FALL 2026 BAIS 7900

# Material: Mathematical Foundations for Online Learning

Lecturer: *Mengxiao Zhang* Date: Aug 26th

## 1 Notation and rates

- The $i$-th coordinate of a vector $w$ is $w_i$. If the vector changes over time, its $i$-th coordinate at round $t$ is $w_{t,i}$.
- The inner product and Euclidean norm are

$$\langle x, y\rangle = x^\top y = \sum_{i=1}^{d} x_i y_i, \qquad \|x\|_2 = \sqrt{\langle x, x\rangle}.$$

- For nonnegative sequences $f(T)$ and $g(T)$:

$$\begin{aligned} f(T) = O(g(T)) &\iff f(T) \le Cg(T) \text{ eventually for some } C > 0, \\ f(T) = o(g(T)) &\iff f(T)/g(T) \longrightarrow 0, \\ f(T) = \Theta(g(T)) &\iff f(T) = O(g(T)) \text{ and } g(T) = O(f(T)). \end{aligned}$$

## 2 Linear algebra and decision geometry

### 2.1 Vectors, matrices, and positive (semi-)definiteness

- If $A \in \mathbb{R}^{m\times d}$ and $x \in \mathbb{R}^d$, then $Ax \in \mathbb{R}^m$, $A^\top A \in \mathbb{R}^{d\times d}$, and $AA^\top \in \mathbb{R}^{m\times m}$.
- The outer product $xy^\top$ is a matrix; the inner product $x^\top y$ is a scalar.
- Useful identities include

$$(AB)^\top = B^\top A^\top, \qquad (AB)^{-1} = B^{-1}A^{-1}, \qquad \operatorname{tr}(AB) = \operatorname{tr}(BA),$$

  whenever dimensions and inverses are valid.
- A symmetric matrix $Q$ is positive semidefinite, written $Q \succeq 0$, if $x^\top Qx \ge 0$ for every $x$. We call a symmetric matrix $Q$ positive definite, written $Q \succ 0$, if $x^\top Qx > 0$ for every nonzero vector $x$.
- A scalar $\lambda$ is an eigenvalue of a square matrix $A$ if there exists a nonzero vector $v$ such that $Av = \lambda v$. We call $v$ a corresponding eigenvector.
- Gram matrices are positive semidefinite:

$$A^\top A \succeq 0, \qquad x^\top A^\top Ax = \|Ax\|_2^2 .$$

- If $Q = Q^\top$, then
$$Q \succeq 0 \quad \Longleftrightarrow \quad \lambda_i(Q) \geq 0 \text{ for every eigenvalue.}$$
The spectral theorem gives $Q = U\Lambda U^\top$ with orthogonal $U$.

- Eigenvalue properties
$$\lambda_{\min}(Q) \leq \frac{x^\top Q x}{\|x\|_2^2} \leq \lambda_{\max}(Q) \qquad (x \neq 0),$$
where $\lambda_{\min}$ and $\lambda_{\max}$ are the minimum and maximum eigenvalues of $Q$.

- If $Q \succ 0$, define the matrix-induced norm
$$\|x\|_Q = \sqrt{x^\top Q x}.$$
The corresponding generalized Cauchy–Schwarz inequality is
$$|\langle x, y\rangle| \leq \|x\|_Q \, \|y\|_{Q^{-1}} .$$

- If $Q \succ 0$, the rank-one identities are
$$\det(Q + xx^\top) = \det(Q)\big(1 + x^\top Q^{-1} x\big)$$
and
$$(Q + xx^\top)^{-1} = Q^{-1} - \frac{Q^{-1} x x^\top Q^{-1}}{1 + x^\top Q^{-1} x}.$$

- For $f(w) = \frac{1}{2}\|Aw - b\|_2^2$,
$$\nabla f(w) = A^\top (Aw - b), \qquad \nabla^2 f(w) = A^\top A \succeq 0.$$

## 2.2 Norms and dual norms

- The norms used most often are
$$\|x\|_1 = \sum_i |x_i| , \qquad \|x\|_2 = \sqrt{\sum_i x_i^2}, \qquad \|x\|_\infty = \max_i |x_i| .$$

- The dual norm is
$$\|y\|_* = \max_{\|x\| \leq 1} \langle x, y\rangle .$$
The dual of $\ell_2$-norm is $\ell_2$-norm; the dual of $\ell_1$-norm is $\ell_\infty$-norm.

