Consider the following problems:

① Numerical linear Algebra:

x solves $Ax = b$

y solves $(A+\Delta A)y = b$

how big can $\|x-y\|$ be for a small change ΔA?

② Optimization:

$$x = \underset{z \in \mathbb{R}^d}{\operatorname{argmin}} \; f(z), \text{ for some function } f$$

$$y = \underset{z \in \mathbb{R}^d}{\operatorname{argmin}} \; f(z_0) + \nabla f(z_0)^T (z - z_0) + \frac{1}{2} (z-z_0)^T \nabla_2 f(z_0) (z-z_0)$$

how big can $\|x-y\|$ be for a small $\|x - z_0\|$??

③ Statistics:

$$\theta(P) = \underset{\theta \in \Theta}{\operatorname{argmin}} \int \ell(\theta; x)\, dP(x)$$

$$\theta(P_n) = \underset{\theta \in \Theta}{\operatorname{argmin}} \int \ell(\theta; x)\, dP_n(x)$$

how big can $\|\theta(P) - \theta(P_n)\|$ be for a small $\|P_n - P\|$?

All these problems fall under the concept of perturbation analysis or sensitivity analysis.

Variational analysis provides us with tools to derive perturbation bounds.

In real analysis, we go from sequences to limits to continuity to differentiation.

In variational analysis, one considers liminf/limsups more carefully and study functions that are not continuous but semi-continuous, functions that are not differentiable but semi-differentiable.

In a way, variational analysis stems from ideas in convex analysis where smoothness may not be present but regularity exists.

Some Examples of Perturbation bounds:

Example 1: Numerical linear algebra.

$$Ax = b$$

$$(A+\Delta A)y = b + \Delta b \Rightarrow (A+\Delta A)y = Ax + \Delta b$$

$$\Rightarrow A^{-1}(A+\Delta A)y = x + A^{-1}\Delta b$$

$$\Rightarrow (I + A^{-1}\Delta A)y = x + A^{-1}\Delta b$$

$$\Rightarrow x - y = (A^{-1}\Delta A)y - A^{-1}\Delta b$$

$$\Rightarrow \quad x - y = (A^{-1}\Delta A)x + (A^{-1}\Delta A)(y-x) - A^{-1}\Delta b$$

$$\Rightarrow \quad \|x-y\| \leq \|A^{-1}\Delta A\|_{op} \cdot \|x\| + \|A^{-1}\Delta A\|_{op} \cdot \|x-y\| + \|A^{-1}\Delta b\|$$

$$\Rightarrow \quad \frac{\|x-y\|}{\|x\|} \leq \|A^{-1}\Delta A\|_{op} + \|A^{-1}\Delta A\|_{op} \cdot \frac{\|x-y\|}{\|x\|} + \frac{\|A^{-1}\Delta b\|}{\|A^{-1}b\|}$$

$$\Rightarrow \quad \frac{\|x-y\|}{\|x\|} \leq \frac{\|A^{-1}\Delta A\|_{op}}{1 - \|A^{-1}\Delta A\|_{op}} + \frac{\|A^{-1}\Delta b\|}{\|A^{-1}b\|}.$$

In numerical linear algebra, this bound is presented as

$$\frac{\kappa(A)\,\varepsilon}{1-\kappa(A)\,\varepsilon} \quad \text{where} \quad \kappa(A) = \|A^{-1}\|_{op} \cdot \|A\|_{op}$$

$$\varepsilon = \frac{\|\Delta A\|_{op}}{\|A\|_{op}} \vee \frac{\|\Delta b\|}{\|b\|}$$

<u>Intuition</u>: If relative change in A is small, relative change in x is small.

<u>Application</u>: Take

$$A = \Sigma = \mathbb{E}[XX^T]$$

$$A + \Delta A = \hat{\Sigma} = \frac{1}{n}\sum_{1}^{n} X_i X_i^T$$

$$b = \Gamma = \mathbb{E}[XY]$$

$$b + \Delta b = \hat{\Gamma} = \frac{1}{n}\sum_{1}^{n} X_i Y_i.$$

Provides model-free bounds for estimation error in linear regression.

