ECE 602 – Section 1
Mathematical preliminaries

- Norms and inner products in $\mathbb{R}^n$ and $\mathbb{R}^{m \times n}$
- Open sets, closed sets and closed functions
- Range, nullspace, orthogonal complement and direct sum
- SVD, EVD, positive definiteness, and pseudo-inverse
- Differential, derivative, gradient, and Hessian

Inner product, Euclidean norm, and angle

- The standard *inner product* on $\mathbf{R}^n$ is given by

$$\langle x, y\rangle = x^T y = \sum_{i=1}^{n} x_i y_i, \quad \text{for } x, y \in \mathbf{R}^n,$$

with x and y viewed as *column vectors*.

- The *Euclidean norm*, or ℓ_2-norm, of $x \in \mathbf{R}^n$ is defined as

$$\|x\|_2 = \langle x, x\rangle^{1/2} = \left(x^T x\right)^{1/2} = \left(x_1^2 + x_2^2 + \ldots + x_n^2\right)^{1/2}.$$

- The *Cauchy-Schwartz inequality* states that

$$|x^T y| \leq \|x\|_2 \|y\|_2, \quad \text{for } x, y \in \mathbf{R}^n.$$

- The *angle* between $x, y \in \mathbf{R}^n$ is defined as

$$\angle(x, y) = \cos^{-1}\left(\frac{x^T y}{\|x\|_2 \, \|y\|_2}\right) \in [0, \pi].$$

- We say x and y are *orthogonal* if $x^T y = 0$.

- The standard inner product and its related norm can also be defined on $\mathbf{R}^{m \times n}$ (i.e., the linear space of $m \times n$ matrices).

- The standard *inner product* on $\mathbf{R}^{m \times n}$ is given by

$$\langle X, Y \rangle = \sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij} Y_{ij} = \mathbf{tr}(X^T Y), \quad \text{for } X, Y \in \mathbf{R}^{m \times n},$$

  where **tr** stands for the *trace* of a matrix.

- The *Frobenius norm* of a matrix $X \in \mathbf{R}^{m \times n}$ is given by

$$\|X\|_F = \langle X, X \rangle^{1/2} = \left(\mathbf{tr}(X^T X) \right)^{1/2} = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} X_{ij}^2 \right)^{1/2},$$

  which makes it analogous to the ℓ_2-norm in $\mathbf{R}^n$.

- Equivalence between the inner products on $\mathbf{R}^n$ and $\mathbf{R}^{m\times n}$ can be established as follows.

- Let vec and mat denote the operations of *vectorization* and *matricization* defined as

$$\operatorname{vec}\left(\begin{bmatrix} a & b & c \\ d & f & g \end{bmatrix}\right) = \begin{bmatrix} a \\ d \\ b \\ f \\ c \\ g \end{bmatrix}, \qquad \operatorname{mat}\left(\begin{bmatrix} a \\ d \\ b \\ f \\ c \\ g \end{bmatrix}\right) = \begin{bmatrix} a & b & c \\ d & f & g \end{bmatrix}.$$

(Note that mat is assumed to "know" the size of the output matrix.)

- Then it is easy to show that, for $x = \operatorname{vec}(X)$ and $y = \operatorname{vec}(Y)$, we have

$$x^T y = \operatorname{vec}(X)^T \operatorname{vec}(Y) = \mathbf{tr}(\operatorname{mat}(x)^T \operatorname{mat}(y)) = \mathbf{tr}(X^T Y).$$

- In general, norms are defied axiomatically.
- A function $f : \mathbf{R}^n \to \mathbf{R}$ with $\mathbf{dom}\, f = \mathbf{R}^n$ is called a *norm* if
  1. $f(x) \geq 0, \ \forall x$ (non-negative)
  2. $f(x) = 0, \ \text{only if } x = 0$ (definite)
  3. $f(tx) = |t| f(x), \ \forall x \in \mathbf{R}^n, t \in \mathbf{R}$ (homogeneous)
  4. $f(x+y) \leq f(x) + f(y), \ \forall x, y \in \mathbf{R}^n$ (obeys the triangle inequality)
- We denote $f(x) = \|x\|$ (which can be interpreted as the "length" of x).
- It turns out there are many possible norms that fit the above definition.

