Probability and Optimization Review

IEMS 402

Abstract. Here is a Review of probability and Optimization Basics for IEMS 402 Statistical Learning.

Table of Content

1 Stochastic Convergence

Guiding question: What does it mean for a sequence of random variables $\{X_n\}_{n\geqslant 1}$ to converge to a random variable X? There will be three versions we will cover. First we give the weakest (also known as weak convergence, just to really drive the idea home):

Definition 1 (Convergence in Distribution)**.** *X_n converges in distribution to X (written $X_n \xrightarrow{d} X$ or $X_n \rightsquigarrow X$ in the book) if for all points $c \in \mathbb{R}^d$ which are continuity points of $x \mapsto \mathbf{Pr}[X \leqslant x]$:*

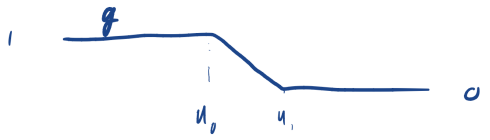
$$\lim_{n\to\infty} \mathbf{Pr}[X_n \leqslant c] = \mathbf{Pr}[X \leqslant c]$$

Lemma 1 (Subset of Portmanteau Lemma)**.**

$$X_n \xrightarrow{d} X \iff \mathbf{E}[f(X_n)] \to \mathbf{E}[f(X)] \text{ for any bounded continuous } f \text{ or any bounded Lipschitz } f$$

Sketch. First the backward direction: Assume $\mathbf{E}[f(X_n)] \to \mathbf{E}[f(X)]$. Fix some arbitrary u_0, u_1 and we can define

$$g(t) = \begin{cases} 1 & t \leqslant u_0 \\ 0 & t > u_1 \\ 1 - \frac{t-u_0}{u_1-u_0} & u_0 < t \leqslant u_1 \end{cases}$$



(a) Function definition (b) Function drawing

Figure 1: Continuous (Lipschitz) Approximation to the Indicator

Then note that $\mathbf{Pr}[X_n \leqslant c] = \mathbf{E}[I_{X_n \leqslant c}] \approx \mathbf{E}[g(X_n)] \to \mathbf{E}[g(X)] = \mathbf{Pr}[X \leqslant c]$ and we can make the approximation arbitrarily good by choice of u_0, u_1.
For the forward direction: Assume $\lim_{n\to\infty} \mathbf{Pr}[X_n \in [a,b]] = \mathbf{Pr}[X \in [a,b]]$ and fix $\epsilon > 0$. Pick a, b such that $\mathbf{Pr}[X \notin [a,b]] < \epsilon$. Since f is bounded and continuous, on this compact set it is uniformly continuous. We can then break the expectation up into a finite number of intervals such that f varies by at most ϵ on the interval. Now define $f_\epsilon = \sum_{i=1}^m f(x_i) 1_{[a_i,b_i]}$ with $x_i \in [a_i, b_i]$ arbitrarily chosen. Then we have:

$$\begin{aligned}
|\mathbf{E}[f(X_n)] - \mathbf{E}[f(X)]| &= |\mathbf{E}[f(X_n)] - \mathbf{E}[f_\epsilon(X_n)] + \mathbf{E}[f_\epsilon(X_n)] - \mathbf{E}[f(X)]| \\
&\leqslant |\mathbf{E}[f(X_n)] - \mathbf{E}[f_\epsilon(X_n)]| + |\mathbf{E}[f_\epsilon(X_n)] - \mathbf{E}[f(X)]| \\
&\leqslant \epsilon + \mathbf{Pr}[X_n \notin [a,b]] + |\mathbf{E}[f_\epsilon(X_n)] - \mathbf{E}[f(X)]| \\
&\leqslant 2\epsilon + |\mathbf{E}[f_\epsilon(X_n)] - \mathbf{E}[f_\epsilon(X)] + \mathbf{E}[f_\epsilon(X)] - \mathbf{E}[f(X)]| \\
&\leqslant 2\epsilon + |\mathbf{E}[f_\epsilon(X_n)] - \mathbf{E}[f_\epsilon(X)]| + |\mathbf{E}[f_\epsilon(X)] - \mathbf{E}[f(X)]| \\
&\leqslant 4\epsilon + |\mathbf{E}[f_\epsilon(X_n)] - \mathbf{E}[f_\epsilon(X)]| \\
&\leqslant 4\epsilon + \sum_{i=1}^m |f(x_i)| |\mathbf{Pr}[X_n \in [a_i,b_i]] - \mathbf{Pr}[X \in [a_i,b_i]]|
\end{aligned}$$