- Hölder's inequality gives
$$|\langle x, y\rangle| \leq \|x\| \cdot \|y\|_* .$$
Cauchy–Schwarz is the Euclidean special case
$$|\langle x, y\rangle| \leq \|x\|_2 \, \|y\|_2 .$$
Important special cases are
$$\langle x, y\rangle \leq \|x\|_2 \, \|y\|_2 , \qquad \langle x, y\rangle \leq \|x\|_1 \, \|y\|_\infty .$$

### 2.3 Convex sets, the simplex, and projection

- A set $\Omega$ is convex if
$$\lambda u + (1-\lambda)v \in \Omega \quad \text{for all } u, v \in \Omega,\ \lambda \in [0,1].$$
- The probability simplex is
$$\Delta(d) = \left\{ w \in \mathbb{R}^d : w_i \geq 0 \text{ for all } i,\ \sum_{i=1}^{d} w_i = 1 \right\}.$$
- A convex combination $\sum_i \lambda_i x_i$ has $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$.
- Euclidean projection onto a nonempty closed convex set is
$$\Pi_\Omega(y) := \arg\min_{w \in \Omega} \frac{1}{2} \|w - y\|_2^2 .$$
- Projection onto the simplex has the form **[derive this offline from the definition if you are not familiar with this operation]**
$$\left[\Pi_{\Delta(d)}(y)\right]_i = \max\{y_i - \tau, 0\},$$
where $\tau$ makes the projected coordinates sum to one.

## 3 Probability for sequential decisions

### 3.1 Conditioning and expectation

- Let $A$ and $B$ be two events with $\mathbb{P}(B) > 0$. Conditional probability and Bayes' rule are
$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \qquad \mathbb{P}(A \mid B) = \frac{\mathbb{P}(B \mid A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$
- Linearity of expectation does not require independence: for $n$ possibly dependent random variables $X_1, \ldots, X_n$, we have
$$\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mathbb{E}[X_i].$$
- For an event $A$, $\mathbb{E}[\mathbf{1}\{A\}] = \mathbb{P}(A)$.
- If $A$ and $B$ are independent and $\mathbb{P}(B) > 0$, then $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. If $X$ and $Y$ are independent, then $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$ when the expectations exist. The reverse implications do not generally hold.
- Variance and covariance are defined as
$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2], \qquad \mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)].$$
- For a random vector $X$ with mean $\mu$, its covariance matrix is positive semidefinite:
$$\mathrm{Cov}(X) = \mathbb{E}[(X - \mu)(X - \mu)^\top] = \mathbb{E}[XX^\top] - \mu\mu^\top \succeq 0.$$
- If $X \sim \mathcal{N}(\mu, \Sigma)$, then every linear transformation $a^\top X$ is Gaussian with mean $a^\top \mu$ and variance $a^\top \Sigma a$.

### 3.2 Martingales

In sequential problems, observations may depend on the past, so independence is often too strong an assumption. Martingales provide a natural framework for modeling such dependence while retaining a conditional notion of zero drift.

**Definition 3.1** (Martingale)**.** *A stochastic process $\{M_t\}_{t\geq 0}$ is a martingale if $\mathbb{E}[|M_t|] < \infty$ for every $t$ and $\mathbb{E}[M_t \mid M_0, \ldots, M_{t-1}] = M_{t-1}$. Equivalently, its increments $X_t := M_t - M_{t-1}$ satisfy $\mathbb{E}[X_t \mid M_0, \ldots, M_{t-1}] = 0$. The increments need not be independent.*

**Example 3.2.** *Let $Y_1, Y_2, \ldots$ be independent random variables taking values $+1$ and $-1$ with equal probability, and define $M_t \triangleq \sum_{s=1}^{t} Y_s$ and $M_0 = 0$. Then, given the previous outcomes, the next step still has conditional mean zero, so $\mathbb{E}[M_t \mid M_0, \ldots, M_{t-1}] = M_{t-1}$. Hence $\{M_t\}_{t\geq 0}$ is a martingale.*

Two important concentration inequalities are Hoeffding's inequality for independent bounded random variables and Azuma–Hoeffding's inequality for martingales.