Examples of norms

- The *sum-absolute-value*, or ℓ_1-*norm*, is defined as

$$\|x\|_1 = |x_1| + \ldots + |x_n|.$$

- The *Chebyshev*, or ℓ_∞-*norm*, is defined as

$$\|x\|_\infty = \max\{|x_1|, \ldots, |x_n|\}.$$

- The ℓ_p-norm is defined as

$$\|x\|_p = (|x_1|^p + \ldots + |x_n|^p)^{1/p}.$$

- The *quadratic norm* w.r.t. some $P \in \mathbf{S}^n_{++}$ (i.e., the set of symmetric positive definite matrices) is defined as

$$\|x\|_P = (x^T P x)^{1/2} = \|P^{1/2} x\|_2,$$

where $P^{1/2}$ is the *square root* of P, i.e., $P^{1/2} P^{1/2} = P$.

- In addition to the Frobenius norm, for any $X \in \mathbf{R}^{m\times n}$, one can define the *sum-absolute-value norm* to be

$$\|X\|_{\text{sav}} = \sum_{i=1}^{m}\sum_{j=1}^{n} |X_{ij}|.$$

- The *maximum-absolute-value norm* is defined as

$$\|X\|_{\text{mav}} = \max\left\{|X_{ij}| \mid i = 1, \ldots, m,\ j = 1, \ldots, n\right\}.$$

- Note that, in finite dimensional spaces (like $\mathbf{R}^n$ or $\mathbf{R}^{m\times n}$), *all norms are equivalent*, which means that:

$$\text{For any } \|\cdot\|_a \text{ and } \|\cdot\|_b :\ \exists\, 0 < A, B < \infty, \text{ such that}$$

$$A\,\|\xi\|_a \le \|\xi\|_b \le B\,\|\xi\|_a, \quad \forall \xi,$$

with the norms and ξ being defined either in $\mathbf{R}^n$ or $\mathbf{R}^{m\times n}$.

- Norms are useful for defining distances.
- The *distance* between x and y can be defined as

$$\mathbf{dist}(x,y) = \|x - y\|.$$

- The *unit ball* of the norm $\|\cdot\|$ is defined as

$$\mathcal{B} = \left\{x \in \mathbf{R}^n \mid \|x\| \leq 1\right\}.$$

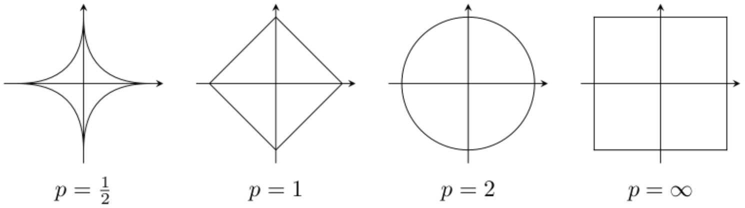

Figure: Unit balls for different ℓ_p-norms

- Suppose $\|\cdot\|_a$ and $\|\cdot\|_b$ are norms on $\mathbf{R}^m$ and $\mathbf{R}^n$, respectively. Then, the *operator norm* of $\mathbf{R}^{m\times n}$ is defined as

$$\|X\|_{a,b} = \sup\left\{\|Xu\|_a \mid \|u\|_b \leq 1\right\}.$$

- For $a = b = 2$, we get the *spectral norm* of X defined as its maximum *singular value*

$$\|X\|_2 = \sigma_{\max}(X).$$

- For $a = b = \infty$, we get the *max-row-sum norm* of X defined as

$$\|X\|_\infty = \max_i \sum_{j=1}^{n} |X_{ij}|.$$

- For $a = b = 1$, we get the *max-column-sum norm* of X defined as

$$\|X\|_1 = \max_j \sum_{i=1}^{m} |X_{ij}|.$$

Dual norm

- Let $\|\cdot\|$ be a norm on $\mathbf{R}^n$. Its associated *dual norm* is defined as

$$\|z\|_* = \sup\left\{z^T x \mid \|x\| \leq 1\right\}.$$

- For all x and z we have: $|z^T x| \leq \|x\|\|z\|_*$ (tight).
- The dual of $\|\cdot\|_*$ is $\|\cdot\|$.
- The dual of $\|\cdot\|_2$ is $\|\cdot\|_2$ (Cauchy-Schwarz).
- The dual of $\|\cdot\|_p$ is $\|\cdot\|_q$, where $1/p + 1/q = 1$ (Hölder).
- $\|\cdot\|_\infty$ and $\|\cdot\|_1$ are dual w.r.t. each other.