Then taking limits in n and noting ϵ arbitrary we have the result. □

Definition 2 (Convergence in Probability)**.** *We say X_n converges in probability to X (written $X_n \xrightarrow{p} X$) if for all $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbf{Pr}[\|X_n - X\| > \epsilon] = 0$$

Note that we left the norm unspecified for a reason. You can use any norm for the particular metric space you want. But remember that both X_n and X need to be defined on the same probability space.

Definition 3 (Convergence almost surely)**.** *We say X_n converges almost surely to X (written $X_n \xrightarrow{a.s.} X$) if for all $\epsilon > 0$,*

$$\mathbf{Pr}[\lim_{n\to\infty} \|X_n - X\| > \epsilon] = 0$$

Example 1. *Suppose $X_1, X_2, \ldots$ are iid from a distribution $\mathcal{P}$ with $\mathbf{E}[X_i] = \mu$ and $\mathbf{Cov}(X_i) = \Sigma$. let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$.*

Theorem 1 (Strong Law of Large Numbers)**.**

$$\overline{X}_n \xrightarrow{a.s.} \mu$$

Theorem 2 (Central Limit Theorem)**.**

$$\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$$

We will use (1) and (2) very often in this course. Know them well.
Note that we covered convergence in order of increasing strength:

Proposition 1.

$$X_n \xrightarrow{a.s.} X \implies X_n \xrightarrow{p} X \implies X_n \xrightarrow{d} X$$

2 Useful tools (CMT, Slutsky, etc.)

The rest of lecture is devoted to some useful tools that we will often use.

Theorem 3 (Continuous Mapping Theorem)**.** *If g is continuous on a set B satisfying $\mathbf{Pr}[X \in B] = 1$, then*

$$X_n \xrightarrow{*} X \implies g(X_n) \xrightarrow{*} g(X)$$

where $* \in \{d, p, a.s.\}$

This is super useful! We will apply functions to variables all the time so being able to say things about convergence of the functions of the variables is powerful.

Sketch for $\xrightarrow{d}$. If f is continuous and bounded and g is continuous then $h = f \circ g$ is also continuous and bounded. Then

$$\mathbf{E}[h(X_n)] \to \mathbf{E}[h(X)]$$

by assumption and (1). Thus we have

$$\mathbf{E}[f(g(X_n))] \to \mathbf{E}[f(g(X))]$$

for any f continuous and bounded. Thus, again by (1), we have $g(X_n) \xrightarrow{d} g(X)$. □

Application: Suppose we can show that $T_n \to \theta$ and we have some loss ℓ. Can we say that $\ell(T_n) \to \ell(\theta)$? For instance in linear classifier example, if we know that $\hat{w} \to w$ what can we say about $\ell(\hat{w}) \to \ell(w)$?

The next theorem is useful if we have some convergence to a mean and then other convergence. What can we say when we combine those two results? There are two things colloquially known as Slutsky's. We will give both but only call one real Slutsky's.

Theorem 4 (Slutsky's)**.** *1. If c is a constant, $X_n \xrightarrow{d} c \implies X_n \xrightarrow{p} c$.*

2. *If Y_n is a sequence of random variables and $Z_n = \|X_n - Y_n\|$ satisfies $Z_n \xrightarrow{p} 0$ then $X_n \xrightarrow{d} X \implies Y_n \xrightarrow{d} X$.*

3. *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$ (a constant), then*

$$\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}$$

Sketch. We will approach these in order

1. By definition of weak convergence, for all $\epsilon > 0$, $\lim_{n\to\infty} \mathbf{Pr}[X_n \leqslant c + \epsilon\mathbf{1}] = \mathbf{Pr}[c \leqslant c + \epsilon\mathbf{1}]$, where $\mathbf{1}$ is the unit all-1 vector. Then note that c is deterministic so $\mathbf{Pr}[c \leqslant c + \epsilon\mathbf{1}] = 1$. And similarly: $\lim_{n\to\infty} \mathbf{Pr}[X_n > c - \epsilon\mathbf{1}] = \mathbf{Pr}[c > c - \epsilon\mathbf{1}] = 1$ so

$$\lim_{n\to\infty} \mathbf{Pr}[\|X - c\| < \epsilon] = \lim_{n\to\infty} \mathbf{Pr}[c + \epsilon\mathbf{1} \geqslant X_n \geqslant c - \epsilon\mathbf{1}] = 1$$

   Thus by law of total probability, $\lim_{n\to\infty} \mathbf{Pr}[\|X_n - c\| > \epsilon] = 0$.