**Theorem 3.3** (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_T$ be independent random variables such that $X_t \in [a_t, b_t]$ almost surely. Then, for every $\epsilon > 0$,*

$$\Pr\left(\left|\sum_{t=1}^{T} X_t - \sum_{t=1}^{T} \mathbb{E}[X_t]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{2\epsilon^2}{\sum_{t=1}^{T}(b_t - a_t)^2}\right).$$

**Theorem 3.4** (Azuma–Hoeffding's inequality)**.** *Let $\{M_t\}_{t=0}^{T}$ be a martingale such that $|M_t - M_{t-1}| \leq c_t$ almost surely for every $t = 1, \ldots, T$. Then, for every $\epsilon > 0$,*

$$\Pr\left(|M_T - M_0| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2}{2\sum_{t=1}^{T} c_t^2}\right).$$

The two inequalities have a similar interpretation: when the individual bounds are constants, the sum of bounded random fluctuations is unlikely to deviate from its mean by much more than order $\sqrt{T}$. The main difference is that Hoeffding's inequality assumes independence, whereas Azuma–Hoeffding allows the increments to depend on the past, as long as their conditional means are zero.

## 4 Calculus and optimization

### 4.1 Derivatives in several dimensions

- For $f : \mathbb{R}^d \to \mathbb{R}$, the gradient of $f$ is defined as

$$\nabla f(w) \triangleq \left(\frac{\partial f}{\partial w_1}, \ldots, \frac{\partial f}{\partial w_d}\right)^{\top}.$$

- The directional derivative along $v$ is $\langle \nabla f(w), v\rangle$.
- The Hessian of $f$ is a $d \times d$ matrix with the $(i, j)$-th entry defined as

$$[\nabla^2 f(w)]_{ij} = \frac{\partial^2 f(w)}{\partial w_i \partial w_j}.$$

- For $g : \mathbb{R}^d \to \mathbb{R}^m$, the Jacobian matrix $Dg(w)$ is the $m \times d$ matrix whose $i$-th row is $\nabla g_i(w)^\top$.
- Useful derivatives are

| $f(w)$ | $\nabla f(w)$ |
|---|---|
| $a^\top w$ | $a$ |
| $\frac{1}{2}\|w\|_2^2$ | $w$ |
| $\frac{1}{2}\|Aw - b\|_2^2$ | $A^\top(Aw - b)$ |
| $-\ln(a^\top w)$ | $-a/(a^\top w)$ |
| $\ln \sum_i e^{w_i}$ | $\left(e^{w_i}/\sum_j e^{w_j}\right)_{i=1}^d$ |

  whenever the logarithms are well defined.

## 4.2 Convexity and first-order methods

- A function $f : \Omega \to \mathbb{R}$ is convex if $\Omega$ is convex and, for any $x, y \in \Omega$ and $\lambda \in [0, 1]$, we have $\lambda f(x) + (1 - \lambda) f(y) \geq f(\lambda x + (1 - \lambda) y)$. A function is concave if $-f$ is convex.
- If $f$ is convex, Jensen's inequality gives $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$. Equivalently, if $f$ is concave, then $f(\mathbb{E}[X]) \geq \mathbb{E}[f(X)]$.
- A differentiable function is convex if and only if

$$f(v) \geq f(u) + \langle \nabla f(u), v - u \rangle \qquad \text{for all } u, v \in \Omega.$$

- A vector $g \in \partial f(u)$ is a subgradient if

$$f(v) \geq f(u) + \langle g, v - u \rangle \qquad \text{for all } v \in \Omega.$$

- If $f$ is twice differentiable on an open convex set $\Omega$, then $f$ is convex if and only if $\nabla^2 f(w) \succeq 0$ for every $w \in \Omega$.
- If $f$ is $\alpha$-strongly convex, then for every $u, v \in \Omega$ and $g \in \partial f(u)$,

$$f(v) \geq f(u) + \langle g, v - u \rangle + \frac{\alpha}{2}\|v - u\|^2.$$

- A convex combination of convex functions is convex. The pointwise maximum of convex functions is convex, and the pointwise minimum of concave functions is concave.
- In unconstrained differentiable convex optimization, $\nabla f(w^\star) = 0$ implies that $w^\star$ is a global minimizer.
- For differentiable convex $f$ on a convex set $\Omega$, $w^\star$ minimizes $f$ over $\Omega$ if and only if

$$\langle \nabla f(w^\star), w - w^\star \rangle \geq 0 \qquad \text{for all } w \in \Omega.$$

# 5 Questions and Solutions

1. Let $A \in \mathbb{R}^{m\times d}$ and $x \in \mathbb{R}^d$. State the dimensions of $Ax$, $A^\top A$, $AA^\top$, and $xx^\top$. Prove that $A^\top A \succeq 0$.

