Introduction of inverse problems

Reference:
[1] A. Kirsch, An Introduction to the Mathematical Theory of Inverse Problems, Applied Mathematical Sciences (1996).
[2] L. Lu, R. Pestourie, W. Yao, Z. Wang, F. Verdugo, and S. G. Johnson, Physics-Informed Neural Networks with Hard Constraints for Inverse Design, SIAM Journal on Scientific Computing 43 (6) (2021) B1105-B1132.
[3] G. Bao, X. Ye, Y. Zang, and H. Zhou, Numerical solution of inverse problems by weak adversarial networks, Inverse Problems 36 (11) (2020) 115003.
[4] Y. Zang and G. Bao, ParticleWNN: a Novel Neural Networks Framework for Solving Partial Differential Equations, ArXiv abs/2305.12433 (2023).

Examples of inverse problems

EX 1 (Backwards heat eqation)
Consider the one-dimensional heat equation

$$\frac{\partial u(x,t)}{\partial t} = \frac{\partial^2 u(x,t)}{\partial x^2}, \quad (x,t) \in (0,\pi) \times \mathbb{R}_{>0}$$

with boundary conditions

$$u(0,t) = u(\pi,t) = 0, \quad t \geq 0$$

and initial condition

$$u(x,0) = u_0(x), \quad 0 \leq x \leq \pi$$

The separation of variables leads to the (formal) solution

$$u(x,t) = \sum_{n=1}^{\infty} a_n e^{-n^2 t} \sin(nx) \quad \text{with} \quad a_n = \frac{2}{\pi} \int_0^{\pi} u_0(y) \sin(ny) dy$$

The direct problem: Given the initial temperature distribution u_0 and the final time T, determine $u(\cdot, T)$.
The inverse problem: one measures the final temperature distribution $u(\cdot, T)$ and tries to determine the temperature at earlier times $t < T$, for example, the initial temperature $u(\cdot, 0)$.
From the solution formula, we see that we have to determine $u_0 := u(\cdot, 0)$ from the following integral equation:

$$u(x,T) = \frac{2}{\pi} \int_0^{\pi} k(x,y) u_0(y) dy, \quad 0 \leq x \leq \pi,$$

where

$$k(x,y) := \sum_{n=1}^{\infty} e^{-n^2 T} \sin(nx) \sin(ny).$$

Ex 2 (Inverse scattering problem)
The direct problem: Let a bounded region $D \subset \mathbb{R}^N (N = 2 \text{ or } 3)$ be given with smooth boundary ∂D (the scattering object) and a plane incident wave $u^i(x) = e^{ik\hat{\theta}\cdot x}$, where $k > 0$ denotes the wave number and $\hat{\theta}$ is a unit vector that describes the direction of the incident wave. The direct problem is to find the total field $u = u^i + u^s$ as the sum of the incident field u^i and the scattered field u^s such that

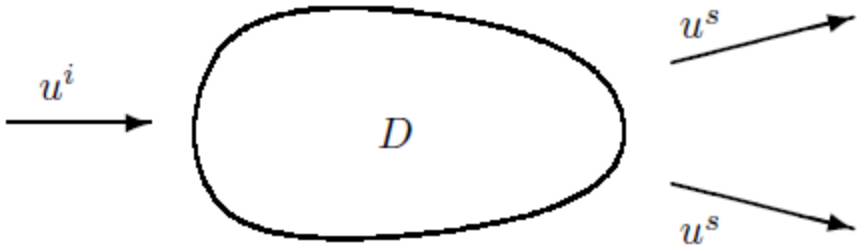
$$\Delta u + k^2 u = 0 \quad \text{in } \mathbb{R}^N \backslash \bar{D}, \quad u = 0 \quad \text{on } \partial D$$
$$\frac{\partial u^s}{\partial r} - iku^s = \mathcal{O}\left(r^{-(N+1)/2}\right) \quad \text{for } r = |x| \to \infty \text{ uniformly in } \frac{x}{|x|}.$$

For acoustic scattering problems, $v(x,t) = u(x)e^{-i\omega t}$ describes the pressure and $k = \omega/c$ is the wave number with speed of sound c. For suitably polarized time harmonic electromagnetic scattering problems, Maxwell's equations reduce to the two-dimensional Helmholtz equation $\Delta u + k^2 u = 0$ for the components of the electric (or magnetic) field u. The wave number k is given in terms of the dielectric constant ε and permeability μ by $k = \sqrt{\varepsilon\mu}\omega$.

In both cases, the radiation condition yields the following asymptotic behavior:

$$u^s(x) = \frac{\exp(ik|x|)}{|x|^{(N-1)/2}} u_\infty(\hat{x}) + \mathcal{O}\left(|x|^{-(N+1)/2}\right) \quad \text{as } |x| \to \infty$$

where $\hat{x} = x/|x|$. **The inverse problem** is to determine the shape of D when the far field pattern $u_\infty(\hat{x})$ is measured for all $\hat{x}$ on the unit sphere in $\mathbb{R}^N$.



Ill-Posed Problems

A mathematical model for a physical problem with properly posed or well-posed in the sense that it has the following three properties:

1. There exists a solution of the problem (existence).
2. There is at most one solution of the problem (uniqueness).

3. The solution depends continuously on the data (stability).

However, most inverse problem is ill-posed.
In backwards heat eqation, the problem is to find a function $u(x,t)$ satisfying the heat equation and the homogeneous lateral boundary conditions

$$\partial_t u - \partial_x^2 u = 0 \text{ in } \Omega \times (0,T), \quad u = 0 \text{ on } \partial\Omega \times (0,T)$$

where Ω is the unit interval $(0,1)$, from the final data

$$u(x,T) = u_T(x), \quad x \in (0,1)$$

The functions $u_k(x,t) = e^{-\pi^2 k^2 t}\sin(\pi k x)$ satisfy the heat equation and the boundary conditions. The initial data are $u_k(x,0) = \sin(\pi k x)$. They have C^0- norm equal to 1 and L_2-norm $(1/2)^{\frac{1}{2}}$. The final data have C^0-norm $e^{-\pi^2 k^2 T}$ and H^m-norm $e^{-\pi^2 k^2 T}\left(\left(1 + \cdots + (\pi k)^{2m}\right)/2\right)^{1/2}$. If we define $Au_0 = u_T$, then the bound $\|u_0\|_X \leq C\|u_T\|_Y$ is impossible when X, Y are classical function spaces: the norms of u_{Tk} go to zero exponentially when the norms of the u_{0k} are greater than $1/2$. Therefore, the problem of finding the initial data from the final data is exponentially unstable in all classical function spaces. This phenomenon is quite typical for many important inverse problems in partial differential equations.

The operator of inverse problem is often unbounded.
Definition (Singular Value) Let $A : H \mapsto \tilde{H}$ be a linear compact operator with adjoint $A^* : \bar{H} \mapsto H, H$ and $\tilde{H}$ be Hilbert spaces. The square root $\sigma_n := \sqrt{\mu_n}$ of the eigenvalue $\mu_n > 0$ of the self-adjoint operator $A^*A : H \mapsto H$ is called the singular value of the operator A.
singular value expansion:

$$Au = \sum_{n=1}^{\infty} \sigma_n \left(u, u_n\right) v_n, u \in H$$

We obtain the following formulae:

$$Au_n = \sigma_n v_n, \quad A^* v_n = \sigma_n u_n$$

The triple $\{\sigma_n, u_n, v_n\}$ is called the singular system for the non-self-adjoint operator $A : H \mapsto \tilde{H}$.

Theorem (Picard) Let H and $\tilde{H}$ be Hilbert spaces and $A : H \mapsto \tilde{H}$ be a linear compact operator with the singular system $\{\sigma, u_n, v_n\}$. Then the equation $Au = f$ has a solution if and only if

$$f \in \mathcal{N}(A^*)^{\perp} \text{ and } \sum_{n=1}^{\infty} \frac{1}{\sigma_n^2} |(f, v_n)|^2 < +\infty$$

In this case

$$u := A^{\dagger} f = \sum_{n=1}^{\infty} \frac{1}{\sigma_n} (f, v_n) u_n$$

is the solution of the equation $Au = f$.

Since $\mu_n > 0, n \in \mathbb{N}$ are eigenvalues of the self-adjoint operator A^*A (as well as AA^*) and $\sigma_n := \sqrt{\mu_n}$, we have: $\sigma_n \to 0$, as $n \to \infty$, if $\dim \mathcal{R}(A) = \infty$. Then it follows from formulae that $A^\dagger$ is an unbounded operator. Indeed, for any fixed eigenvector v_k, with $\|v_k\| = 1$, we have:

$$A^\dagger v_k \quad = \frac{1}{\sigma_k} \to \infty, \text{ as } n \to \infty$$

Regularization

A General Regularization Theory

When analysis the inverse problem, we need to consider the equation $Kx = y$. If we assume the operator K is compact and injective, the inverse problem is to find x when we know y.
We make the assumption that there exists a solution $x^* \in X$ of the unperturbed equation $Kx^* = y^*$. In other words, we assume that $y^* \in \mathcal{R}(K)$. The injectivity of K implies that this solution is unique.