Open and closed sets

- $x \in C \subseteq \mathbf{R}^n$ is an *interior point* if $\exists\, \epsilon > 0$ such that

$$\{y \mid \|y - x\|_2 \leq \epsilon\} \subseteq C.$$

  All such points constitute the *interior* of C, $\mathbf{int}\, C$.

- A set C is *open* if $\mathbf{int}\, C = C$.

- A set C is *closed* if $\mathbf{R}^n \backslash C$ is open.

- The *closure* of a set C is defined as

$$\mathbf{cl}\, C = \mathbf{R}^n \backslash \mathbf{int}\, (\mathbf{R}^n \backslash C).$$

- The *boundary* of the set C is defined as

$$\mathbf{bd}\, C = \mathbf{cl}\, C \backslash \mathbf{int}\, C.$$

Functions

- We denote by
$$f : A \to B$$
a *function* defined on the set $\mathbf{dom}\, f \subseteq A$ into set B.

- As an example consider the function $f : \mathbf{S}^n \to \mathbf{R}$, given by
$$f(X) = \log\det X,$$
with $\mathbf{dom}\, f = \mathbf{S}^n_{++}$.

- A function f is *closed* if, for each $\alpha \in \mathbf{R}$, the sublevel set
$$\{x \in \mathbf{dom}\, f \mid f(x) \le \alpha\}$$
is closed.

- Any closed function f approaches infinity, as its argument approaches the boundary of $\mathbf{dom}\, f$.

- Let $A \in \mathbf{R}^{m\times n}$. The *range* of A is a subspace of $\mathbf{R}^m$ defined as

$$\mathcal{R}(A) = \{Ax \mid x \in \mathbf{R}^n\}.$$

- The dimension of $\mathcal{R}$ is the *rank* of A, denoted $\mathbf{rank}\, A$.

- We say A has *full rank* if $\mathbf{rank}\, A = \min\{n, m\}$.

- The *nullspace* (or *kernel*) of A is defined as

$$\mathcal{N}(A) = \{x \mid Ax = 0\},$$

  which is a subspace of $\mathbf{R}^n$.

- If $\mathcal{V}$ is a subspace of $\mathbf{R}^n$, its *orthogonal complement* is defined as

$$\mathcal{V}^{\perp} = \left\{ x \mid z^T x = 0, \forall z \in \mathcal{V} \right\}.$$

- For any $A \in \mathbf{R}^{m \times n}$, we have

$$\mathcal{N}(A) = \mathcal{R}(A^T)^{\perp}, \quad \mathcal{R}(A) = \mathcal{N}(A^T)^{\perp}.$$

- The above results can also be stated as

$$\mathcal{N}(A) \oplus \mathcal{R}(A^T) = \mathbf{R}^n,$$

where the symbol $\oplus$ refers to *orthogonal direct sum*.

Eigenvalue decomposition

- Any real symmetric matrix $A \in \mathbf{S}^n$ can be factored as
$$A = Q \Lambda Q^T,$$
where Q is *orthogonal* (i.e., $Q^T Q = I$), and $\Lambda = \mathbf{diag}\,(\lambda_1, \ldots, \lambda_n)$.

- This is called the *eigenvalue decomposition* of A, with λ_i being the *eigenvalues* of A.