   **Solution.** Since $A \in \mathbb{R}^{m\times d}$ and $x \in \mathbb{R}^d$,

   $$Ax \in \mathbb{R}^m, \qquad A^\top A \in \mathbb{R}^{d\times d}, \qquad AA^\top \in \mathbb{R}^{m\times m}, \qquad xx^\top \in \mathbb{R}^{d\times d}.$$

   Moreover, $A^\top A$ is symmetric, and for every $z \in \mathbb{R}^d$,

   $$z^\top A^\top Az = (Az)^\top (Az) = \|Az\|_2^2 \geq 0.$$

   Hence $A^\top A \succeq 0$.

2. Let $n_1, \ldots, n_K$ be nonnegative integers satisfying $\sum_{i=1}^K n_i = T$, where $T$ is a positive integer. Prove that

   $$\sum_{i=1}^K \sum_{s=1}^{n_i} \frac{1}{\sqrt{s}} \leq 2\sqrt{KT},$$

   where a sum with upper limit zero is interpreted as zero.

   **Solution.** For every integer $n \geq 1$, since $x \mapsto x^{-1/2}$ is decreasing,

   $$\sum_{s=1}^n \frac{1}{\sqrt{s}} \leq 1 + \int_1^n \frac{1}{\sqrt{x}}\, dx = 2\sqrt{n} - 1 \leq 2\sqrt{n}.$$

   The same upper bound is trivially true when $n = 0$. Therefore,

   $$\sum_{i=1}^K \sum_{s=1}^{n_i} \frac{1}{\sqrt{s}} \leq 2\sum_{i=1}^K \sqrt{n_i}.$$

   By Cauchy–Schwarz inequality, we have

   $$\sum_{i=1}^K \sqrt{n_i} \leq \sqrt{K\sum_{i=1}^K n_i} = \sqrt{KT}.$$

   Combining the above two inequalities gives the desired result.

3. Let $C, T > 0$. Over all choices of $a_i > 0$ and $n_i \geq 0$, $i = 1, \ldots, K$, satisfying $\sum_{i=1}^K n_i \leq T$ and $n_i a_i^2 \leq C$ for all $i \in [K]$. Determine the largest possible value of $\sum_{i=1}^K n_i a_i$ and characterize all equality cases, i.e., when this largest possible value is achieved.

   **Solution.** For each $i$, we have

   $$n_i a_i = \sqrt{n_i}\sqrt{n_i a_i^2} \leq \sqrt{Cn_i}.$$

   Hence

   $$\sum_{i=1}^K n_i a_i \leq \sqrt{C}\sum_{i=1}^K \sqrt{n_i} \leq \sqrt{CK\sum_{i=1}^K n_i} \leq \sqrt{CKT},$$

   where the second inequality is Cauchy–Schwarz. Thus the largest possible value is $\sqrt{CKT}$.

We now characterize equality. Equality in the final inequality requires $\sum_i n_i = T$. Equality in Cauchy–Schwarz requires $\sqrt{n_1} = \cdots = \sqrt{n_K}$, so $n_i = \frac{T}{K}$ for all $i \in [K]$. Since these values are positive, equality in the first inequality further requires $n_i a_i^2 = C$ for every $i$. Therefore, when equality holds, we have $a_i = \sqrt{\frac{KC}{T}}$ for all $i \in [K]$. Conversely, these choices satisfy all constraints and attain $\sqrt{CKT}$. Hence they are exactly the equality cases.

4. A message is junk with prior probability 0.2. A keyword appears in 70% of junk messages and 10% of legitimate messages. Compute the probability that a message is junk given that the keyword appears.

   **Solution.** Let $J$ be the event that the message is junk and $K$ the event that the keyword appears. Then
   $$\Pr(J) = 0.2, \qquad \Pr(K \mid J) = 0.7, \qquad \Pr(K \mid J^c) = 0.1.$$
   By Bayes' rule,
   $$\Pr(J \mid K) = \frac{\Pr(K \mid J)\Pr(J)}{\Pr(K \mid J)\Pr(J) + \Pr(K \mid J^c)\Pr(J^c)} = \frac{0.7 \cdot 0.2}{0.7 \cdot 0.2 + 0.1 \cdot 0.8} = \frac{7}{11}.$$

5. Let
   $$\ell(w) = \log(1 + e^{-y\langle w, x\rangle})$$
   with $x = (1, 2)^\top$, $y = 1$, and $w = (0, -1)^\top$.

   (a) Compute the gradient at the given $w$.