2. Now suppose f is bounded by 1 and 1 Lipschitz. Thus

$$f(Y_n) = f(X_n) \pm \min\{\|X_n - Y_n\|, 1\}$$

   so using monotonicity of the expectation we have

$$\mathbf{E}[f(Y_n)] = \mathbf{E}[f(X_n)] \pm \mathbf{E}[\min\{\|X_n - Y_n\|, 1\}]$$

   but that second term goes to 0 as $n \to \infty$ so

$$\lim_{n\to\infty} \mathbf{E}[f(Y_n)] = \lim_{n\to\infty} \mathbf{E}[f(X_n)] = \mathbf{E}[f(X)]$$

   so by (1) we have that $Y_n \xrightarrow{d} X$.

3. First note that

$$\begin{pmatrix} X_n \\ c \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}$$

   since for every continuous and bounded $f(x, y)$, the mapping $x \mapsto f(x, c)$ is also continuous and bounded. Similarly,

$$\left\| \begin{pmatrix} X_n \\ Y_n \end{pmatrix} - \begin{pmatrix} X_n \\ c \end{pmatrix} \right\| = \|Y_n - c\| \xrightarrow{p} 0$$

   so by (1) and (2) of this theorem, we get the result.

□

Corollary 1. *If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{p} c$, then*

1. $Y_n X_n \xrightarrow{d} cX$
2. $Y_n + X_n \xrightarrow{d} c + X$
3. *for $c \neq 0$,* $Y_n^{-1} X_n \xrightarrow{d} c^{-1} X$

Proof. Use (4) to conclude $\begin{pmatrix} X_n \\ Y_n \end{pmatrix} \xrightarrow{d} \begin{pmatrix} X \\ c \end{pmatrix}$ then use (3) □

Example 2 (t-type statistic)**.** *Let $X_1, X_2, \ldots \overset{i.i.d.}{\sim} \mathcal{P}$, $\mathbb{E}X_i = \mu$, $\mathbf{Cov}(X_i) = \Sigma > 0$ (i.e. positive definite so we can apply part (iii) of the Corollary). Let $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$, $S_n = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T$, and define $T_n = S_n^{-1/2}\sqrt{n}(\overline{X}_n - \mu)$.*

<u>Claim:</u> By Slutsky's, $T_n \xrightarrow{d} \mathcal{N}(0, I)$. Why is this true? First, by the Strong Law of Large Numbers (Lecture 1, Theorem 2.6), $\overline{X}_n \xrightarrow{\text{a.s.}} \mu$ and $S_n \xrightarrow{\text{a.s.}} \Sigma$ (formally, we can write $S_n = \frac{1}{n}\sum_{i=1}^n (X_i - \overline{X}_n)(X_i - \overline{X}_n)^T = \frac{1}{n}\sum X_i X_i^T - \overline{X}_n \overline{X}_n^T$ and apply LLN to each of the two terms separately before combining). By the Continuous Mapping Theorem (Lecture 1, Theorem 3.1), $S_n^{1/2} \xrightarrow{\text{a.s.}} \Sigma^{1/2}$. Also, by Central Limit Theorem (Lecture 1, Theorem 2.7), $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0, \Sigma)$. Applying part (iii) of the Corollary,

$$T_n = S_n^{-1/2}\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \Sigma^{-1/2}\mathcal{N}(0, \Sigma) \overset{d}{=} \mathcal{N}(0, I)$$

Remark: As an aside, we note that while sample covariance is commonly written with a factor of $1/(n-1)$ to ensure unbiasedness, since we are in an asymptotic regime, it doesn't really matter whether we divide by $n-1$ or n.