The main problem we focus on:

In practice, the right-hand side $y^* \in Y$ is **never known exactly** but only up to an error of, say, $\delta > 0$. Therefore, we assume that we know $\delta > 0$ and $y^\delta \in Y$ with

$$y^* - y^\delta \ _Y \leq \delta.$$

It is our aim to **"solve" the perturbed equation**

$$Kx^\delta = y^\delta$$

Necessity to regularization:

In general, the perturbed equation is not solvable because we cannot assume that the measured data y^δ are in the range $\mathcal{R}(K)$ of K. Therefore, the best we can hope is to determine an approximation $x^\delta \in X$ to the exact solution x^* that is "not much worse" than the worst-case error $\mathcal{F}\left(\delta, E, \|\cdot\|_{\hat{X}}\right)$.
An additional requirement is that the approximate solution x^δ should depend continuously on the data y^δ. In other words, it is our aim to construct a suitable bounded approximation $R : Y \to X$ of the (unbounded) inverse operator $K^{-1} : \mathcal{R}(K) \to X$.

正则化算子在极限意义下收敛真正的反问题
Definition 2.1 A regularization strategy is a family of linear and bounded operators

$$R_\alpha : Y \longrightarrow X, \quad \alpha > 0$$

such that

$$\lim_{\alpha \to 0} R_\alpha K x = x \quad \text{for all } x \in X$$

that is, the operators $R_\alpha K$ converge pointwise to the identity.
From this definition and the compactness of K, we conclude the following.

正则化算子并非一致有界，因此也不能一致收敛
Theorem 2.2 Let R_α be a regularization strategy for a compact and injective operator $K : X \to Y$ where $\dim X = \infty$. Then we have
(1) The operators R_α are not uniformly bounded; that is, there exists a sequence (α_j) with $\|R_{\alpha_j}\|_{\mathcal{L}(Y,X)} \to \infty$ for $j \to \infty$.
(2) The sequence $(R_\alpha K x)$ does not converge uniformly on bounded subsets of X; that is, there is no convergence $R_\alpha K$ to the identity I in the operator norm.

正则化算子作用时候的误差来源?
The notation of a regularization strategy is based on unperturbed data; that is, the regularizer $R_\alpha y^*$ converges to x^* for the exact right-hand side $y^* = Kx^*$.
Now let $y^* \in \mathcal{R}(K)$ be the exact right-hand side and $y^\delta \in Y$ be the measured data with $\|y^* - y^\delta\|_Y \leq \delta$. We define

$$x^{\alpha,\delta} := R_\alpha y^\delta$$

as an approximation of the solution x^* of $Kx^* = y^*$. Then the error splits into two parts by the following obvious application of the triangle inequality:

$$\begin{aligned} \|x^{\alpha,\delta} - x^*\|_X &\leq \|R_\alpha y^\delta - R_\alpha y^*\|_X + \|R_\alpha y^* - x^*\|_X \\ &\leq \|R_\alpha\|_{\mathcal{L}(Y,X)} \|y^\delta - y^*\|_Y + \|R_\alpha K x^* - x^*\|_X \end{aligned}$$

and thus

$$\|x^{\alpha,\delta} - x^*\|_X \leq \delta \|R_\alpha\|_{\mathcal{L}(Y,X)} + \|R_\alpha K x^* - x^*\|_X$$

Analogously, for the defect in the equation we have

$$\|Kx^{\alpha,\delta} - y^*\|_Y \leq \delta \|K R_\alpha\|_{\mathcal{L}(Y)} + \|K R_\alpha y^* - y^*\|_Y$$

These are our fundamental estimates, which we use often in the following.
误差来源：正则化算子自身条件数+对真解的逼近误差

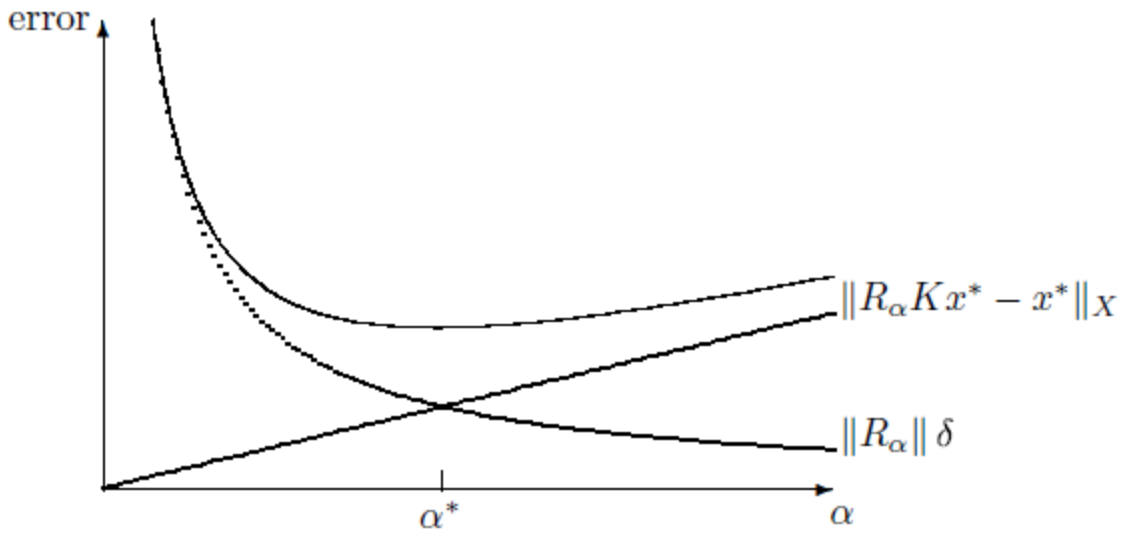


Figure 2.1: Behavior of the total error

Note: 也就是说，α 选择不是越大/小就越好。

如何选择 α?
Definition 2.3 A parameter choice $\alpha = \alpha(\delta)$ for the regularization strategy R_α is called **admissible** if $\lim_{\delta\to 0}\alpha(\delta) = 0$ and

$$\sup\left\{ \left\| R_{\alpha(\delta)} y^\delta - x \right\|_X : y^\delta \in Y,\ \left\| Kx - y^\delta \right\|_Y \le \delta \right\} \to 0, \quad \delta \to 0$$

for every $x \in X$.
原则：随着观测数据越来越准，反演的近似解也要越来越准

如何构造一列正则化算子，满足 admissible 条件？借助奇异值理论
Let $K : X \to Y$ be a linear compact operator, and let $\{\mu_j, x_j, y_j : j \in J\}$ be a singular system for K. As readily seen, **the solution x of $Kx = y$ is given by Picard's theorem** as

$$x = \sum_{j=1}^{\infty} \frac{1}{\mu_j} (y, y_j)_Y x_j$$

provided the series converges, that is, $y \in \mathcal{R}(K)$. This result illustrates again the influence of errors in y because the large factors $1/\mu_j$ (note that $\mu_j \to 0$ as $j \to \infty$) amplify the errors in the expansion coefficients $(y, y_j)_Y$. **We construct regularization strategies by damping the factors $1/\mu_j$.**

通过调控奇异值的系数来构造好的正则化算子
Theorem 2.6 Let $K : X \to Y$ be compact and one-to-one with singular system $\{\mu_j, x_j, y_j : j \in \mathbb{N}\}$ and

$$q : (0, \infty) \times \left(0, \|K\|_{\mathcal{L}(X,Y)}\right] \longrightarrow \mathbb{R}$$

be a function with the following properties:
(1) $|q(\alpha,\mu)| \le 1$ for all $\alpha > 0$ and $0 < \mu \le \|K\|_{\mathcal{L}(X,Y)}$.
(2) For every $\alpha > 0$, there exists $c(\alpha)$ such that

$$|q(\alpha,\mu)| \le c(\alpha)\mu \quad \text{for all} \quad 0 < \mu \le \|K\|_{\mathcal{L}(X,Y)}$$

(3a) $\lim_{\alpha\to 0} q(\alpha,\mu) = 1$ for every $0 < \mu \le \|K\|_{\mathcal{L}(X,Y)}$.
Then the operator $R_\alpha : Y \to X, \alpha > 0$, defined by

$$R_\alpha y := \sum_{j=1}^{\infty} \frac{q(\alpha,\mu_j)}{\mu_j}(y,y_j)_Y x_j, \quad y \in Y$$

is a regularization strategy with $\|R_\alpha\|_{\mathcal{L}(Y,X)} \le c(\alpha)$ and $\|KR_\alpha\|_{\mathcal{L}(Y)} \le 1$. **A choice $\alpha = \alpha(\delta)$ is admissible if $\alpha(\delta) \to 0$ and $\delta c(\alpha(\delta)) \to 0$ as $\delta \to 0$.** The function q is called a regularizing filter for K.

几个例子
Theorem 2.8 The following three functions q satisfy the assumptions (1), (2), (3a), and (3b) of Theorems 2.6 or 2.7 , respectively:

- (a) $q(\alpha,\mu) = \mu^2/\left(\alpha+\mu^2\right)$. This choice satisfies (2) with $c(\alpha) = 1/(2\sqrt{\alpha})$. Assumption (3b) holds with $\omega_\sigma(\alpha) = c_\sigma \alpha^{\sigma/2}$ if $\sigma \le 2$ and $\omega_\sigma(\alpha) \le c_\sigma \alpha$ if $\sigma > 2$. Here c_σ is independent of α. It is $c_1 = 1/2$ and $c_2 = 1$.
- (b) $q(\alpha,\mu) = 1 - \left(1 - a\mu^2\right)^{1/\alpha}$ for some $0 < a < 1/\|K\|^2_{\mathcal{L}(X,Y)}$. In this case (2) holds with $c(\alpha) = \sqrt{a/\alpha}$, and (3b) is satisfied with $\omega_\sigma(\alpha) = \left(\frac{\sigma}{2a}\right)^{\sigma/2}\alpha^{\sigma/2}$ for all $\sigma, \alpha > 0$.
- (c) Let q be defined by

  $$q(\alpha,\mu) = \begin{cases} 1, & \mu^2 \ge \alpha \\ 0, & \mu^2 < \alpha \end{cases}$$

  In this case (2) holds with $c(\alpha) = 1/\sqrt{\alpha}$, and (3b) is satisfied with $\omega_\sigma(\alpha) = \alpha^{\sigma/2}$ for all $\sigma, \alpha > 0$.

