z}$ is finite. Assume further that the constrained optimization problem:

$$\min_{z \in T \cap C} f(z)$$

has a finite solution.
Then:

- If $\tilde{z} \in C$, then $z^* = \tilde{z} = \arg\min_{z \in T \cap C} f(z)$.
- If $\tilde{z} \notin C$, then there exists a z^* in $T \cap C_{\text{edge}}$ such that $z^* = \arg\min_{z \in T \cap C} f(z)$, where C_{edge} denotes the boundary of C.

Proof:

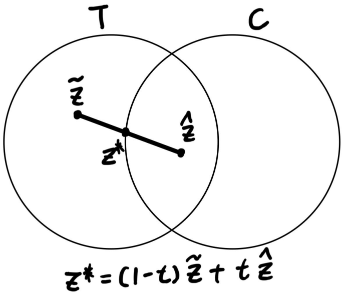

Figure 1: Lemma 5 illustration

Part 1: If $\tilde{z} \in C$, then since $\tilde{z}$ is the minimum of f over T, it is also the minimum over $T \cap C$ (which is a subset of T), so $z^* = \tilde{z}$.

Part 2: If $\tilde{z} \notin C$: Let $\hat{z} \in T \cap C$ be finite and $\hat{z} = \arg\min_{T \cap C} f(z)$. Then if $\hat{z} \in T \cap C_{edge}$ then $z^* = \hat{z}$. Otherwise, with reference to the diagram above, suppose $\hat{z} \in T \cap C_{int}$ where C_{int} is the interior point of C. Note that $\tilde{z} \notin C, \tilde{z} \in T$ ($\tilde{z}$ does not lie in the intersection as shown), then, the value can be further minimized by considering the following convex combination:

$$z^* = (1-t)\tilde{z} + t\hat{z} \in T \cap C_{edge} \quad \text{for some } t \in (0,1) \text{ and } z^* \in T \cap C_{edge}$$

Graphically, think of this problem as try to approach and minimize the distance to $\tilde{z}$,then for any $\hat{z} \in C_{int}$, we can connect the two points and find z^* as the intersection to C_{edge}, which reduces the distance.
Furthermore, $f(\tilde{z}) < f(\hat{z})$ (minimizing over a set minimizes all its subsets as well):

$$\begin{aligned} f(z^*) &\leq (1-t)f(\tilde{z}) + tf(\hat{z})) \\ &\leq (1-t)f(\hat{z}) + tf(\hat{z})) \\ &\leq f(\hat{z}) \end{aligned}$$

With $\hat{z} = \arg\min_{T \cap C} f(z)$ ($\hat{z}$ is already the minimum so that there is no smaller value), we have $f(z^*) = f(\hat{z})$, and thus $z^* = \arg\min_{z \in T \cap C} f(z)$.

Proof of Theorem 4

Let $G \in \mathbb{R}^{m \times k}$, $g_{k+1} \in \mathbb{R}^m$, and $b \in \mathbb{R}^m$ be given, where G and $\{G, g_{k+1}\}$ are of full column rank. The unique solution of the rank-k NLS problem is given as:

$$s(G, b) \in \mathbb{R}^k$$

Then, the unique solution of the rank-$(k+1)$ NLS problem:

$$\begin{bmatrix} y^* \\ y^*_{k+1} \end{bmatrix} = \arg \min_{y \geq 0, y_{k+1} \geq 0} \left\| \begin{bmatrix} G & g_{k+1} \end{bmatrix} \begin{bmatrix} y \\ y_{k+1} \end{bmatrix} - b \right\|$$

is given by:

$$y^*_{k+1} = \frac{1}{\|g_{k+1}\|^2} [g^T_{k+1}(b - G \times s(G - \frac{g_{k+1}g^T_{k+1}}{||g_{k+1}||^2})G, b - \frac{g_{k+1}g^T_{k+1}}{||g_{k+1}||^2}b]_+$$

$$y^* = s(G, b - g_{k+1}y^*_{k+1})$$

Proof:
Let's consider

$$\begin{bmatrix} \tilde{y} \\ \tilde{y_{k+1}} \end{bmatrix} = \arg \min_{y \geq 0, y_{k+1} \in \mathbb{R}} \left\| \begin{bmatrix} G & g_{k+1} \end{bmatrix} \begin{bmatrix} y \\ y_{k+1} \end{bmatrix} - b \right\|^2_F$$