Example 2: Optimization (Kantorovich's theorem).

$$x = \operatorname*{argmin}_{z \in \mathbb{R}^d} f(z)$$

$$y = \operatorname*{argmin}_{z \in \mathbb{R}^d} f(z_0) + \nabla f(z_0)^T (z - z_0) + \frac{1}{2}(z - z_0)^T \nabla_2 f(z_0)(z - z_0)$$

Then $\nabla f(x) = 0$

and $\nabla f(z_0) + \nabla_2 f(z_0)(y - z_0) = 0.$

Kantorovich Theorem:

If $\nabla_2 f(z_0)$ is non-singular,

$$\|(\nabla_2 f(z_0))^{-1}(\nabla_2 f(z) - \nabla_2 f(z_0))\|_{op} \leq L \|z - z_0\|_2$$

whenever $\|z - z_0\|_2 \leq \frac{1}{3L}$,

and $\|(\nabla_2 f(z_0))^{-1} \nabla f(z_0)\|_2 \leq \frac{2}{9L}$,

then there exists a unique solution

$x \in B(z_0, \frac{3}{2}\|(\nabla_2 f(z_0))^{-1} \nabla f(z_0)\|_2)$ satisfying

$\nabla f(x) = 0$ and

$$\|x - z_0 - (\nabla_2 f(z_0))^{-1} \nabla f(z_0)\|_2 \leq \frac{9L}{4} \|(\nabla_2 f(z_0))^{-1} \nabla f(z_0)\|_2^2$$

Application: (Asymp. Normality of MLE)

Suppose $X_1, \dots, X_n \overset{iid}{\sim} p_{\theta_0}$.

Set $L_n(\theta) = -\sum_{i=1}^{n} \log p_\theta(X_i)$.

If $\nabla_2 L_n(\theta_0)$ is non-singular
(Empirical Fisher information is non-singular at the truth)

and $\theta \mapsto \nabla_2 L_n(\theta)$ is Lipschitz continuous at θ_0
($\approx \theta \mapsto \log p_\theta(x)$ is thrice diff. with bnd third derivative)

and $\|\nabla L_n(\theta_0)\|$ is small,

then there exists a unique solution to $\nabla L_n(\theta) = 0$ in the nbd of θ_0 such that

$$\left\| \hat{\theta} - \theta_0 - (\nabla_2 L_n(\theta_0))^{-1} \nabla L_n(\theta_0) \right\|_2 \lesssim \left\| (\nabla_2 L_n(\theta_0))^{-1} \nabla L_n(\theta_0) \right\|^2$$

$$\Rightarrow \quad \sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, I^{-1}(\theta_0)).$$

Example 3: (Hjort and Pollard (2011, arXiv: 1107.3806))

Suppose $A(\cdot)$ is a convex function and

$$\alpha \in \underset{\theta \in \Theta}{\operatorname{argmin}}\ A(\theta).$$

Suppose $B(\cdot)$ is any function with minimizer β:

$$\beta = \underset{\theta \in \Theta}{\operatorname{argmin}}\ B(\theta). \subseteq \Theta.$$

Define

$$U(\delta) = \sup_{|\theta-\beta| \leq \delta} |A(\theta) - B(\theta)|$$

and $$\mathbb{C}(\delta) = \inf_{|\theta-\beta| = \delta} B(\theta) - B(\beta).$$

Then for any $\delta > 0$ such that

$$U(\delta) \leq \frac{1}{2}\mathbb{C}(\delta),$$

we have $|\alpha - \beta| \leq \delta$.

The proof only relies on the definition of convexity. The principle is general. Uniform convergence and curvature imply closeness of solutions.

Pf:

Take any point

$\theta = \beta + lu$ for some $l > \delta$ and $u \in S^{d-1}$.

$(\|\theta - \beta\| > \delta)$. We shall show that

$A(\theta) > A(\beta)$ which implies that any minimizer of $A(\cdot)$ cannot be outside $\mathcal{B}(\beta, \delta)$.