Therefore, all of the functions q defined in (a), (b), and (c) are regularizing filters.

Tikhonov Regularization

最佳逼近
Lemma 2.10 Let X and Y be Hilbert spaces, $K : X \to Y$ be linear and bounded, and $y^* \in Y$. There exists $\hat{x} \in X$ with $\|K\hat{x} - y^*\|_Y \le \|Kx - y^*\|_Y$ for all $x \in X$ if and only if $\hat{x} \in X$ solves the normal equation $K^*K\hat{x} = K^*y^*$. Here, $K^* : Y \to X$ denotes the adjoint of K.

Tikhonov 正则化关注的修正能量泛函
Given the linear, bounded operator $K : X \to Y$ and $y \in Y$, determine $x^\alpha \in X$ that minimizes the **Tikhonov functional**

$$J_\alpha(x) := \|Kx - y\|_Y^2 + \alpha\|x\|_X^2 \quad \text{for } x \in X$$

We prove the following theorem.

Theorem 2.11 Let $K : X \to Y$ be a linear and bounded operator between Hilbert spaces and $\alpha > 0$. Then the Tikhonov functional J_α has a unique minimum $x^\alpha \in X$. This minimum x^α is the unique solution of the normal equation

$$\alpha x^\alpha + K^*Kx^\alpha = K^*y$$

The operator $\alpha I + K^*K$ is an isomorphism from X onto itself for every $\alpha > 0$.

为什么这个格式有效？用正则化理论分析该算子
The solution x^α of equation (2.16) can be written in the form $x^\alpha = R_\alpha y$ with

$$R_\alpha := (\alpha I + K^*K)^{-1}K^* : Y \longrightarrow X$$

Choosing a singular system $\{\mu_j, x_j, y_j : j \in \mathbb{N}\}$ for the compact and injective operator K, we see that $R_\alpha y$ has the representation

$$R_\alpha y = \sum_{n=0}^{\infty} \frac{\mu_j}{\alpha + \mu_j^2}(y, y_j)_Y x_j = \sum_{n=0}^{\infty} \frac{q\,(\alpha, \mu_j)}{\mu_j}(y, y_j)_Y x_j, \quad y \in Y$$

with $q(\alpha, \mu) = \mu^2/\left(\alpha + \mu^2\right)$. This function q is exactly the filter function that was studied in Theorem 2.8, part (a). Therefore, applications of Theorems 2.6 and 2.7 yield the following.

Theorem 2.12 Let $K : X \to Y$ be a linear, compact, and injective operator and $\alpha > 0$ and $x^* \in X$ be the exact solution of $Kx^* = y^*$. Furthermore, let $y^\delta \in Y$ with $\left\|y^\delta - y^*\right\|_Y \leq \delta$.
(a) The operators $R_\alpha : Y \to X$ from (2.18) form a regularization strategy with $\|R_\alpha\|_{\mathcal{L}(Y,X)} \leq 1/(2\sqrt{\alpha})$. It is called the Tikhonov regularization method. $R_\alpha y^\delta$ is determined as the unique solution $x^{\alpha,\delta} \in X$ of the equation of the second kind

$$\alpha x^{\alpha,\delta} + K^*Kx^{\alpha,\delta} = K^*y^\delta$$

Every choice $\alpha(\delta) \to 0 (\delta \to 0)$ with $\delta^2/\alpha(\delta) \to 0 (\delta \to 0)$ is admissible.

唯一性?
Theorem 2.13 Let $K : X \to Y$ be linear, compact, and one-to-one such that the range $\mathcal{R}(K)$ is infinite-dimensional. Furthermore, let $x \in X$, and assume that there exists a continuous function $\alpha : [0, \infty) \to [0, \infty)$ with $\alpha(0) = 0$ such that

$$\lim_{\delta \to 0} \left\|x^{\alpha(\delta),\delta} - x\right\|_X \delta^{-2/3} = 0$$

for every $y^\delta \in Y$ with $\left\|y^\delta - Kx\right\|_Y \leq \delta$, where $x^{\alpha(\delta),\delta} \in X$ solves (2.20) for $\alpha = \alpha(\delta)$. Then $x = 0$.

Landweber Iteration

Theorem 2.15 Again let $K : X \to Y$ be a compact and injective operator and let $0 < a < 1/\|K\|^2_{\mathcal{L}(X,Y)}$. Let $x^* \in X$ be the exact solution of $Kx^* = y^*$. Furthermore, let $y^\delta \in Y$ with $\|y^\delta - y^*\|_Y \le \delta$.
(a) Define the linear and bounded operators $R_m : Y \to X$ by (2.24). These operators R_m define a regularization strategy with discrete regularization parameter $\alpha = 1/m, m \in \mathbb{N}$, and $\|R_m\|_{\mathcal{L}(Y,X)} \le \sqrt{am}$. **The sequence $x^{m,\delta} = R_m y^\delta$ is computed by the iteration (2.22), that is,**

$$x^{0,\delta} = 0 \quad \text{and} \quad x^{m,\delta} = (I - aK^*K)x^{m-1,\delta} + aK^*y^\delta$$

for $m = 1, 2, \ldots$. Every strategy $m(\delta) \to \infty (\delta \to 0)$ with $\delta^2 m(\delta) \to 0 (\delta \to 0)$ is admissible.
利用观测数据做类似不动点迭代，可以让迭代变量数值上逼近真解

The Discrepancy Principle of Morozov

Tikhonov正则化中正则项要选取多大?
The Tikhonov regularization of this equation was investigated in Section 2.2. It corresponds to the regularization operators

$$R_\alpha = (\alpha I + K^*K)^{-1}K^* \quad \text{for } \alpha > 0$$

that approximate the unbounded inverse of K on $\mathcal{R}(K)$. We have seen that $x^\alpha = R_\alpha y$ exists and is the unique minimum of the Tikhonov functional

$$J_\alpha(x) := \|Kx - y\|_Y^2 + \alpha\|x\|_X^2, \quad x \in X, \quad \alpha > 0$$

More facts about the dependence on α and y are proven in the following theorem.

Theorem 2.16 Let $y \in Y, \alpha > 0$, and x^α be the unique solution of the equation

$$\alpha x^\alpha + K^*Kx^\alpha = K^*y$$

Then x^α depends continuously on y and α. The mapping $\alpha \mapsto \|x^\alpha\|_X$ is monotonously nonincreasing and

$$\lim_{\alpha \to \infty} x^\alpha = 0$$

The mapping $\alpha \mapsto \|Kx^\alpha - y\|_Y$ is monotonically nondecreasing and

$$\lim_{\alpha \to 0} Kx^\alpha = y$$

If $y \neq 0$, then strict monotonicity holds in both cases.
α 所决定的近似解随着 α 的变化依范数单调。

Now we consider the determination of $\alpha(\delta)$ from the discrepancy principle. We compute $\alpha = \alpha(\delta) > 0$ such that the corresponding Tikhonov solution $x^{\alpha,\delta}$, that is, the solution of the equation

$$\alpha x^{\alpha,\delta} + K^*Kx^{\alpha,\delta} = K^*y^\delta,$$

that is, the minimum of

$$J_{\alpha,\delta}(x) := \quad Kx - y^\delta \quad {}^2_Y + \alpha\|x\|^2_X$$

satisfies the equation

$$Kx^{\alpha,\delta} - y^\delta \quad {}_Y = \delta$$

Note that this choice of α by the discrepancy principle guarantees that, on the one side, the error of the defect is δ and, **on the other side, α is not too small**.
Furthermore, $\alpha \mapsto \quad Kx^{\alpha,\delta} - y^\delta \quad {}_Y$ is continuous and strictly increasing.

如何确定α
Theorem 2.17 Let $K : X \to Y$ be linear, compact, and one-to-one with dense range in Y. Let $Kx^* = y^*$ with $x^* \in X, y^* \in Y$, and $y^\delta \in Y$ such that $\quad y^\delta - y^* \quad {}_Y \le \delta < \quad y^\delta \quad {}_Y$. Let the Tikhonov solution $x^{\alpha(\delta)}$ satisfy $\quad Kx^{\alpha(\delta),\delta} - y^\delta \quad {}_Y = \delta$ for all $\delta \in (0, \delta_0)$. Then
(a) $x^{\alpha(\delta),\delta} \to x^*$ for $\delta \to 0$; that is, **the discrepancy principle is admissible**.
(b) Let $x^* = K^*z \in K^*(Y)$ with $\|z\|_Y \le E$. Then

$$x^{\alpha(\delta),\delta} - x^* \quad {}_X \le 2\sqrt{\delta E}$$

Therefore, the discrepancy principle is an **optimal regularization strategy** under the information $\quad (K^*)^{-1}x^* \quad {}_Y \le E$.

