3 Ways Optimization

is

Well-Conditioned

or

Ill-Conditioned

Howard Heaton

Typal Academy

Given input data d and a function f_d, consider

$$\min_x f_d(x).$$

Assume $x_d^\star$ is the unique solution to this problem, *i.e.*

$$x_d^\star = \arg\min_x f_d(x).$$

Three matters of interest:

- How solutions $x_d^\star$ change with input data d
- How the landscape (*e.g.* gradients) change with x
- How ratios of singular values affect matrix behavior

Conditioning of Problem

The problem's relative condition number is

$$\kappa_f(d) = \lim_{\delta \to 0^+} \sup_{\|p\| \leq \delta} \frac{\|x^\star_{d+p} - x^\star_d\|}{\|x^\star_d\|} \bigg/ \frac{\|p\|}{\|d\|}.$$

This can be viewed as the limit of the supremum over all infinitesimal perturbations p. Differences in solutions are divided by the size of the solution itself; in the denominator, perturbations p are considered relative to the norm of input data d. If $x^\star_d$ is differentiable, then

$$\kappa_f(d) = \left\| \frac{\partial x^\star_d}{\partial d} \right\| \cdot \frac{\|d\|}{\|x^\star_d\|}.$$

Conditioning of Landscape (Operator)

The landscape of a function is here described in terms of how it is traversed (*e.g.* gradient descent). If T_d is a map from each point x to a point $T_d(x)$, then the relative condition number for the landscape is

$$\kappa_{f,d}(x) = \lim_{\delta \to 0^+} \sup_{\|p\| \leq \delta} \frac{\|T_d(x+p) - T_d(x)\|}{\|T_d(x)\|} \Bigg/ \frac{\|p\|}{\|x\|}.$$

Note: It would be more proper to call this the condition number for the operator T_d, but "function landscapes" are widely known and discussed (unlike operators).

Special Case – Gradient Descent

If f_d is twice differentiable with Hessian $H_d(x)$ and the operator T_d gives the update in gradient descent, *i.e.*

$$T_d(x) = x - \alpha \nabla f_d(x)$$

for a step size $\alpha > 0$, then

$$\kappa_{f,d}(x) = \|\mathrm{I} - \alpha H_d(x)\| \cdot \frac{\|x\|}{\|x - \alpha \nabla f_d(x)\|},$$

where I is the identity matrix. For $x_d^\star \neq 0$, this implies

$$\kappa_{f,d}(x_d^\star) = \|\mathrm{I} - \alpha H_d(x_d^\star)\|.$$

Conditioning of Matrix

The condition number of a square and invertible matrix

A is defined to be

$$\kappa(A) = \|A\|\|A^{-1}\|.$$

When using the Euclidean norm (*i.e.* $\|\cdot\| = \|\cdot\|_2$),

$$\kappa(A) = \frac{\sigma_{max}(A)}{\sigma_{min}(A)}.$$

If A is singular, we set $\kappa(A) = \infty$.

Example – Linear System

Given a square and nonsingular matrix A and vector d, consider solving the linear system $Ax = d$. This can be formulated as a minimization problem:

$$\min_x \|Ax - d\|^2.$$

Here $x_d^\star = A^{-1}d$, and so

$$\kappa_f(d) = \underbrace{\left\| \frac{\partial x_d^\star}{\partial d} \right\|}_{\|A^{-1}\|} \frac{\|d\|}{\|x_d^\star\|} = \|A^{-1}\| \cdot \frac{\|Ax_d^\star\|}{\|x_d^\star\|} \leq \|A^{-1}\|\|A\|.$$

Thus, $\kappa_f(d) \leq \kappa(A)$.

Consider solving the problem via gradient descent, *i.e.*

$$T_d(x) = x - \alpha A^\top(Ax - d),$$

where $\alpha > 0$ is a step size. Then

$$\begin{aligned}\kappa_{f,d}(x) &= \lim_{\delta\to 0^+} \sup_{\|p\|\leq\delta} \frac{\|p - \alpha A^\top A p\|}{\|p\|} \cdot \frac{\|x\|}{\|x - \alpha A^\top(Ax - d)\|} \\ &= \frac{\|\mathrm{I} - \alpha A^\top A\| \cdot \|x\|}{\|(\mathrm{I} - \alpha A^\top A)x + \alpha A^\top d\|},\end{aligned}$$

which implies

$$\boxed{\kappa_{f,d}(x_d^\star) = \left\|\mathrm{I} - \alpha A^\top A\right\|}$$

and

$$\boxed{\lim_{\|x\|\to\infty} \kappa_{f,d}(x) \leq \kappa\left(\mathrm{I} - \alpha A^\top A\right).}$$

- Both the problem and landscape condition numbers relate to matrix condition numbers.
- The condition number of the problem is bounded by the condition number of the matrix A.
- If A is singular, then $\kappa(A) = \infty$ and problem may not have a unique solution. However, the landscape can still be “well-behaved” in this case, *e.g.* consider