   (b) Derive the Hessian for general $w$ and prove that it is positive semidefinite.

   **Solution.** Let $z = y\,\langle w, x\rangle$. Since
   $$\frac{d}{dz}\log(1 + e^{-z}) = -\frac{1}{1 + e^z},$$
   the chain rule gives
   $$\nabla \ell(w) = -\frac{yx}{1 + e^{y\langle w, x\rangle}}.$$
   At the specified point,
   $$\langle w, x\rangle = 0 \cdot 1 + (-1) \cdot 2 = -2,$$
   and therefore
   $$\nabla \ell(w) = -\frac{1}{1 + e^{-2}} \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$
   For general $w$, differentiating once more gives
   $$\nabla^2 \ell(w) = \frac{y^2 e^{y\langle w, x\rangle}}{(1 + e^{y\langle w, x\rangle})^2} xx^\top.$$
   The scalar coefficient is nonnegative. Moreover, for every $v \in \mathbb{R}^d$,
   $$v^\top xx^\top v = (x^\top v)^2 \geq 0.$$
   Hence $xx^\top \succeq 0$, and therefore $\nabla^2 \ell(w) \succeq 0$.

6. Consider the rates

$$100T^{2/3}, \quad \log\log T, \quad 0.01T, \quad \frac{T}{\log T}, \quad \sqrt{T}, \quad 3(\log T)^2,$$
$$263\sqrt{T\log T}, \quad \frac{T}{\log\log T}, \quad 159T^{\frac{1}{\log T}}, \quad \log T, \quad T + 314\sqrt{T}.$$

Assume $T$ is sufficiently large that all expressions are defined.

(a) Arrange the rates from smallest to largest, using $\Theta(\cdot)$ when two expressions have the same asymptotic order.

(b) Identify exactly which rates are $o(T)$ and therefore yield a no-regret guarantee.

(c) Identify exactly which rates are $O(\sqrt{T})$.

**Solution.** First note that, taking log to denote the natural logarithm,

$$T^{1/\log T} = e,$$

so $159T^{1/\log T} = \Theta(1)$. Thus, from smallest to largest,

$$\begin{aligned} 159T^{1/\log T} &< \log\log T < \log T < 3(\log T)^2 < \sqrt{T} \\ &< 263\sqrt{T\log T} < 100T^{2/3} < \frac{T}{\log T} < \frac{T}{\log\log T} \\ &< 0.01T = \Theta\big(T + 314\sqrt{T}\big), \end{aligned}$$

for all sufficiently large $T$. In particular, the last two expressions are both $\Theta(T)$.

The rates that are $o(T)$ are

$$100T^{2/3}, \quad \log\log T, \quad \frac{T}{\log T}, \quad \sqrt{T}, \quad 3(\log T)^2, \quad 263\sqrt{T\log T}, \quad \frac{T}{\log\log T}, \quad 159T^{1/\log T}, \quad \log T.$$

The two rates that are not $o(T)$ are $0.01T$ and $T + 314\sqrt{T}$.

The rates that are $O(\sqrt{T})$ are

$$\log\log T, \qquad \sqrt{T}, \qquad 3(\log T)^2, \qquad 159T^{1/\log T}, \qquad \log T.$$

7. Let $A, B, C \in \mathbb{R}^{d\times d}$, and assume that every inverse appearing below exists. For each statement, either prove it or give a concrete counterexample:

(a) $(AB)^{-1} = B^{-1}A^{-1}$,
(b) $(I + A)^{-1} = I - A$,
(c) $\operatorname{tr}(ABC) = \operatorname{tr}(CAB)$,
(d) $(AB)^\top = A^\top B^\top$.

For every false identity, state a correct replacement or a sufficient additional condition.

**Solution.**

(a) This identity is true. Indeed,

$$(AB)(B^{-1}A^{-1}) = A(BB^{-1})A^{-1} = I,$$

and similarly $(B^{-1}A^{-1})(AB) = I$. Hence $(AB)^{-1} = B^{-1}A^{-1}$.

(b) This identity is false in general. For example, in dimension one, take $A = 1$. Then

$$(I + A)^{-1} = \frac{1}{2}, \qquad I - A = 0.$$

A correct general statement is the Neumann-series identity

$$(I + A)^{-1} = I - A + A^2 - A^3 + \cdots$$

whenever, for example, $\|A\| < 1$ for a submultiplicative matrix norm. The displayed identity $(I + A)^{-1} = I - A$ does hold under the sufficient condition $A^2 = 0$.