$$\tilde{y} = \arg \min_{y \geq 0} ||Gy - (b - g_{k+1}\tilde{y}_{k+1})||^2_F$$

Assuming $\tilde{y}_{k+1}$ ahs been found,

$$\tilde{y} = S(G, b - g_{k+1}\tilde{y}_{k+1})$$

For any given y,

$$\begin{aligned} \tilde{y}_{k+1} &= \arg \min_{y_{k+1} \in \mathbb{R}} ||g_{k+1}y_{k+1} - (b - G\tilde{y})||^2_F \quad \text{(standard least square problem)} \\ &= \frac{g^T_{k+1}(b - G\tilde{y})}{||g_{k+1}||^2} \end{aligned}$$

This means:

$$\begin{aligned} \tilde{y} &= \arg \min_{y \geq 0} ||Gy + g_{k+1}\frac{g_{k+1}(b - G\tilde{y})}{||g_{k+1}||^2} - b||^2_F \\ &= \arg \min_{y \geq 0} ||\tilde{G}y - \tilde{b}||^2_F, \text{where } \tilde{G} = G - \frac{g_{k+1}g^T_{k+1}}{||g_{k+1}||^2}G, \tilde{b} = b - \frac{g_{k+1}g^T_{k+1}}{||g_{k+1}||^2}b \\ &= S(\tilde{G}, \tilde{b}) \quad \text{(both } \tilde{G} \text{ and } \tilde{b} \text{ are constant)} \end{aligned}$$

Note that if $rank(\tilde{G}) = k$, then $rank(\tilde{G}, g_{k+1}) = k + 1$. With reference to lemma 5, consider two sets T and C defined as:

$$T = \{ \begin{bmatrix} y \\ y_{k+1} \end{bmatrix}, y \in \mathbb{R}^k, y \geq 0, y_{k+1} \in \mathbb{R}\}$$

$$C = \{ \begin{bmatrix} y \\ y_{k+1} \end{bmatrix}, y \in \mathbb{R}^k, y \geq 0, \boldsymbol{y_{k+1} \geq 0}\}$$

Then $C = T \cap C$ since C is a subset of T and $C \in T$.

$$\begin{aligned} \tilde{z} &= \begin{bmatrix} \tilde{y} \\ \tilde{y_{k+1}} \end{bmatrix} \\ &= \arg \min_{z \in T} ||G \begin{bmatrix} y \\ y_{k+1} \end{bmatrix} - b||^2_F \\ \hat{z} &= \begin{bmatrix} \hat{y} \\ \hat{y_{k+1}} \end{bmatrix} \\ &= \arg \min_{z \in C} ||G \begin{bmatrix} y \\ y_{k+1} \end{bmatrix} - b||^2_F \\ &= \arg \min_{\boldsymbol{z \in T \cap C}} ||G \begin{bmatrix} y \\ y_{k+1} \end{bmatrix} - b||^2_F \end{aligned}$$

By lemma 5, if:

$$\tilde{y}_{k+1} = \frac{g_{k+1}^T(b - G \times S(\tilde{G}, \tilde{b}))}{||g_{k+1}||^2} \geq 0$$

then $\tilde{z} \in C \Rightarrow \tilde{z} = \hat{z}$. Otherwise, if $\tilde{y}_{k+1} \notin C$ then $\hat{z} \in T \cap C_{edge} \Rightarrow g_{k+1} = 0$
Thus:

$$\hat{y} = \arg\min_{y \geq 0} ||Gy - b||_F^2 = S(G, b)$$

To summarize:

- If $\tilde{y}_{k+1} = \frac{g_{k+1}^T(b - G \times S(\tilde{G}, \tilde{b}))}{||g_{k+1}||^2} \geq 0$, then $\hat{y} = \tilde{y} = S(\tilde{G}, \tilde{b})$, $\hat{y}_{k+1} = \tilde{y}_{k+1} = \frac{g_{k+1}^T(b - G \times S(\tilde{G}, \tilde{b}))}{||g_{k+1}||^2}$
- Otherwise: $\hat{y} = S(G, b)$, $\hat{y}_{k+1} = 0$