- The determinant and trace of A can be expressed as
$$\det A = \prod_{i=1}^{n} \lambda_i, \qquad \mathbf{tr}\, A = \sum_{i=1}^{n} \lambda_i.$$

- The spectral and Frobenius norms of $A \in \mathbf{S}^n$ can be expressed as
$$\|A\|_2 = \|\lambda\|_\infty, \qquad \|A\|_F = \|\lambda\|_2.$$

Definiteness

- For any $A \in \mathbf{S}^n$, we have

$$\lambda_{\max}(A) = \sup_{\|x\|_2=1} x^T Ax, \quad \lambda_{\min}(A) = \inf_{\|x\|_2=1} x^T Ax.$$

- This suggests that, for any x with $\|x\| = 1$, one has

$$\lambda_{\min}(A) \leq x^T A\, x \leq \lambda_{\max}(A).$$

- If $x^T Ax > 0, \forall x$, then A is called *positive definite* ($A \in \mathbf{S}^n_{++}$ or $A \succ 0$). In this case, $\lambda_{\min}(A) > 0$.

- If $x^T Ax \geq 0, \forall x$, then A is called *positive semi-definite* ($A \in \mathbf{S}^n_{+}$ or $A \succeq 0$). In this case, $\lambda_{\min}(A) \geq 0$.

- For $A, B \in \mathbf{S}^n$, we use $A \succ B$ to mean that $A - B \succ 0$.

- Let $A \in \mathbf{S}^n_+$, with eigenvalue decomposition $A = Q\,\mathbf{diag}(\lambda_1, \ldots, \lambda_n)Q^T$.

- The *symmetric square root* of A is defined as

$$A^{1/2} = Q\,\mathbf{diag}(\lambda_1^{1/2}, \ldots, \lambda_n^{1/2})Q^T.$$

- This is the *unique* symmetric positive semidefinite solution of

$$X^2 = A.$$

Singular value decomposition

- Suppose $A \in \mathbf{R}^{m \times n}$ with $\mathbf{rank}\, A = r$. Then A can be factored as

$$A = U \Sigma V^T,$$

  where
  - $U \in \mathbf{R}^{m \times r}$ satisfies $U^T U = I$
  - $V \in \mathbf{R}^{n \times r}$ satisfies $V^T V = I$
  - $\Sigma = \mathbf{diag}(\sigma_1, \ldots, \sigma_r)$, with $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$.

- This factorization is called the *singular value decomposition* (SVD) of A, with $\sigma_i \geq 0$ being the *singular values* of A.

- The SVD of A is closely related to the eigenvalue decomposition of $A^T A$ and $A A^T$ (how?).

Pseudo-inverse

- Let $A = U\Sigma V^T$, with $\mathbf{rank}\, A = r$.
- The *pseudo inverse* or *Moore-Penrose inverse* of A is defined as

$$A^\dagger = V\Sigma^{-1}U^T \in \mathbf{R}^{m\times n}.$$

- If $\mathbf{rank}\, A = n$, then $A^\dagger = (A^TA)^{-1}A^T$.
- If $\mathbf{rank}\, A = m$, then $A^\dagger = A^T(AA^T)^{-1}$.
- If A is square and has a full rank, then $A^\dagger = A^{-1}$.
- $A^\dagger b$ is a solution to the *least-square (LS) problem*

$$\min_x \|Ax - b\|_2^2.$$

External derivative and gradient

- Given a real-valued $f \in \mathcal{C}^1(\mathbf{R}^n)$, its *total differential* at $x^* \in \operatorname{dom} f$ is defined as

$$df(x^*) = \frac{\partial f}{\partial x_1}(x^*)dx_1 + \frac{\partial f}{\partial x_2}(x^*)dx_2 + \ldots + \frac{\partial f}{\partial x_n}(x^*)dx_n = \sum_{i=1}^{n} \frac{\partial f}{\partial x_i}(x^*)dx_i.$$

- Note that $df(x^*)$ has the form of an inner product, namely

$$df(x^*) = \langle g(x^*), dx \rangle,$$

where $dx = [dx_1, dx_2, \ldots, dx_n]^T$ and

$$g(x^*) = \nabla f(x^*) = \left[\frac{\partial f}{\partial x_1}(x^*), \frac{\partial f}{\partial x_2}(x^*), \ldots, \frac{\partial f}{\partial x_n}(x^*) \right]^T,$$

which is called the *gradient* of f at x^*.

- The expression $df(x) = \langle g(x), dx \rangle$ is known as the *external definition of the gradient*.