Theorem 2.18 Let K be one-to-one and compact and assume that there exists $\sigma > 0$ such that for every $x \in \mathcal{R}\left((K^*K)^{\sigma/2}\right)$ with $y = Kx \neq 0$, and all sequences $\delta_n \to 0$ and $y^{\delta_n} \in Y$ with $\quad y - y^{\delta_n} \quad {}_Y \le \delta_n$ and $\quad y^{\delta_n} \quad {}_Y > \delta_n$ for all n, the Tikhonov solutions $x^n = x^{\alpha(\delta_n),\delta_n}$ (where $\alpha\,(\delta_n)$ is chosen by the discrepancy principle) converge to x faster than $\sqrt{\delta_n}$ to zero, that is,

$$\frac{1}{\sqrt{\delta_n}}\|x^n - x\|_X \longrightarrow 0 \quad \text{as } n \to \infty$$

Then the range $\mathcal{R}(K)$ has to be finite-dimensional.
依赖于Discrepancy准则构造的近似解列收敛于真解速度快于噪声的平方根。

Landweber's Iteration Method with Stopping Rule

迭代格式什么时候停止?
Let $r > 1$ be a fixed number. Stop the algorithm at the first occurrence of $m \in \mathbb{N}_0$ with $\quad Kx^{m,\delta} - y^\delta \quad {}_Y \le r\delta$. ----- 自然的想法，依赖于误差逼近

Theorem 2.19 Let $K : X \to Y$ be linear, compact, and one-to-one with dense range. Let $Kx^* = y^*$ and $y^\delta \in Y$ be perturbations with $\quad y^* - y^\delta \quad {}_Y \le \delta$ and $\quad y^\delta \quad {}_Y \ge r\delta$ for all $\delta \in (0, \delta_0)$

where $r > 1$ is some fixed parameter (independent of δ). Let the sequence $x^{m,\delta}, m = 0, 1, 2, \ldots$, be determined by Landweber's method; that is, $x^{0,\delta} = 0$ and

$$x^{m+1,\delta} = x^{m,\delta} + aK^* \left(y^\delta - Kx^{m,\delta}\right), \quad m = 0, 1, 2, \ldots$$

for some $0 < a < 1/\|K\|^2_{\mathcal{L}(X,Y)}$. Then the following assertions hold:
(1) $\lim_{m\to\infty} \left\| Kx^{m,\delta} - y^\delta \right\|_Y = 0$ for every $\delta > 0$; that is, the following stopping rule is well-defined: Let $m = m(\delta) \in \mathbb{N}_0$ be the smallest integer with $\left\| Kx^{m,\delta} - y^\delta \right\|_Y \leq r\delta$.
(2) $\delta^2 m(\delta) \to 0$ for $\delta \to 0$, that is, **this choice of $m(\delta)$ is admissible**. Therefore, by the assertions of Theorem 2.15, the sequence $x^{m(\delta),\delta}$ converges to x^* as δ tends to zero.

The Conjugate Gradient Method

考虑优化中的共轭梯度法
Theorem 2.20 (Fletcher-Reeves)
Let $K : X \to Y$ be a compact, linear, and injective operator between Hilbert spaces X and Y. Then the conjugate gradient method is well-defined and either stops or produces sequences $(x^m), (p^m)$ in X with the properties

$$\left(\nabla f\left(x^m\right), \nabla f\left(x^j\right)\right)_X = 0 \quad \text{for all } j \neq m,$$

and

$$\left(Kp^m, Kp^j\right)_Y = 0 \quad \text{for all } j \neq m;$$

that is, **the gradients are orthogonal** and **the directions p^m are K-conjugate**. Furthermore,

$$\left(\nabla f\left(x^j\right), K^*Kp^m\right)_X = 0 \quad \text{for all } j < m.$$

Define again the function

$$f(x) := \|Kx - y\|_Y^2 = (Kx - y, Kx - y)_Y, \quad x \in X.$$

We abbreviate $\nabla f(x) := 2K^*(Kx - y) \in X$ and note that $\nabla f(x)$ is indeed the Riesz representation (see Theorem A.23) of the Fréchet derivative $f'(x)$ of f at x (see Lemma 2.14). We call two elements $p, q \in X K$-conjugate if $(Kp, Kq)_Y = 0$. If K is one-to-one, this bilinear form has the properties of an inner product on X.

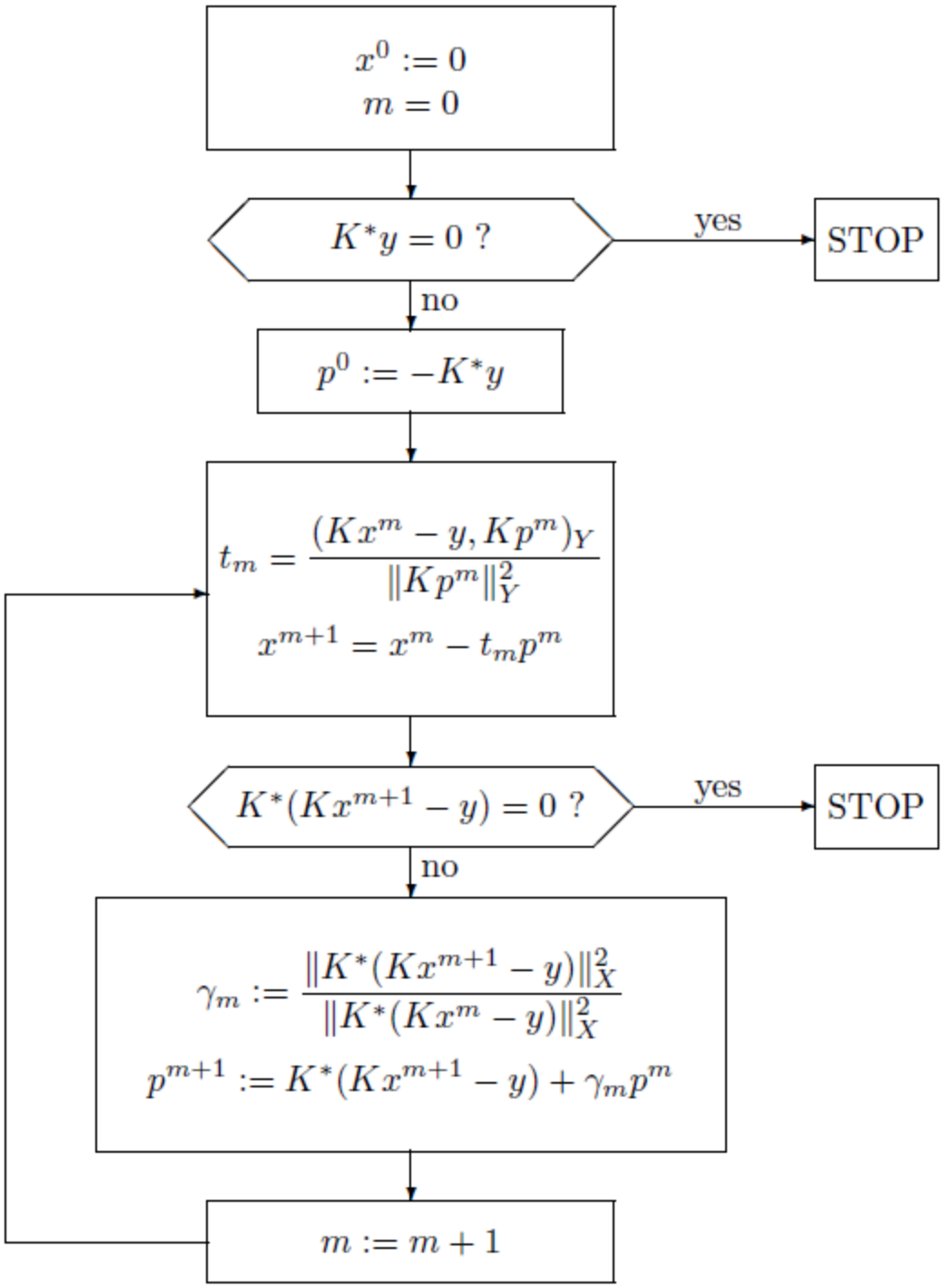


Figure 2.2: The conjugate gradient method

步长张成的子空间于正交梯度张成的子空间相同。相当于每一步都是找到了扩张子空间中的最佳逼近。

Theorem 2.21 Let (x^m) and (p^m) be the sequences of the conjugate gradient method. Define the space $V_m := \operatorname{span}\{p^0, \ldots, p^m\}$. Then we have the following equivalent characterizations of V_m :

$$\begin{aligned} V_m &= \operatorname{span}\{\nabla f(x^0), \ldots, \nabla f(x^m)\} \\ &= \operatorname{span}\{p^0, K^*Kp^0, \ldots, (K^*K)^m p^0\} \end{aligned}$$

for $m = 0, 1, \ldots$. The spaces V_m are called Krylov spaces. Furthermore, x^m is the minimum of f on V_{m-1} for every $m \geq 1$.

共轭梯度法的有效性

Theorem 2.23 Let K and K^* be one-to-one, and assume that the conjugate gradient method does not stop after finitely many steps. Then

$$Kx^m \longrightarrow y \quad \text{as} \quad m \to \infty$$

for every $y \in Y$.

停时选取?

We stop the algorithm with the smallest $m = m(\delta)$ such that the defect $\|Kx^{m,\delta} - y^\delta\|_Y \leq r\delta$, where $r > 1$ is some given parameter. Then our stopping rule is find m s.t.

$$\|Kx^{m(\delta),\delta} - y^\delta\|_Y \leq r\delta < \|Kx^{m(\delta)-1,\delta} - y^\delta\|_Y.$$

Regularization by Discretization

Projection Methods

Definition 3.1 Let X be a normed space over the field $\mathbb{K}$ where $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$. Let $U \subset X$ be a closed subspace. A linear bounded operator $P : X \to X$ is called a projection operator on U if

- $Px \in U$ for all $x \in X$ and
- $Px = x$ for all $x \in U$.

