

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

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

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

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

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

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

whenever dimensions and inverses are valid.

- A symmetric matrix Q is positive semidefinite, written $Q \succeq 0$, if $x^\top Qx \geq 0$ 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 λ 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:

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

- If $Q = Q^\top$, then

$$Q \succeq 0 \iff \lambda_i(Q) \geq 0 \text{ for every eigenvalue.}$$

The spectral theorem gives $Q = U\Lambda U^\top$ with orthogonal U .

- Eigenvalue properties

$$\lambda_{\min}(Q) \leq \frac{x^\top Q x}{\|x\|_2^2} \leq \lambda_{\max}(Q) \quad (x \neq 0),$$

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

$$\det(Q + x x^\top) = \det(Q)(1 + x^\top Q^{-1} x)$$

and

$$(Q + x x^\top)^{-1} = Q^{-1} - \frac{Q^{-1} x x^\top Q^{-1}}{1 + x^\top Q^{-1} x}.$$

- For $f(w) = \frac{1}{2} \|Aw - b\|_2^2$,

$$\nabla f(w) = A^\top (Aw - b), \quad \nabla^2 f(w) = A^\top A \succeq 0.$$

2.2 Norms and dual norms

- The norms used most often are

$$\|x\|_1 = \sum_i |x_i|, \quad \|x\|_2 = \sqrt{\sum_i x_i^2}, \quad \|x\|_\infty = \max_i |x_i|.$$

- The dual norm is

$$\|y\|_* = \max_{\|x\| \leq 1} \langle x, y \rangle.$$

The dual of ℓ_2 -norm is ℓ_2 -norm; the dual of ℓ_1 -norm is ℓ_∞ -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

$$\langle x, y \rangle \leq \|x\|_2 \|y\|_2, \quad \langle x, y \rangle \leq \|x\|_1 \|y\|_\infty.$$

2.3 Convex sets, the simplex, and projection

- A set Ω 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]**

$$[\Pi_{\Delta(d)}(y)]_i = \max\{y_i - \tau, 0\},$$

where τ 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

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad \mathbb{P}(A | B) = \frac{\mathbb{P}(B | A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

- Linearity of expectation does not require independence: for n possibly dependent random variables $X_1, \dots, X_n$, 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 $\mathbb{P}(A | B) = \mathbb{P}(A)$. If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$ when the expectations exist. The reverse implications do not generally hold.
- Variance and covariance are defined as

$$\text{Var}(X) = \mathbb{E}[(X - \mathbb{E}X)^2], \quad \text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}X)(Y - \mathbb{E}Y)].$$

- For a random vector X with mean μ , its covariance matrix is positive semidefinite:

$$\text{Cov}(X) = \mathbb{E}[(X - \mu)(X - \mu)^\top] = \mathbb{E}[XX^\top] - \mu\mu^\top \succeq 0.$$

- 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 $\mathbb{E}[M_t \mid M_0, \dots, M_{t-1}] = M_{t-1}$. Equivalently, its increments $X_t := M_t - M_{t-1}$ satisfy $\mathbb{E}[X_t \mid M_0, \dots, M_{t-1}] = 0$. The increments need not be independent.*

Example 3.2. *Let $Y_1, Y_2, \dots$ 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 $\mathbb{E}[M_t \mid M_0, \dots, M_{t-1}] = M_{t-1}$. 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 $X_1, \dots, X_T$ 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 $t = 1, \dots, T$. Then, for every $\epsilon > 0$,*

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

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

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

- 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 $g : \mathbb{R}^d \rightarrow \mathbb{R}^m$, 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 $f : \Omega \rightarrow \mathbb{R}$ is convex if Ω 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

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

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

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

- If f is twice differentiable on an open convex set Ω , then f is convex if and only if $\nabla^2 f(w) \succeq 0$ for every $w \in \Omega$.
- If f is α -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, $\nabla f(w^*) = 0$ implies that w^* is a global minimizer.
- For differentiable convex f on a convex set Ω , w^* minimizes f over Ω if and only if

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

5 Questions

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 $A^\top A \succeq 0$.
2. Let $n_1, \dots, n_K$ 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.

3. Let $C, T > 0$. Over all choices of $a_i > 0$ and $n_i \geq 0$, $i = 1, \dots, K$, satisfying

$$\sum_{i=1}^K n_i \leq T \quad \text{and} \quad n_i a_i^2 \leq C \quad \text{for every } i,$$

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.

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.
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.
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 $O(\sqrt{T})$.
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) $\text{tr}(ABC) = \text{tr}(CAB)$,
- (d) $(AB)^\top = A^\top B^\top$.

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

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 2×2 counterexample for every false claim:

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

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

9. Let

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

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.

10. Consider

$$\min_{x \in \Omega} F(x),$$

where $\Omega \subseteq \mathbb{R}^d$ is convex and $F : \mathbb{R}^d \rightarrow \mathbb{R}$ is differentiable and convex. Show that every minimizer x^* satisfies

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