- Recall that, for a single-variate f, we have $df(x) = f'(x)dx$.
- f' is just a function of x that has its total differential defined as

$$df'(x) = f''(x)dx,$$

  with f'' being the second-order derivative of f.
- In the case when f is multi-variate, we have

$$df(x) = \langle g(x), dx \rangle = g^T(x)dx.$$

- When $f \in \mathcal{C}^2(\mathbf{R}^n)$, the gradient $g(x)$ can be viewed as a function from $\mathbf{R}^n$ to $\mathbf{R}^n$ that obeys the differential form given by

$$dg(x) = H(x)dx,$$

  where $H(x) \in \mathbf{S}^n$ is called the *Hessian* of $f(x)$ at x.
- The above formula gives the *external definition of the Hessian*.

Hessian (cont.)

- Explicitly, the Hessian matrix $H(x)$ can be defined as

$$H(x) = \nabla^2 f(x) = \begin{pmatrix} \frac{\partial^2 f(x)}{\partial x_1 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n \partial x_n} \end{pmatrix}.$$

- Alternatively, we can write

$$H(x) = \begin{pmatrix} \frac{\partial g_1(x)}{\partial x_1} & \frac{\partial g_1(x)}{\partial x_2} & \cdots & \frac{\partial g_1(x)}{\partial x_n} \\ \frac{\partial g_2(x)}{\partial x_1} & \frac{\partial g_2(x)}{\partial x_2} & \cdots & \frac{\partial g_2(x)}{\partial x_n} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial g_n(x)}{\partial x_1} & \frac{\partial g_n(x)}{\partial x_2} & \cdots & \frac{\partial g_n(x)}{\partial x_n} \end{pmatrix} = \begin{pmatrix} \nabla g_1^T(x) \\ \nabla g_2^T(x) \\ \cdots \\ \nabla g_n^T(x) \end{pmatrix},$$

where

$$g(x) = \begin{bmatrix} g_1(x) \\ g_2(x) \\ \cdots \\ g_n(x) \end{bmatrix} = \begin{bmatrix} \partial f(x)/\partial x_1 \\ \partial f(x)/\partial x_2 \\ \cdots \\ \partial f(x)/\partial x_n \end{bmatrix}.$$

Examples

- Differential of a linear operator $y = Ax$.

$$dy = A(x + dx) - Ax = A\,dx.$$

- Differential of a linear function $f(x) = b^T x$.

$$df(x) = b^T(x + dx) - b^T x = b^T dx.$$

Comparing with $df(x) = g^T(x)dx$ reveals that $\nabla f(x) = b$.

- Differential of the quadratic form $f(x) = x^T Ax$.

$$df(x) = (x + dx)^T A(x + dx) - x^T Ax \simeq x^T Adx + dx^T Ax =$$
$$= x^T Adx + x^T A^T dx = (x^T A + x^T A^T)dx = \Big((A + A^T)x\Big)^T dx.$$

Consequently, $\nabla f(x) = g(x) = (A + A^T)x$ and $H(x) = A + A^T$.

- Let $X \in \mathbf{R}^{m \times n}$ and let $f(X)$ be a scalar-valued function of X.
- Explicitly, the gradient of f is defined as

$$G(X) = \nabla f(X) = \begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \frac{\partial f}{\partial x_{12}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \frac{\partial f}{\partial x_{21}} & \frac{\partial f}{\partial x_{22}} & \cdots & \frac{\partial f}{\partial x_{2n}} \\ \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f}{\partial x_{n1}} & \frac{\partial f}{\partial x_{n2}} & \cdots & \frac{\partial f}{\partial x_{nn}} \end{pmatrix}$$

- With dX defined as

$$dX = \begin{pmatrix} dx_{11} & dx_{12} & \cdots & dx_{1n} \\ dx_{21} & dx_{22} & \cdots & dx_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ dx_{n1} & dx_{n2} & \cdots & dx_{nn} \end{pmatrix},$$

the *total differential* of $f(X)$ is given by

$$\boxed{df(X) = \sum_{i,j} \frac{\partial f(X)}{\partial x_{ij}} dx_{ij} = \mathbf{tr}\, G^T(X) dX = \langle G(X), dX \rangle}$$