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

How to (Loosely) Classify Conditioning

- Well-Conditioned

  small condition number (*e.g.* 1, 10, 100)

- Ill-Conditioned

  large condition number (*e.g.* 10^5, 10^{20})

Example – Quadratic Function

With scalar data d, consider the problem

$$\min_x \frac{(x_1 - 2^d)^2}{2} + \frac{x_2^2}{2}$$

The solution is $x_d^\star = (2^d, 0)$. Letting f_d denote the objective, the gradient is $\nabla f_d(x) = x - x_d^\star$, and the Hessian H_d is the identity matrix. Consider use of gradient descent with step size equal to one half, *i.e.*

$$T_d(x) = x - \frac{1}{2} \nabla f_d(x) = \frac{1}{2}\left(x + x_d^\star\right).$$

Landscape is Well-Conditioned

$$\kappa_{f,d}(x) = \frac{\|H_d(x)\|}{2} \cdot \frac{\|x\|}{\|x - \nabla f_d(x)/2\|} = \frac{\|x\|}{\|x + x_d^\star\|}.$$

Thus, if $x_1 \geq 0$, then $\kappa_{f,d}(x) \leq 1$. In particular,

$$\lim_{x \to x_d^\star} \kappa_{f,d}(x) = \frac{1}{2} \quad \text{and} \quad \lim_{\|x\| \to \infty} \kappa_{f,d}(x) = 1.$$

Problem is Ill-Conditioned

$$\kappa_f(d) = \left\| \frac{\partial x_d^\star}{\partial d} \right\| \cdot \frac{|d|}{\|x_d^\star\|} = \ln(2) \cdot 2^d \cdot \frac{|d|}{2^d} = \ln(2) \cdot |d|.$$

This implies $\kappa_f(d)$ gets large as d increases, *i.e.*

$$\lim_{d \to \infty} \kappa_f(d) = \infty.$$

Ill-conditioned as $x_d^\star$ moves far with small change in d.

Example – Rosenbrock Function

With scalar data d, consider the problem

$$\min_x \frac{(x_1 - 1)^2}{2} + \frac{d(x_2 - x_1^2)^2}{2}.$$

For each choice of d, the solution is $x_d^\star = (1,1)$. Hence

$$\kappa_f(d) = \left\| \frac{\partial x_d^\star}{\partial d} \right\| \cdot \frac{|d|}{\|x_d^\star\|} = 0 \cdot \frac{d}{\sqrt{2}} = 0,$$

and so the problem is well-conditioned.

Yet, estimating $x_d^\star$ numerically is difficult as d increases...

Here the gradient is

$$\nabla f_d(x) = \begin{bmatrix} (x_1 - 1) + 2dx_1(x_1^2 - x_2) \\ d(x_2 - x_1^2) \end{bmatrix},$$

and the Hessian is

$$H_d(x) = \begin{bmatrix} 1 + 2d(3x_1^2 - x_2) & -2dx_1 \\ -2dx_1 & d \end{bmatrix}.$$

To show ill-conditioning, it suffices to consider a gradient descent step at $z = (-1, 1)$ with $\alpha = 1/2$. Here

$$\nabla f_d(z) = \begin{bmatrix} -2 \\ 0 \end{bmatrix} \quad \text{and} \quad H_d(z) = \begin{bmatrix} 1 + 4d & 2d \\ 2d & d \end{bmatrix}.$$

Consequently,

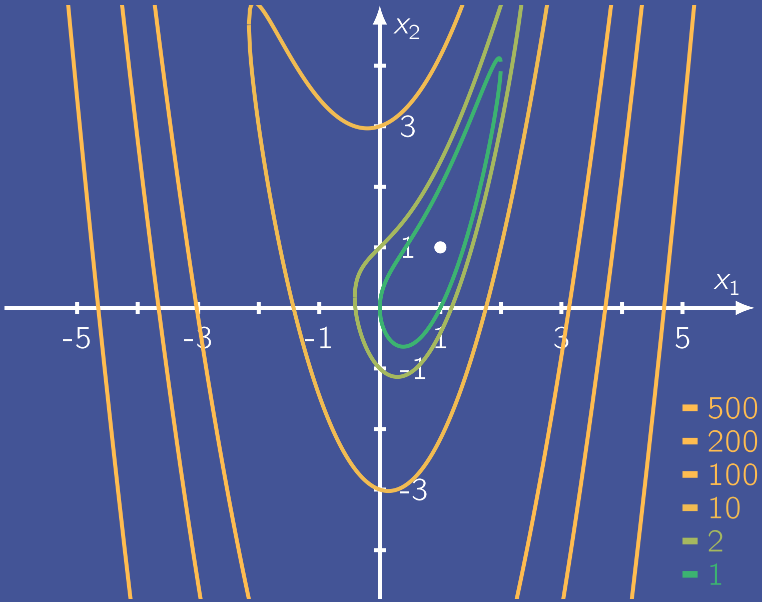
$$\kappa_{f,d}(z) = \frac{\|\mathrm{I} - \alpha H_d(z)\| \cdot \|z\|}{\|z - \alpha \nabla f_d(z)\|} \approx \frac{1 + 5d}{2} \cdot \frac{\sqrt{2}}{1},$$