(c) This identity is true. By cyclicity of the trace,

$$\operatorname{tr}(ABC) = \operatorname{tr}(BCA) = \operatorname{tr}(CAB).$$

(d) This identity is false in general. The correct identity is

$$(AB)^\top = B^\top A^\top.$$

For example, let

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad B = \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}.$$

Then

$$(AB)^\top = \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \qquad A^\top B^\top = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}.$$

8. Let $P, Q \in \mathbb{R}^{d \times d}$ be symmetric and positive semidefinite. Decide whether each matrix below must be positive semidefinite. Give a proof for every true claim and an explicit $2 \times 2$ counterexample for every false claim:

$$P + Q, \qquad \alpha P \ \ (\alpha \geq 0), \qquad P - Q, \qquad PQ, \qquad PQP.$$

Pay attention to the symmetry requirement in the definition of positive semidefiniteness.

**Solution.**

- $P + Q$ must be positive semidefinite. It is symmetric, and for every $x$,

$$x^\top (P + Q)x = x^\top Px + x^\top Qx \geq 0.$$

- $\alpha P$ must be positive semidefinite for every $\alpha \geq 0$. It is symmetric, and

$$x^\top (\alpha P)x = \alpha x^\top Px \geq 0.$$

- $P - Q$ need not be positive semidefinite. Take

$$P = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}, \qquad Q = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Then $P, Q \succeq 0$, but $P - Q = -I \not\succeq 0$.

- $PQ$ need not be positive semidefinite. Take

$$P = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \qquad Q = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}.$$

  Both matrices are symmetric and positive semidefinite, but we have $PQ = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}$, which is not symmetric. Hence it is not positive semidefinite under the stated definition.

- $PQP$ must be positive semidefinite. It is symmetric because

$$(PQP)^\top = P^\top Q^\top P^\top = PQP,$$

  and for every $x$,

$$x^\top PQPx = (Px)^\top Q(Px) \geq 0.$$

9. Let

$$f(w) = w^\top Aw + b^\top w + c,$$

where $A \in \mathbb{R}^{d \times d}$ is not assumed symmetric.

(a) Derive $\nabla f(w)$ and $\nabla^2 f(w)$.

(b) Give a necessary and sufficient condition on $A$ for $f$ to be convex.

(c) Suppose $A + A^\top \succ 0$. Find the unique minimizer.

**Solution.** Since

$$w^\top Aw = \sum_{i,j} A_{ij} w_i w_j,$$

differentiation gives

$$\nabla f(w) = (A + A^\top)w + b, \qquad \nabla^2 f(w) = A + A^\top.$$

Therefore, $f$ is convex if and only if

$$A + A^\top \succeq 0,$$

equivalently, if and only if the symmetric part $(A + A^\top)/2$ is positive semidefinite.

If $A + A^\top \succ 0$, then $f$ is strictly convex and therefore has at most one minimizer. The first-order condition is

$$(A + A^\top)w^\star + b = 0,$$

so the unique minimizer is

$$w^\star = -(A + A^\top)^{-1} b.$$

10. Consider $\min_{x \in \Omega} F(x)$, where $\Omega \subseteq \mathbb{R}^d$ is convex and $F : \mathbb{R}^d \to \mathbb{R}$ is differentiable and convex. Show that every minimizer $x^\star$ satisfies

$$\langle \nabla F(x^\star), x - x^\star \rangle \geq 0 \qquad \text{for all } x \in \Omega.$$

**Solution.** Fix any $x \in \Omega$. Since $\Omega$ is convex,

$$x^\star + t(x - x^\star) \in \Omega \qquad \text{for every } t \in [0, 1].$$

Because $x^\star$ is a minimizer,

$$F\big(x^\star + t(x - x^\star)\big) - F(x^\star) \geq 0 \qquad \text{for every } t \in (0, 1].$$

Dividing by $t > 0$ and letting $t$ approach 0, differentiability gives

$$\langle \nabla F(x^\star), x - x^\star \rangle \geq 0.$$

Since $x \in \Omega$ was arbitrary, the result follows. Note that in fact, we have never used the fact that $F(x)$ is convex here. Therefore, first-order optimality does **not** require the convexity of the function, but only requires the convexity of the domain.