3 Uniform Tightness

Definition 4. *A collection $\{X_\alpha\}_{\alpha \in A}$ is **uniformly tight** if for all $\varepsilon > 0$, there exists $M < \infty$ such that*

$$P(\|X_\alpha\| > M) \leqslant \varepsilon \text{ for all } \alpha \in A$$

Example 3 (Markov's inequality)**.** *If all X_α satisfy $\mathbb{E}\|X_\alpha\|^\ell = k < \infty$, $\ell > 0$ then by Markov's inequality*

$$P[\|X_\alpha\| > M] = P[\|X_\alpha\|^\ell > M^\ell] \leqslant \frac{\mathbb{E}\|X_\alpha\|^\ell}{M^\ell} = \frac{k}{M^\ell}$$

Choosing $M > (\frac{k}{\varepsilon})^{1/\ell}$ gives the desired bound.

Example 4 (Non-example: **not** uniformly tight collection)**.** *If $X_n \sim \mathcal{N}(n, 1)$, then $\{X_n\}_{n \in \mathbb{N}}$ is not uniformly tight. In general, if X_n is a sequence with increasing means, it will not be uniformly tight.*

Uniform tightness gives us nice vocabulary and is an important concept that will come up later in the course when we talk about optimality. The following is a theorem which will be used from time to time in this couse. It essentially gives necessary and sufficient conditions for uniform tightness.

Theorem 5 (Prohorov). *A collection $\{X_\alpha\}_{\alpha\in A}$ is uniformly tight if and only if for all sequences $\{X_n\}_{n\in\mathbb{N}} \subset \{X_\alpha\}_{\alpha\in A}$, there is a subsequence n_k and a random variable X such that $X_{n_k} \xrightarrow{d} X$.*

Proof. A sketch of one direction follows from the fact that every weakly converging sequence is uniformly tight. Let $\{X_n\}$ such that $X_n \xrightarrow{d} X$. First, note that any random vector X is tight: for all $\varepsilon > 0$, there exists a constant M such that $P(\|X\| > M) < \varepsilon$. By definition of convergence in distribution (Portmanteau lemma), there exists some N_0 such that for all $n > N_0$, $P[\|X_n\| > M] \leqslant \varepsilon + P[\|X\| > M] < \varepsilon + \varepsilon$. Since there are only finitely many $n \leqslant N_0$, and each of the X_n is tight, we can pick M (by increasing its value if necessary) such that $P[\|X_n\| > M] < 2\varepsilon$ for every n. See van der Vaart for the proof of the other direction (Section 2.1, page 9). □

4 Big-O, little-o notation

We define shorthand notation to help clean things up. First, recall the definition of general (non-stochastic) big-O and little-O: $f(x) = O(g(x))$ if $\lim_{\varepsilon\to 0} \frac{f(\varepsilon)}{g(\varepsilon)} < \infty$ and $f(\varepsilon) = o(g(\varepsilon))$ if $\lim_{\varepsilon\to 0} \frac{f(\varepsilon)}{g(\varepsilon)} = 0$.

Definition 5 (Little-o_p). *Let X_n be vector-valued and R_n real-valued sequences of random variables. We write $X_n = o_p(R_n)$ if there exist vector-valued random variables Y_n such that $X_n = R_n \cdot Y_n$ and $Y_n \xrightarrow{p} 0$; i.e. the magnitude of X_n is bounded in probability by R_n.*

Definition 6 (Big-O_p). *We write $X_n = O_p(R_n)$ if there exist vector-valued Y_n such that $X_n = R_n \cdot Y_n$ and Y_n are uniformly tight (equiv. $Y_n = O_p(1)$); i.e. in probability, X_n takes values proportional to R_n up to a constant.*

Example 5. $\overline{X}_n - \mu = o_p(1)$ *since by LLN,* $\overline{X}_n \xrightarrow{p} \mu$. $\sqrt{n}(\overline{X}_n - \mu) = O_p(1)$ *since by CLT,* $\sqrt{n}(\overline{X}_n - \mu) \xrightarrow{d} \mathcal{N}(0,1)$, *which has finite variance and is thus uniformly tight.*

Remark: The following lemma will come in handy for dealing with remainders (such as in the proof of the Delta Method, which will come up shortly).