where the approximation holds when d is large.[1] Thus,

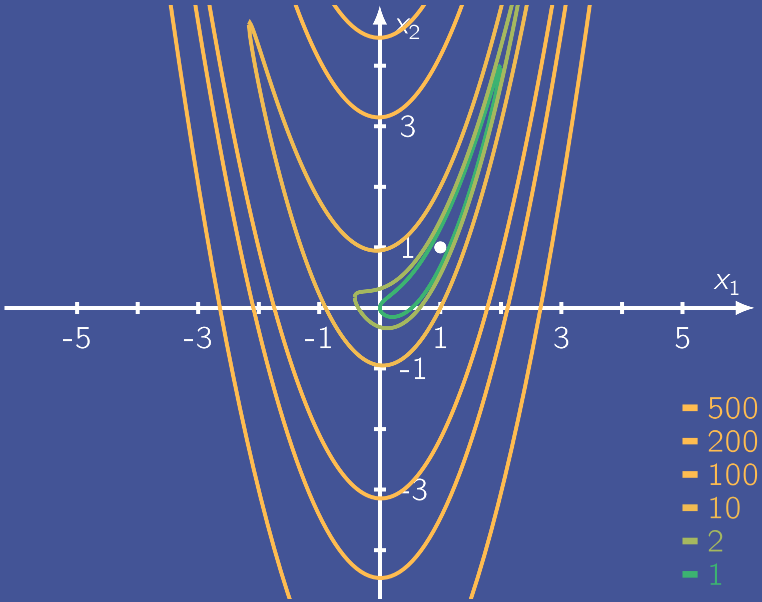
$$\lim_{d \to \infty} \kappa_{f,d}(z) = \infty.$$

Generally, $\kappa_{f,d}(x)$ is large when d is large and $x_2 = x_1^2$, *i.e.* the landscape is ill-conditioned in the "valley" about this curve. The following plots show this "valley" becomes narrower and gets steeper sides as d increases.

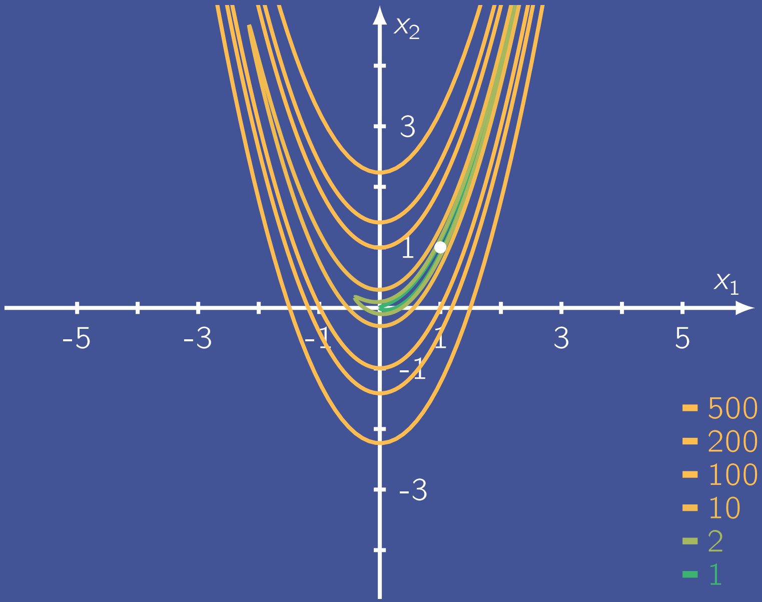
[1] The exact formula for $\|\mathrm{I} - \alpha H_d(z)\|$ is omitted to keep clean presentation.



Rosenbrock function contours for $d = 1$. Dot = $x_d^\star$.



Rosenbrock function contours for $d = 10$. Dot $= x_d^\star$.



Rosenbrock function contours for $d = 100$. Dot $= x_d^\star$.

Howard Heaton

Well-Conditioned Concepts in Optimization

- Problem Condition Number $\kappa_f(d)$

  small changes in d $\rightarrow$ small changes in solution $x_d^\star$

- Landscape Condition Number $\kappa_{f,d}(x)$

  small changes in x $\rightarrow$ small changes in $T_d(x)$

- Matrix Condition Number $\kappa(A)$

  the ratio $\frac{\sigma_{max}(A)}{\sigma_{min}(A)}$ of singular values is small

Found this useful?

➕ Follow me for more

♻️ Repost to share with friends