- Let $F : \mathbf{R}^n \to \mathbf{R}^m$ be a continuously differentiable function of the form

$$F(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \dots \\ f_m(x) \end{pmatrix}.$$

- The *Jacobian* of F is defined as

$$J_F(x) = \begin{pmatrix} \frac{\partial f_1(x)}{\partial x_1} & \frac{\partial f_1(x)}{\partial x_2} & \dots & \frac{\partial f_1(x)}{\partial x_n} \\ \frac{\partial f_2(x)}{\partial x_1} & \frac{\partial f_2(x)}{\partial x_2} & \dots & \frac{\partial f_2(x)}{\partial x_n} \\ \dots & \dots & \dots & \dots \\ \frac{\partial f_m(x)}{\partial x_1} & \frac{\partial f_m(x)}{\partial x_2} & \dots & \frac{\partial f_m(x)}{\partial x_n} \end{pmatrix},$$

which is an $m \times n$ matrix.

- The total differential of F is then defined as

$$dF(x) = J_F(x)dx,$$

which can be viewed as a vector of the total differentials of f_i.

- First note that, if $m = 1$, than instead of a Jacobian, $F : \mathbf{R}^m \to \mathbf{R}$ has a gradient which can formally be defined as

$$\nabla F(x) = J_F^T(x).$$

- Suppose we are given $F : \mathbf{R}^n \to \mathbf{R}^m$ and $\varphi : \mathbf{R}^m \to \mathbb{R}^k$. Consider

$$\psi(x) = \varphi(F(x)).$$

- Note that $dF = J_F dx$ and $d\varphi = J_\varphi dF$. Then,

$$d\psi = J_\varphi \underbrace{dF}_{J_F\,dx} = \underbrace{J_\varphi J_F}_{J_\psi} dx,$$

and therefore

$$\boxed{J_\psi = J_\varphi J_F}$$

- This is called the *chain rule*.

Important example

- Consider a function of the form $p = h(g(f(x)))$, where h, g, and f are differentiable everywhere within their respective domains, and

$$f : \mathbf{R}^n \to \mathbf{R}^m, \; g : \mathbf{R}^m \to \mathbf{R}^m, \; h : \mathbf{R}^m \to \mathbf{R},$$

  and, therefore, $p : \mathbf{R}^n \to \mathbf{R}$.

- Moreover, function g is assumed to be *diagonal* in the sense that, for any $y \in \mathbf{R}^m$, we have

$$g(y_1, y_2, \ldots, y_m) = \left[\varphi(y_1), \varphi(y_2), \ldots, \varphi(y_m)\right],$$

  for some real-valued $\varphi \in \mathcal{C}^1(\mathbf{R})$.

- The gradient of p is given by $\nabla p(x) = J_p^T(x)$, where

$$J_p = \overbrace{\underbrace{J_h}_{1 \times m} \cdot \underbrace{J_g}_{m \times m} \cdot \underbrace{J_f}_{m \times n}}^{1 \times n}.$$

- In computations, we frequently need to compute $J_p(x)$ for any given $\hat{x} \in \mathbf{R}^n$. In this case, we compute

$$\boxed{J_p(\hat{x}) = \nabla h^T(z)\Big|_{z=g(y)} \cdot J_g(y)\Big|_{y=f(x)} \cdot J_f(x)\Big|_{x=\hat{x}}}$$

- In the special case when, for some fixed values of $A \in \mathbf{R}^{m \times n}$, $b \in \mathbf{R}^m$ and $c \in \mathbf{R}^m$, f and h are defined as

$$f(x) = Ax - b, \quad h(z) = c^T z,$$

we have

$$J_p = \underbrace{c^T}_{J_h} \underbrace{\operatorname{diag}\left(\varphi'(Ax - b)\right)}_{J_g} \underbrace{A}_{J_f}.$$

- The computation of J_p starts on the right by computing J_f first, with its subsequent (left-) multiplicative (recursive) update. Note how, starting with $\hat{x} = x$, we *back-propagate* it to $y = f(x)$ and, subsequently, to $z = g(y) = g(f(x))$.