Lemma 2. *If $f : \mathbb{R}^d \to \mathbb{R}^k$ satisfying $f(0) = 0$ and $X_n \xrightarrow{p} 0$, then*

(1) If $f(\varepsilon) = o(\|\varepsilon\|^\ell)$ as $\varepsilon \to 0$ for some ℓ, then $f(X_n) = o_p(\|X_n\|^\ell)$

(2) If $f(\varepsilon) = O(\|\varepsilon\|^\ell)$ as $\varepsilon \to 0$ for some ℓ, then $f(X_n) = O_p(\|X_n\|^\ell)$

See Van der Vaart for a proof of the lemma (Section 2.2, page 13).

5 Optimization

5.1 Fenchel-Legendre Biconjugate and Bi-Dual

The Fenchel-Legendre transform (or convex conjugate) of a function $f : \mathbb{R}^n \to \mathbb{R} \cup \{\infty\}$ is defined as:

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

The function f^* maps $y \in \mathbb{R}^n$ to the supremum of the affine functions $\langle y, x\rangle - f(x)$ and is always convex, regardless of whether f itself is convex.

The biconjugate (or Fenchel-Legendre biconjugate) of f, denoted f^{**}, is defined as the conjugate of f^*:

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

The biconjugate f^{**} is the largest lower semi-continuous convex function that does not exceed f. By the Fenchel-Moreau theorem, we have:

$$f^{**}(x) = f(x) \quad \text{if and only if } f \text{ is convex and lower semi-continuous.}$$

5.2 Lagrangian Duality Theory

Duality is a fundamental concept in convex optimization that provides a powerful framework for analyzing and solving optimization problems. It involves formulating a *dual problem* associated with a given *primal problem* and studying the relationships between them.

Primal Problem Consider the following convex optimization problem, known as the **primal problem**:

$$\begin{aligned} \underset{x \in R^n}{\text{minimize}} \quad & f_0(x) \\ \text{subject to} \quad & f_i(x) \leqslant 0, \quad i = 1, \ldots, m, \\ & h_j(x) = 0, \quad j = 1, \ldots, p, \end{aligned}$$

where:

- $f_0, f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ are convex functions.
- $h_1, \ldots, h_p : \mathbb{R}^n \to \mathbb{R}$ are affine (linear) functions.

Lagrangian The **Lagrangian** $L(x, \lambda, \nu)$ combines the objective function and the constraints:

$$L(x, \lambda, \nu) = f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{j=1}^{p} \nu_j h_j(x),$$

where:

- $\lambda_i \geqslant 0$ are the Lagrange multipliers for the inequality constraints.
- ν_j are the Lagrange multipliers for the equality constraints.

Dual Function The **dual function** $g(\lambda, \nu)$ is defined as the infimum of the Lagrangian over x:

$$g(\lambda, \nu) = \inf_x L(x, \lambda, \nu).$$

This function provides a lower bound on the optimal value of the primal problem for any $\lambda \geqslant 0$ and ν.

Dual Problem The **dual problem** is formulated as:

$$\underset{\lambda \succcurlyeq 0, \nu}{\text{maximize}} \quad g(\lambda, \nu).$$

The dual problem is always a concave maximization problem, even if the primal problem is not convex.

Weak Duality The **weak duality** theorem states that for any primal feasible x and dual feasible (λ, ν):

$$g(\lambda, \nu) \leqslant f_0(x).$$

This implies that the optimal value of the dual problem $d^\star$ is a lower bound to the optimal value of the primal problem $p^\star$:

$$d^\star \leqslant p^\star.$$

Strong Duality Under certain conditions, such as Slater's condition (which requires that there exists a strictly feasible point x where $f_i(x) < 0$ and $h_j(x) = 0$), **strong duality** holds:

$$d^\star = p^\star.$$

Strong duality allows us to solve the dual problem instead of the primal problem, which can be more tractable.