Example

(a) (Orthogonal projection) Let X be a pre-Hilbert space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$ and $U \subset X$ be a complete subspace. Let $Px \in U$ be the best approximation to x in U; that is, Px satisfies

$$\|Px - x\|_X \leq \|u - x\|_X \quad \text{for all } u \in U$$

By the projection theorem, $P : X \to U$ is linear and $Px \in U$ is characterized by the abstract "normal equation" $(x - Px, u)_X = 0$ for all $u \in U$, that is, $x - Px \in U^\perp$. In this example, by the binomial theorem we have

$$\begin{aligned} \|x\|_X^2 &= \|Px + (x - Px)\|_X^2 \\ &= \|Px\|_X^2 + \|x - Px\|_X^2 + \underbrace{2\mathrm{Re}(x - Px, Px)_X}_{=0} \geq \|Px\|_X^2 \end{aligned}$$

that is, $\|P\|_{\mathcal{L}(X)} = 1$. Important examples of subspaces U are spaces of splines or finite elements.

(b) (Interpolation operator) Let $X = C[a, b]$ be the space of real-valued continuous functions on $[a, b]$ supplied with the supremum norm. Then X is a normed space over $\mathbb{R}$. Let $U = \operatorname{span}\{u_1, \ldots, u_n\}$ be an n-dimensional subspace and $t_1, \ldots, t_n \in [a, b]$ such that the

interpolation problem in U is uniquely solvable; that is, $\det\left[u_j\left(t_k\right)\right] \neq 0$. We define $Px \in U$ by the interpolant of $x \in C[a,b]$ in U, i.e., $u = Px \in U$ satisfies $u\left(t_j\right) = x\left(t_j\right)$ for all $j = 1, \ldots, n$. Then $P : X \to U$ is a projection operator.

Let $a = t_1 < \cdots < t_n = b$ be given points, and let $U \subset C[a,b]$ be defined by

$$\begin{aligned} U &= \mathcal{S}_1\left(t_1, \ldots, t_n\right) \\ &:= \left\{x \in C[a,b] : x|_{[t_j, t_{j+1}]} \in \mathcal{P}_1, j = 1, \ldots, n-1\right\} \end{aligned}$$

where $\mathcal{P}_1$ denotes the space of polynomials of degree at most one. Then the interpolation operator $Q_n : C[a,b] \to \mathcal{S}_1\left(t_1, \ldots, t_n\right)$ is given by

$$Q_n x = \sum_{j=1}^{n} x\left(t_j\right) \hat{y}_j \quad \text{for } x \in C[a,b]$$

where the basis functions $\hat{y}_j \in \mathcal{S}_1\left(t_1, \ldots, t_n\right)$ are defined by

$$\hat{y}_j(t) = \begin{cases} \frac{t - t_{j-1}}{t_j - t_{j-1}}, & t \in [t_{j-1}, t_j] \quad (\text{ if } j \geq 2) \\ \frac{t_{j+1} - t}{t_{j+1} - t_j}, & t \in [t_j, t_{j+1}] \quad (\text{ if } j \leq n-1) \\ 0, & t \notin [t_{j-1}, t_{j+1}] \end{cases}$$

for $j = 1, \ldots, n$. In this example $\|Q_n\|_{\mathcal{L}(C[a,b])} = 1$ (see Problem 3.1).
For general interpolation operators, $\|Q_n\|_{\mathcal{L}(X)}$ exceeds one and $\|Q_n\|_{\mathcal{L}(X)}$ does not have to be bounded with respect to n.

Projections methods that come from orthogonal projection and interpolation operator.
Example
Let $K : X \to Y$ be bounded and one-to-one.
(a) (Galerkin method) Let X and Y be pre-Hilbert spaces and $X_n \subset X$ and $Y_n \subset Y$ be finite-dimensional subspaces with $\dim X_n = \dim Y_n = n$. Let $Q_n : Y \to Y_n$ be **the orthogonal projection**. Then the projected equation $Q_n K x_n = Q_n y$ is equivalent to

$$\left(Kx_n, z_n\right)_Y = \left(y, z_n\right)_Y \quad \text{for all } z_n \in Y_n$$

Again let $X_n = \operatorname{span}\left\{\hat{x}_1, \ldots, \hat{x}_n\right\}$ and $Y_n = \operatorname{span}\left\{\hat{y}_1, \ldots, \hat{y}_n\right\}$. Looking for a solution of (3.9a) in the form $x_n = \sum_{j=1}^{n} \alpha_j \hat{x}_j$ leads to the system

$$\sum_{j=1}^{n} \alpha_j \left(K\hat{x}_j, \hat{y}_i\right)_Y = \left(y, \hat{y}_i\right)_Y \quad \text{for } i = 1, \ldots, n$$

or $A\alpha = \beta$, where $A_{ij} := \left(K\hat{x}_j, \hat{y}_i\right)_Y$ and $\beta_i = \left(y, \hat{y}_i\right)_Y$. This corresponds to (3.7) with $\hat{y}_i^* = \hat{y}_i$ after the identification of Y^* with Y (Theorem A. 23 of Riesz).
(b) (Collocation method) Let X be a Banach space, $Y = C[a,b]$, and $K : X \to C[a,b]$ be a bounded operator. Let $a = t_1 < \cdots < t_n = b$ be given points (collocation points) and

$Y_n = \mathcal{S}_1(t_1, \ldots, t_n)$ be the corresponding space (3.2) of linear splines with interpolation operator $Q_n y = \sum_{j=1}^{n} y(t_j)\hat{y}_j$. Let $y \in C[a,b]$ and some n-dimensional subspace $X_n \subset X$ be given. Then $Q_n K x_n = Q_n y$ is equivalent to

$$(Kx_n)(t_i) = y(t_i) \text{ for all } i = 1, \ldots, n.$$

对应数值格式如何计算

We are particularly interested in the study of integral equations of the form

$$\int_a^b k(t,s)x(s)ds = y(t), \quad t \in [a,b]$$

in $L^2(a,b)$ or $C[a,b]$ for some continuous or weakly singular function k. (3.9b) and (3.10b) now take the form

$$A\alpha = \beta$$

where $x = \sum_{j=1}^{n} \alpha_j \hat{x}_j$ and

$$A_{ij} = \int_a^b \int_a^b k(t,s)\hat{x}_j(s)\hat{y}_i(t)dsdt$$
$$\beta_i = \int_a^b y(t)\hat{y}_i(t)dt$$

for the Galerkin method, and

$$A_{ij} = \int_a^b k(t_i,s)\hat{x}_j(s)ds$$
$$\beta_i = y(t_i)$$

for the collocation method.

对 Galerkin 方法的分析

Assumption 3.6 Let $K : X \to Y$ be a linear, bounded, and injective operator between Banach spaces, $X_n \subset X$ and $Y_n \subset Y$ be finite-dimensional subspaces of dimension n, and $Q_n : Y \to Y_n$ be a projection operator. We assume that $\bigcup_{n\in\mathbb{N}} X_n$ is dense in X and that $Q_n K|_{X_n} : X_n \to Y_n$ is one-to-one and thus invertible. Let $x \in X$ be the solution of

$$Kx = y$$

By $x_n \in X_n$, we denote the unique solutions of the equations

$$Q_n K x_n = Q_n y$$

for $n \in \mathbb{N}$.

We can represent the solutions $x_n \in X_n$ of (3.15) in the form $x_n = R_n y$, where $R_n : Y \to X_n \subset X$ is defined by

$$R_n := \left(Q_n K|_{X_n}\right)^{-1} Q_n : Y \longrightarrow X_n \subset X$$

The projection method is called convergent if the approximate solutions $x_n \in X_n$ of (3.15) converge to the exact solution $x \in X$ of (3.14) for every $y \in \mathcal{R}(K)$, that is, if

$$R_n K x = \left(Q_n K|_{X_n}\right)^{-1} Q_n K x \longrightarrow x, \quad n \to \infty$$

for every $x \in X$.
We observe that this definition of convergence coincides with Definition 2.1 of a regularization strategy for the equation $Kx = y$ with regularization parameter $\alpha = 1/n$. Therefore, **the projection method converges if and only if R_n is a regularization strategy for the equation $Kx = y$.**

收敛性判定定理
Theorem 3.7 Let Assumption 3.6 be satisfied. The solution $x_n = R_n y \in X_n$ of (3.15) converges to x for every $y = Kx$ if and only if there exists $c > 0$ such that

$$\|R_n K\|_{\mathcal{L}(X)} \leq c \quad \text{for all } n \in \mathbb{N}$$

If (3.18) is satisfied the following error estimate holds:

$$\|x_n - x\|_X \leq (1 + c) \min_{z_n \in X_n} \|z_n - x\|_X$$

with the same constant c as in (3.18).