Convexity of $A(\cdot)$ implies

$$\left(1 - \frac{\delta}{l}\right) A(\beta) + \frac{\delta}{l} A(\theta) \geq A(\beta + \delta u).$$

$$\Rightarrow \left(\frac{\delta}{l}\right)\left(A(\theta) - A(\beta)\right) \geq A(\beta + \delta u) - A(\beta)$$

$$= B(\beta + \delta u) - B(\beta)$$

$$+ \{A(\beta + \delta u) - B(\beta + \delta u)\}$$

$$- \{A(\beta) - B(\beta)\}$$

$$\geq \mathbb{C}(\delta) - 2U(\delta)$$

$\Rightarrow A(\theta) - A(\beta) > 0$ whenever $U(\delta) < \frac{1}{2}\mathbb{C}(\delta)$.

In all these problems, there are certain restrictions on uniquess of solutions. The results presented cannot account for solution sets.
They also cannot handle constraints very well.

Suppose we have

$$\alpha_n = \underset{\theta \in \Theta_n}{\operatorname{argmin}} \; A(\theta)$$

$$\beta_n = \underset{\theta \in \Theta}{\operatorname{argmin}} \; B(\theta).$$

Suppose Θ_n converges to Θ in some sense.
How close/far are α_n & β_n?
Several interesting problems fall into this category.

1. Grid search optimization
2. High-dimensional Dantzig type estimators.

We can always write these as unconstrained problems by introducing the "optimization indicator" function.

For any set Θ define

$$\delta_{\Theta}(\theta) = \begin{cases} 0 & \text{if } \theta \in \Theta \\ \infty & \text{if } \theta \notin \Theta. \end{cases}$$

Then

$$\alpha_n = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \; A(\theta) + \delta_{\Theta_n}(\theta)$$

$$\beta_n = \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \; A(\theta) + \delta_{\Theta}(\theta).$$

But uniform convergence cannot be expected. Additionally, these functions are non-smooth around the boundary of Θ_n & Θ.

Epi-convergence was introduced as a way to tackle such problems. This is also referred to as Γ-convergence or Mosco convergence.

<u>Some Basics:</u>

A function $f: \mathbb{R}^d \to \bar{\mathbb{R}}$ is lower semicontinuous at x_0 if $\liminf\limits_{x \to x_0} f(x) \geq f(x_0)$.

Equivalently, $\liminf = f(x_0)$.

A function f is upper semicontinuous at x_0 if

$$\limsup_{x \to x_0} f(x) = f(x_0).$$

A sequence of functions $\{f_n\}_{n\geq 1}$ is said to epi-converge to a function f if $\forall\ x_0 \in \mathbb{R}^d$

① $\forall$ sequences $\{x_n\}_{n\geq 1}$, $x_n \to x_0$,

$$\liminf_{n\to\infty} f_n(x_n) \geq f(x_0)$$

② for some sequence $\{y_n\}$, $y_n \to x_0$,

$$\limsup_{n\to\infty} f_n(y_n) \leq f(x_0).$$

(or Eq $=$).

<u>Thm</u>: Epi-convergence implies convergence of argmin. If $x_n \in \operatorname{argmin} f_n$ and x^* is a limit point of $\{x_n\}$, then $x^* \in \operatorname{argmin} f$

<u>Pf</u>: Let $\{y_n\}$ be any sequence such that $f(y_n) \to \inf_x f(x)$. By ①

$$\liminf_{n\to\infty} f_n(x_n) \geq f(x^*)$$

By ②

$$\liminf_{n\to\infty} f_n(x_n) \leq \limsup_{n\to\infty} f_n(y_n) \leq \inf_x f(x)$$

$$\Rightarrow f(x^*) = \inf_x f(x).$$

Study of Consistency versus asy. distributions:

Suppose

$$x \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \; f(\theta)$$

$$y \in \underset{\theta \in \mathbb{R}^d}{\operatorname{argmin}} \; g(\theta)$$

then

$$r(x-y) = u \Rightarrow \quad x = y + \frac{u}{r}$$

Solves

$$u \in \underset{\delta \in \mathbb{R}^d}{\operatorname{argmin}} \; f\left(y + \frac{u}{r}\right) - f(y)$$

Hence, studying local convergence of f implies some understanding of asymptotic dist. of u.