Karush-Kuhn-Tucker (KKT) Conditions The KKT conditions provide necessary (and under convexity, sufficient) conditions for optimality:

1. **Primal Feasibility**:
$$f_i(x^\star) \leqslant 0, \quad h_j(x^\star) = 0.$$
2. **Dual Feasibility**:
$$\lambda_i^\star \geqslant 0.$$
3. **Complementary Slackness**:
$$\lambda_i^\star f_i(x^\star) = 0.$$
4. **Stationarity**:
$$\nabla f_0(x^\star) + \sum_{i=1}^{m} \lambda_i^\star \nabla f_i(x^\star) + \sum_{j=1}^{p} \nu_j^\star \nabla h_j(x^\star) = 0.$$

5.3 Example 1: Minmum Norm Linear Regression

Consider the following optimization problem (the *primal problem*):

$$\begin{aligned} \min_{x} \quad & x^\top A x \\ \text{subject to} \quad & Bx = y, \end{aligned}$$

where:

- $x \in \mathbb{R}^n$ is the variable vector,
- $A \in \mathbb{R}^{n \times n}$ is a symmetric positive definite matrix,
- $B \in \mathbb{R}^{m \times n}$ is a matrix representing the constraints,
- $y \in \mathbb{R}^m$ is a constant vector.

Formulating the Lagrangian To incorporate the constraints into the objective function, we introduce Lagrange multipliers $\lambda \in \mathbb{R}^m$ and define the Lagrangian $L(x, \lambda)$:

$$L(x, \lambda) = x^\top A x + \lambda^\top (Bx - y).$$

Deriving the Dual Function The dual function $g(\lambda)$ is obtained by minimizing the Lagrangian with respect to x:

$$g(\lambda) = \inf_{x} L(x, \lambda).$$

To find $g(\lambda)$, we set the gradient of $L(x, \lambda)$ with respect to x to zero:

$$\frac{\partial L}{\partial x} = 2Ax + B^\top \lambda = 0.$$

Solving for x:

$$x = -\frac{1}{2} A^{-1} B^\top \lambda.$$

Computing the Dual Function Substitute x back into the Lagrangian:

$$\begin{aligned} g(\lambda) &= L\left(-\frac{1}{2} A^{-1} B^\top \lambda, \lambda\right) \\ &= \left((\frac{1}{2} \lambda^\top B A^{-1}) A (\frac{1}{2} A^{-1} B^\top \lambda)\right) + \lambda^\top \left(B\left(-\frac{1}{2} A^{-1} B^\top \lambda\right) - y\right) \\ &= \frac{1}{4} \lambda^\top B A^{-1} B^\top \lambda - \frac{1}{2} \lambda^\top B A^{-1} B^\top \lambda - \lambda^\top y \\ &= -\frac{1}{4} \lambda^\top B A^{-1} B^\top \lambda - \lambda^\top y. \end{aligned}$$

Formulating the Dual Problem The dual problem is: $max_\lambda \quad g(\lambda)$. Simplify $g(\lambda)$: $g(\lambda) = -\frac{1}{4} \lambda^\top Q \lambda - \lambda^\top y$, where $Q = B A^{-1} B^\top$.

Solving the Dual Problem To find the optimal λ^*, take the gradient of $g(\lambda)$ with respect to λ and set it to zero: $\frac{\partial g}{\partial \lambda} = -\frac{1}{2}Q\lambda - y = 0$. Solve for λ: $\lambda^* = -2Q^{-1}y$.

Recovering the Primal Solution Substitute λ^* back into the expression for x:

$$x^* = -\frac{1}{2}A^{-1}B^\top \lambda^* = -\frac{1}{2}A^{-1}B^\top \left(-2Q^{-1}y\right) = A^{-1}B^\top Q^{-1}y,$$

where $Q = BA^{-1}B^\top$.

6 Taylor expansions

6.1 Real-valued functions

For $f : \mathbb{R}^d \to \mathbb{R}$ differentiable at $x \in \mathbb{R}^d$,

$$f(y) = f(x) + \nabla f(x)^T(y - x) + o(\|y - x\|). \quad \text{(Remainder version)}$$

$$f(y) = f(x) + \nabla f(\tilde{x})^T(y - x). \quad \text{(Mean value version)}$$

If f is twice differentiable,

$$f(y) = f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T\nabla^2 f(x)(y - x) + o(\|y - x\|^2). \quad \text{(Remainder version)}$$

$$f(y) = f(x) + \nabla f(x)^T(y - x) + \frac{1}{2}(y - x)^T\nabla^2 f(\tilde{x})(y - x). \quad \text{(Mean value version)}$$

6.2 Vector-valued functions

Let $f : \mathbb{R}^d \to \mathbb{R}^k$, $f(x) = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_k \end{bmatrix}$. Define $Df(x) = \begin{bmatrix} \nabla f_1^T(x) \\ \nabla f_2^T(x) \\ \vdots \\ \nabla f_k^T(x) \end{bmatrix} \in \mathbb{R}^{k\times d}$ to be the Jacobian of f.