A perturbation result:
Theorem 3.9 Let Assumption 3.6 be satisfied and let again $R_n = \left(Q_n K|_{X_n}\right)^{-1}$
$Q_n : Y \to X_n \subset X$. Let $x^* \in X$ the solution of the unperturbed equation $Kx^* = y^*$. Furthermore, we assume that the projection method converges; that is by Theorem 3.7, $\|R_n K\|_{\mathcal{L}(X)}$ are uniformly bounded with respect to n. Furthermore, let $y^\delta \in Y$ with $\|y^\delta - y^*\|_Y \leq \delta$ and $x_n^\delta = R_n y^\delta$ the solution of the projected equation $Q_n K x_n^\delta = Q_n y^\delta$. Then

$$\begin{aligned} \|x_n^\delta - x^*\|_X &\leq \|x_n^\delta - R_n y^*\|_X + \|R_n y^* - x^*\|_X \\ &\leq \|R_n\|_{\mathcal{L}(Y,X)} \|y^\delta - y^*\|_Y + \|R_n K x^* - x^*\|_X . \end{aligned}$$

In practice, one solves the discrete systems (3.6) or (3.7) where the coefficients β_i are replaced by perturbed coefficients $\beta_i^\delta \in \mathbb{K}$; that is, one solves

$$\sum_{j=1}^{n} A_{ij} \alpha_j^\delta = \beta_i^\delta, \quad i = 1, \ldots, n; \quad \text{that is,} \quad A\alpha^\delta = \beta^\delta$$

instead of $A\alpha = \beta$ where now

$$\|\beta^\delta - \beta\|^2 = \sum_{i=1}^{n} |\beta_i^\delta - \beta_i|^2 \leq \delta^2$$

Recall, that the elements A_{ij} of the matrix $A \in \mathbb{K}^{n\times n}$ and the exact coefficients β_i of $\beta \in \mathbb{K}^n$ are given by (3.5) or (3.8). We call this the discrete perturbation of the right-hand side. Then $x_n^\delta \in X_n$ is defined by

$$x_n^\delta = \sum_{j=1}^n \alpha_j^\delta \hat{x}_j.$$

Galerkin Methods

In this section, we assume that X and Y are (real or complex) Hilbert spaces; $K : X \to Y$ is linear, bounded, and one-to-one; $X_n \subset X$ and $Y_n \subset Y$ are finite-dimensional subspaces with $\dim X_n = \dim Y_n = n$; and $Q_n : Y \to Y_n$ is the orthogonal projection operator onto Y_n. Then equation $Q_n K x_n = Q_n y$ reduces to the Galerkin equations (see Example 3.5)

$$(Kx_n, z_n)_Y = (y, z_n)_Y \quad \text{for all } z_n \in Y_n$$

If we choose bases $\{\hat{x}_1, \dots, \hat{x}_n\}$ and $\{\hat{y}_1, \dots, \hat{y}_n\}$ of X_n and Y_n, respectively, then this leads to a finite system for the coefficients of $x_n = \sum_{j=1}^n \alpha_j \hat{x}_j$ (compare with (3.9b)):

$$\sum_{i=1}^n A_{ij}\alpha_j = \beta_i, \quad i = 1, \dots, n$$

where

$$A_{ij} = \left(K\hat{x}_j, \hat{y}_i\right)_Y \quad \text{and} \quad \beta_i = \left(y, \hat{y}_i\right)_Y.$$

Theorem 3.10 Let Assumption 3.6 be satisfied and let again $R_n = \left(Q_n K|_{X_n}\right)^{-1} Q_n : Y \to X_n \subset X$ as in (3.16). Let $x^* \in X$ the solution of the unperturbed equation $Kx^* = y^*$. Furthermore, we assume that the Galerkin method converges; that is by Theorem 3.7, $\|R_n K\|_{\mathcal{L}(X)}$ are uniformly bounded with respect to n.
(a) Let $y^\delta \in Y$ with $\|y^\delta - y^*\|_Y \le \delta$ and $x_n^\delta = R_n y^\delta$ the solution of the projected equation $Q_n K x_n^\delta = Q_n y^\delta$. Then

$$\|x_n^\delta - x^*\|_X \le \|R_n\|_{\mathcal{L}(Y,X)}\delta + \|R_n K x^* - x^*\|_X$$

(b) Let $Q_n y^* = \sum_{i=1}^n \beta_i \hat{y}_i$ and $\beta_i^\delta \in \mathbb{K}$ with $\|\beta^\delta - \beta\| = \sqrt{\sum_{i=1}^n |\beta_i^\delta - \beta_i|^2} \le \delta$ and let $\alpha^\delta \in \mathbb{K}^n$ be the solution of $A\alpha^\delta = \beta^\delta$. Then, with $x_n^\delta = \sum_{j=1}^n \alpha_j^\delta \hat{x}_j$,

$$\|x_n^\delta - x^*\|_X \le \frac{a_n}{\lambda_n}\delta + \|R_n K x^* - x^*\|_X$$
$$\|x_n^\delta - x^*\|_X \le \|R_n\|_{\mathcal{L}(Y,X)} b_n \delta + \|R_n K x^* - x^*\|_X$$

where

$$a_n = \max\left\{ \left\| \sum_{j=1}^{n} \rho_j \hat{x}_j \right\|_X : \sum_{j=1}^{n} |\rho_j|^2 = 1 \right\}$$
$$b_n = \max\left\{ \sqrt{\sum_{i=1}^{n} |\rho_i|^2} : \left\| \sum_{i=1}^{n} \rho_i \hat{y}_i \right\|_Y = 1 \right\},$$

and $\lambda_n > 0$ denotes the smallest singular value of the matrix A. We note that if X or Y are Hilbert spaces and $\{\hat{x}_j : j = 1, \ldots, n\}$ or $\{\hat{y}_i : i = 1, \ldots, n\}$, respectively, are orthonormal systems then $a_n = 1$ or $b_n = 1$, respectively.

The Least Squares Method

An obvious method to solve an equation of the kind $Kx = y$ is the following: Given a finite-dimensional subspace $X_n \subset X$, determine $x_n \in X_n$ such that

$$\|Kx_n - y\|_Y \leq \|Kz_n - y\|_Y \quad \text{for all } z_n \in X_n.$$

Existence and uniqueness of $x_n \in X_n$ follow easily because X_n is finite-dimensional and K is one-to-one. The solution $x_n \in X_n$ of this least squares problem is characterized by

$$(Kx_n, Kz_n)_Y = (y, Kz_n)_Y \quad \text{for all } z_n \in X_n$$

We observe that this method is a special case of the Galerkin method when we set
$Y_n := K(X_n)$.
Choosing a basis $\{\hat{x}_j : j = 1, \ldots, n\}$ of X_n leads to the finite system

$$\sum_{j=1}^{n} \alpha_j \big(K\hat{x}_j, K\hat{x}_i\big)_Y = \beta_i = \big(y, K\hat{x}_i\big)_Y \quad \text{for } i = 1, \ldots, n.$$

Again, we study the case where the exact right-hand side y^* is perturbed by an error. For continuous perturbations, let $x_n^\delta \in X_n$ be the solution of

$$\left(Kx_n^\delta, Kz_n\right)_Y = \left(y^\delta, Kz_n\right)_Y \quad \text{for all } z_n \in X_n$$

where $y^\delta \in Y$ is the perturbed right-hand side with $\|y^\delta - y^*\|_Y \leq \delta$.
For the discrete perturbation, we assume that $\beta_i = \big(y^*, K\hat{x}_i\big)_Y, i = 1, \ldots, n$, is replaced by a vector $\beta^\delta \in \mathbb{K}^n$ with $|\beta^\delta - \beta| \leq \delta$, where $|\cdot|$ denotes the Euclidean norm in $\mathbb{K}^n$. This leads to the following finite system of equations for the coefficients of $x_n^\delta = \sum_{j=1}^n \alpha_j^\delta \hat{x}_j$:

$$\sum_{j=1}^{n} \alpha_j^\delta \big(K\hat{x}_j, K\hat{x}_i\big)_Y = \beta_i^\delta \quad \text{for } i = 1, \ldots, n$$

This system is uniquely solvable because the matrix A is positive definite.

The Dual Least Squares Method

We assume in addition to the general assumptions of this section that the range $\mathcal{R}(K)$ of K is dense in Y.
Given any finite-dimensional subspace $Y_n \subset Y$, determine $u_n \in Y_n$ such that

$$(KK^*u_n, z_n)_Y = (y, z_n)_Y \quad \text{for all } z_n \in Y_n$$

where $K^* : Y \to X$ denotes the adjoint of K. Then $x_n := K^*u_n$ is called the **dual least squares solution**. It is a special case of the Galerkin method when we set $X_n := K^*(Y_n)$. Writing equation for $y = Kx$ in the form

$$(K^*u_n, K^*z_n)_X = (x, K^*z_n)_X \quad \text{for all } z_n \in Y_n$$

we observe that the dual least squares method is just the least squares method for the equation $K^*u = x$. This explains the name.
在数据具有扰动时：
We assume again that the exact right-hand side is perturbed. Let $y^\delta \in Y$ with $\|y^\delta - y^*\|_Y \le \delta$. Instead, one determines $x_n^\delta = K^*u_n^\delta \in X_n$ with

$$\left(K^*u_n^\delta, K^*z_n\right)_X = \left(y^\delta, z_n\right)_Y \quad \text{for all } z_n \in Y_n, \tag{3.37}$$

For discrete perturbations, we choose a basis $\{\hat{y}_j : j = 1, \dots, n\}$ of Y_n and assume that the right-hand sides $\beta_i = \left(y^*, \hat{y}_i\right)_Y, i = 1, \dots, n$, of the Galerkin equations are perturbed by a vector $\beta^\delta \in \mathbb{K}^n$ with $|\beta^\delta - \beta| \le \delta$ where $|\cdot|$ denotes the Euclidean norm in $\mathbb{K}^n$. Instead, we determine

$$x_n^\delta = K^*u_n^\delta = \sum_{j=1}^n \alpha_j^\delta K^*\hat{y}_j, \tag{3.38}$$

where $\alpha^\delta \in \mathbb{K}^n$ solves

$$\sum_{j=1}^n \alpha_j^\delta \left(K^*\hat{y}_j, K^*\hat{y}_i\right)_X = \beta_i^\delta \quad \text{for } i = 1, \dots, n.$$

The Bubnov–Galerkin Method for Coercive Operators

In this subsection, we assume that $Y = X$ coincides, and $K : X \to X$ is a linear and bounded operator and $X_n, n \in \mathbb{N}$, are finite-dimensional subspaces. The Galerkin method reduces to the problem of determining $x_n \in X_n$ such that

$$\left(Kx_n, z_n\right)_X = \left(y, z_n\right)_X \quad \text{for all } z_n \in X_n, \tag{3.42}$$