Then,

$$f(y) = f(x) + Df(x)(y - x) + o(\|y - x\|). \quad \text{(Remainder version)}$$

But for the mean value version, we don't necessarily have $\tilde{x}$ such that

$$f(y) = f(x) + Df(\tilde{x})(y - x).$$

Example 2 (Failure of mean value version): Let $f : \mathbb{R} \to \mathbb{R}^k$, $f(x) = \begin{bmatrix} x \\ x^2 \\ \vdots \\ x^k \end{bmatrix}$, then $Df(x) = \begin{bmatrix} 1 \\ 2x \\ \vdots \\ kx^{k-1} \end{bmatrix}$. Take $x = 0$, $y = 1$, then $f(y) - f(x) = 1 = \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$. Yet $Df(\tilde{x}) = \begin{bmatrix} 1 \\ 2\tilde{x} \\ \vdots \\ k\tilde{x}^{k-1} \end{bmatrix} \neq \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix}$.

Example 3 (Quantitative continuity guarantees): Recall the operator norm of A is

$$\|A\|_{\mathsf{op}} = \sup_{\|u\|_2=1} \|Au\|_2,$$

this implied that $\|Ax\|_2 \leqslant \|A\|_{\mathsf{op}}\|x\|_2$. For $f : \mathbb{R}^d \to \mathbb{R}^k$, differentiable, assume that Df is L-Lipschitz, i.e. $\|Df(x) - Df(y)\|_{\mathsf{op}} \leqslant L\|x - y\|_2$. (Roughly, this means that $\|D^2 f(x)\| \leqslant L$.)

Claim 2. We have

$$f(y) = f(x) + Df(x)(y - x) + R(y - x),$$

where R is a remainder matrix (depending on x, y) that satisfy $\|R\|_{\mathsf{op}} \leqslant \frac{L}{2}\|y - x\|$ and $\|R(y - x)\| \leqslant \frac{L}{2}\|y - x\|^2$.

Proof Define $\phi_i(t) = f_i((1 - t)x + ty)$, $\phi_i : [0, 1] \to \mathbb{R}$. Note that $\phi_i(0) = f_i(x)$, $\phi_i(1) = f_i(y)$, and $\phi_i'(t) = (\nabla f_i((1 - t)x + ty))^T(y - x)$. Then

$$Df((1-t)x+ty)(y-x) = \begin{bmatrix} \nabla f_1^T((1-t)x+ty) \\ \nabla f_2^T((1-t)x+ty) \\ \vdots \\ \nabla f_k^T((1-t)x+ty) \end{bmatrix} (y-x) = \begin{bmatrix} \phi_1'(t) \\ \phi_2'(t) \\ \vdots \\ \phi_k'(t) \end{bmatrix}.$$

Since $\phi_i(1) - \phi_i(0) = \int_0^1 \phi_i'(t)\,dt$,

$$f(y)-f(x) = \int_0^1 Df((1-t)x+ty)(y-x)\,dt = \int_0^1 (Df((1-t)x+ty)-Df(x))(y-x)\,dt+Df(x)(y-x).$$

To bound the remainder term,

$$\begin{aligned}
\left\| \int_0^1 (Df((1-t)x+ty) - Df(x))(y-x)\,dt \right\| &\leqslant \int_0^1 \|(Df((1-t)x+ty) - Df(x))(y-x)\|\,dt \\
&\leqslant \int_0^1 \|Df((1-t)x+ty) - Df(x)\|_{\mathsf{op}}\|y-x\|\,dt \\
&\leqslant \int_0^1 L\|t(y-x)\|\|y-x\|\,dt \\
&\leqslant \int_0^1 Lt\|y-x\|^2\,dt \\
&= \frac{L}{2}\|y-x\|^2.
\end{aligned}$$