This special case is called the **Bubnov-Galerkin method**. If $y^\delta \in X$ with $\|y^\delta - y^*\|_X \le \delta$ is a perturbed right-hand side, then instead of (3.42) we study the equation

$$\left(Kx_n^\delta, z_n\right)_X = \left(y^\delta, z_n\right)_X \quad \text{for all } z_n \in X_n, \tag{3.43}$$

The other possibility is to choose a basis $\{\hat{x}_j : j = 1, \dots, n\}$ of X_n and assume that the right-hand sides $\beta_i = \left(y^*, \hat{x}_i\right)_X, i = 1, \dots, n$, of the Galerkin equations are perturbed by a vector $\beta^\delta \in \mathbb{K}^n$ with $\|\beta^\delta - \beta\| \leq \delta$. In this case, instead of (3.42), we have to solve

$$\sum_{j=1}^{n} \alpha_j^\delta \left(K\hat{x}_j, \hat{x}_i\right)_X = \beta_i^\delta \quad \text{for } i = 1, \dots, n, \tag{3.44}$$

Definition A. 26 (a) A **Gelfand triple** (or rigged Hilbert space) $V \subset X \subset V'$ consists of a reflexive Banach space V, a separable Hilbert space X, and the anti-dual space V' of V (all over the same field $\mathbb{K} = \mathbb{R}$ or $\mathbb{K} = \mathbb{C}$) such that V is a dense subspace of X, and the imbedding $j : V \hookrightarrow X$ is bounded. Furthermore, the sesquilinear form $\langle \cdot, \cdot \rangle : V' \times V \to \mathbb{K}$ is an extension of the inner product in X; that is, $\langle x, v \rangle = (x, v)_X$ for all $v \in V$ and $x \in X$.
(b) A linear bounded operator $K : V' \to V$ is called **coercive** if there exists $\gamma > 0$ with

$$|\langle x, Kx \rangle| \geq \gamma \|x\|_{V'}^2 \quad \text{for all } x \in V'$$

The operator K satisfies Gårding's inequality if there exists a linear compact operator $C : V' \to V$ such that $K + C$ is coercive; that is,

$$|\langle x, (K + C)x \rangle| \geq \gamma \|x\|_{V'}^2, \quad \text{for all } x \in V'.$$

Theorem 3.13 Let $V \subset X \subset V'$ be a Gelfand triple, and $X_n \subset V$ be finite dimensional subspaces such that $\bigcup_{n \in \mathbb{N}} X_n$ is dense in X. Let $K : V' \to V$ be coercive in the sense of Definition A. 26 with constant $\gamma > 0$. Let $x^* \in X$ be the solution of $Kx^* = y^*$ for some $y^* \in V$. Then we have the following:
(a) There exist unique solutions of the Galerkin equations (3.42)-(3.44). The Bubnov-Galerkin solutions $x_n \in X_n$ of (3.42) converge in V' with

$$\|x^* - x_n\|_{V'} \leq c \min \{\|x^* - z_n\|_{V'} : z_n \in X_n\}$$

for some $c > 0$.
事实上，这个条件还可以进一步弱化。

Collocation Methods

Let X be a Hilbert space over the field $\mathbb{K}, X_n \subset X$ be finite-dimensional subspaces with $\dim X_n = n$, and $a \leq t_1 < \cdots < t_n \leq b$ be the collocation points. Let $K : X \to C[a, b]$ be bounded and one-to-one. Let $Kx^* = y^*$, and assume that the **collocation equations**

$$\left(Kx_n\right)(t_i) = y(t_i), \quad i = 1, \dots, n, \tag{3.72}$$

are uniquely solvable in X_n for every right-hand side. Choosing a basis $\{\hat{x}_j : j = 1, \dots, n\}$ of X_n , we rewrite this as a system $A\alpha = \beta$, where $x_n = \sum_{j=1}^n \alpha_j \hat{x}_j$ and

$$A_{ij} = \left(K\hat{x}_j\right)(t_i), \quad \beta_i = y(t_i).$$

Main result:

Theorem 3.19 Let $Kx^* = y^*$ and let $\left\{t_1^{(n)}, \ldots, t_n^{(n)}\right\} \subset [a,b], n \in \mathbb{N}$, be a sequence of collocation points. Assume that $\bigcup_{n\in\mathbb{N}} X_n$ is dense in X and that the collocation method converges. Let $x_n^\delta = \sum_{j=1}^n \alpha_j^\delta \hat{x}_j \in X_n$, where α^δ solves $A\alpha^\delta = \beta^\delta$. Here, $\beta^\delta \in \mathbb{K}^n$ satisfies $\|\beta - \beta^\delta\| \leq \delta$ where $\beta_i = y^*(t_i)$. Then the following error estimate holds:

$$\left\| x_n^\delta - x^* \right\|_X \leq c\left(\frac{a_n}{\lambda_n}\delta + \inf\left\{\|x^* - z_n\|_X : z_n \in X_n\right\}\right)$$

where

$$a_n = \max\left\{ \left\| \sum_{j=1}^n \rho_j \hat{x}_j \right\|_X : \sum_{j=1}^n |\rho_j|^2 = 1 \right\}$$

and λ_n denotes the smallest singular value of A.

Machine learning method

PINNs(Strong form)

Inverse design. We consider a physical system governed by partial differential equations (PDEs) defined on a domain $\Omega \subset \mathbb{R}^d$:

$$\mathcal{F}[\mathbf{u}(\mathrm{x}); \gamma(\mathrm{x})] = 0, \quad \mathrm{x} = (x_1, x_2, \cdots, x_d) \in \Omega$$

with suitable boundary conditions (BCs):

$$\mathcal{B}[\mathbf{u}(\mathrm{x})] = 0, \quad \mathrm{x} \in \partial\Omega.$$

What do we focus?
In an inverse-design problem, we search for the best γ by minimizing an objective function $\mathcal{J}$ that depends on $\mathbf{u}$ and γ.The inverse design problem is formulated as a constrained optimization problem:

$$\min_{\mathbf{u},\gamma} \mathcal{J}(\mathbf{u}; \gamma)$$

subject to

$$\begin{cases} \mathcal{F}[\mathbf{u}; \gamma] = \mathbf{0} \\ \mathcal{B}[\mathbf{u}] = 0 \\ h(\mathbf{u}, \gamma) \leq 0 \end{cases}$$

where the last equation is the inequality constraint(s).

Physics-informed neural networks

In a PINN, we employ n fully connected deep neural networks $\hat{\mathbf{u}}\left(\mathbf{x};\boldsymbol{\theta}_u\right)$ to approximate the solution $\mathbf{u}(\mathrm{x})$ (Fig. 1A), where θ_u is the set of trainable parameters in the network. The network takes the coordinates x as the input and outputs the approximate solution $\hat{\mathbf{u}}(\mathrm{x})$. Similarly, we also employ another, independent, fully connected network $\hat{\gamma}\left(\mathrm{x};\boldsymbol{\theta}_\gamma\right)$ for the unknown parameters γ. We then restrict the two networks of $\hat{\mathbf{u}}$ and $\hat{\gamma}$ to satisfy the PDEs by using a PDE-informed loss function:

$$\mathcal{L}_{\mathcal{F}}\left(\boldsymbol{\theta}_u,\boldsymbol{\theta}_\gamma\right)=\frac{1}{MN}\sum_{j=1}^{M}\sum_{i=1}^{N}\mathcal{F}_i\left[\hat{\mathbf{u}}\left(\mathrm{x}_j\right);\hat{\gamma}\left(\mathrm{x}_j\right)\right]^2$$

where $\{\mathrm{x}_1,\mathrm{x}_2,\cdots,\mathrm{x}_M\}$ are a set of M residual points in the domain Ω, and $\mathcal{F}_i\left[\hat{\mathbf{u}}\left(\mathrm{x}_j\right);\hat{\gamma}\left(\mathrm{x}_j\right)\right]$ measures the discrepancy of the i-th $\mathrm{PDE}\,\mathcal{F}_i[\mathbf{u};\gamma]=0$ at the residual point $\mathbf{x}_j$.

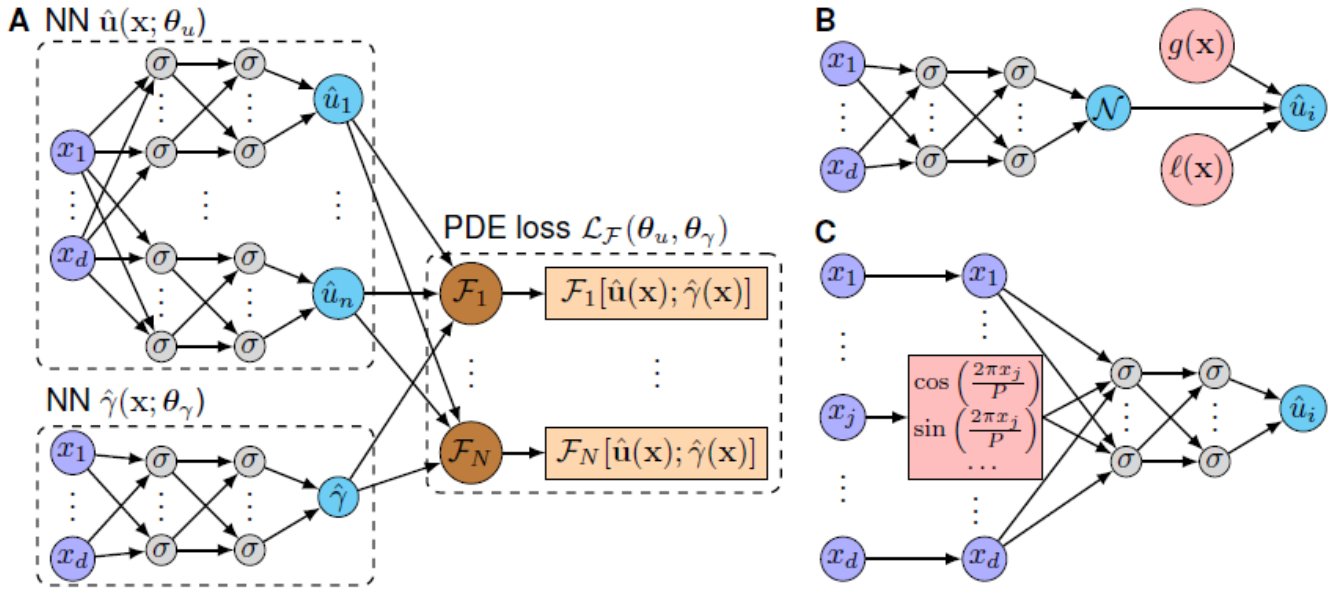


FIG. 1. *Physics-informed neural networks with hard-constraint Dirichlet and periodic boundary conditions. (A) Two independent neural networks* $\hat{\mathbf{u}}(\mathbf{x};\boldsymbol{\theta}_u)$ *and* $\hat{\gamma}(\mathbf{x};\boldsymbol{\theta}_\gamma)$ *are conostructed to approximate* $\mathbf{u}(\mathbf{x})$ *and* $\gamma(\mathbf{x})$*. The gradients in the PDE-informed loss is computed via AD. (B) Dirichlet BCs are strictly imposed into the network architecture by modifying the network output. (C) Periodic BCs are strictly imposed into the network architecture by modifying the network input.*

Hard-constraint boundary conditions

Dirichlet BCs. Let us consider a Dirichlet BC for the solution $u_i(1\leq i\leq n)$:

$$u_i(\mathrm{x})=g(\mathrm{x}),\quad \mathrm{x}\in\Gamma_D$$

where $\Gamma_D\subset\partial\Omega$ is a subset of the boundary. To make the approximate solution $\hat{u}_i\left(\mathrm{x};\boldsymbol{\theta}_u\right)$ satisfy this BC, we construct the solution as (Fig. 1B)

$$\hat{u}_i\left(\mathrm{x};\boldsymbol{\theta}_u\right)=g(\mathrm{x})+\ell(\mathrm{x})\mathcal{N}\left(\mathrm{x};\boldsymbol{\theta}_u\right)$$

where $\mathcal{N}\left(\mathrm{x};\boldsymbol{\theta}_u\right)$ is the network output, and ℓ is a function satisfying the following two conditions:

$$\begin{cases}\ell(\mathrm{x})=0, & \mathrm{x} \in \Gamma_D \\ \ell(\mathrm{x})>0, & \mathrm{x} \in \Omega-\Gamma_D\end{cases}$$

If Γ_D is a simple geometry, then it is possible to choose $\ell(\mathrm{x})$ analytically. For example, when Γ_D is the boundary of an interval $\Omega=[a, b]$, i.e., $\Gamma_D=\{a, b\}$, we can choose $\ell(x)$ as $(x-a)(b-x)$ or $\left(1-e^{a-x}\right)\left(1-e^{x-b}\right)$. For complex domains, it

Periodic BCs. If $u_i(\mathrm{x})$ is a periodic function with respect to x_j of the peribd P, then in the x_j direction, $u_i(\mathrm{x})$ can be decomposed into a weighted summation of the basis functions of the Fourier series $\left\{1, \cos\left(\frac{2\pi x_j}{P}\right), \sin\left(\frac{2\pi x_j}{P}\right), \cos\left(\frac{4\pi x_j}{P}\right), \sin\left(\frac{4\pi x_j}{P}\right), \cdots\right\}$. Hence, we can replace the network input x_j with the Fourier basis functions to impose the periodicity in the x_j direction (Fig. 1C):

$$u_i(\mathrm{x})=\mathcal{N}\left(x_1, \cdots, x_{j-1},\left[\cos\left(\frac{2\pi x_j}{P}\right), \sin\left(\frac{2\pi x_j}{P}\right), \cos\left(\frac{4\pi x_j}{P}\right), \sin\left(\frac{4\pi x_j}{P}\right), \cdots\right], x_{j+1}, \cdots, x_d\right)$$

In classical Fourier analysis, many basis functions may be required approximate an arbitrary periodic function with a good accuracy. We can use as few as two terms $\left\{\cos\left(\frac{2\pi x_j}{P}\right), \sin\left(\frac{2\pi x_j}{P}\right)\right\}$ without loss of accuracy.

Soft constraints

Consider the constraints as soft constraints via loss functions. Specifically, we convert the original constrained optimization to an unconstrained optimization problem:

$$\min_{\boldsymbol{\theta}_u, \boldsymbol{\theta}_\gamma} \mathcal{L}\left(\boldsymbol{\theta}_u, \boldsymbol{\theta}_\gamma\right)=\mathcal{J}+\mu_{\mathcal{F}} \mathcal{L}_{\mathcal{F}}+\mu_h \mathcal{L}_h$$

where $\mathcal{L}_h$ is a quadratic penalty to measure the violation of the hard constraint $h(\mathbf{u}, \gamma) \leq 0$:

$$\mathcal{L}_h\left(\boldsymbol{\theta}_u, \boldsymbol{\theta}_\gamma\right)=\mathbb{1}_{\{h(\hat{\mathbf{u}}, \hat{\gamma})>0\}} h^2(\hat{\mathbf{u}}, \hat{\gamma})$$

and $\mu_{\mathcal{F}}$ and μ_h are the fixed penalty coefficients of the soft constraints. Then the final solution is obtained by minimizing the total loss via gradient-based optimizers:

$$\boldsymbol{\theta}_u^*, \boldsymbol{\theta}_\gamma^*=\arg \min_{\boldsymbol{\theta}_u, \boldsymbol{\theta}_\gamma} \mathcal{L}\left(\boldsymbol{\theta}_u, \boldsymbol{\theta}_\gamma\right)$$

If $\mu_{\mathcal{F}}$ and μ_h are larger, we penalize the constraint violations more severely, thereby forcing the solution satisfying the constraints better. However, when the penalty coefficients are too large, the optimization problem becomes ill-conditioned and hence makes it difficult to converge to a minimum.

Penalty method

The unconstrained problem in the k-th "outer" iteration is

$$\min_{\boldsymbol{\theta}_u, \boldsymbol{\theta}_\gamma} \mathcal{L}^k\left(\boldsymbol{\theta}_u, \boldsymbol{\theta}_\gamma\right)=\mathcal{J}+\mu_{\mathcal{F}}^k \mathcal{L}_{\mathcal{F}}+\mu_h^k \mathcal{L}_h$$

where $\mu_{\mathcal{F}}^k$ and μ_h^k are the penalty coefficients in the k-th iteration. In each iteration, we increase the penalty coefficients by constant factors $\beta_{\mathcal{F}}>1$ and $\beta_h>1$:

$$\mu_{\mathcal{F}}^{k+1}=\beta_{\mathcal{F}} \mu_{\mathcal{F}}^k, \quad \mu_h^{k+1}=\beta_h \mu_h^k .$$

Here, we need to choose the initial coefficients $\mu_{\mathcal{F}}^0$ and μ_h^0 and the factors $\beta_{\mathcal{F}}$ and β_h.

Augmented Lagrangian method

It adds new terms designed to mimic Lagrange multipliers. The unconstrained problem in the k-th iteration is

$$\begin{aligned}\min_{\boldsymbol{\theta}_u, \boldsymbol{\theta}_\gamma} \mathcal{L}^k\left(\boldsymbol{\theta}_u, \boldsymbol{\theta}_\gamma\right)= & \mathcal{J} \\ & +\mu_{\mathcal{F}}^k \mathcal{L}_{\mathcal{F}} \\ & +\mu_h^k \mathbb{1}_{\{h>0 \vee \lambda_h^k>0\}} h^2 \\ & +\frac{1}{M N} \sum_{j=1}^M \sum_{i=1}^N \lambda_{i, j}^k \mathcal{F}_i\left[\hat{\mathbf{u}}\left(\mathbf{x}_j\right) ; \hat{\gamma}\left(\mathbf{x}_j\right)\right] \\ & +\lambda_h^k h,\end{aligned}$$

where the symbol " V " in the third term is the logical OR, and $\lambda_{i, j}^k$ and λ_h^k are multipliers.

Weak form

An inverse problem defined on Ω can be presented in a general form as:

$$\begin{aligned}\mathcal{A}[u, \gamma] & =0, \quad \text { in } \Omega, \\ \mathcal{B}[u, \gamma] & =0, \quad \text { on } \partial \Omega\end{aligned}.$$

Use the weak form of the PDE, then we can deal with the PDE in lower order and lower regularity.
E.g.(IP 1)

$$-\nabla \cdot(\gamma \nabla u)=f, \quad u, \partial_n u|_{\partial \Omega}=0. \Rightarrow(\gamma \nabla u, \nabla v)=(f, v) \quad \text { for test function } v.$$

WAN(weak adversarial networks)

The main idea is to compute the weak form via a **min-max** form.

To obtain the weak formulation of the PDE, we multiply both sides by an arbitrary test function $\varphi \in H_0^1(\Omega)$ (the Hilbert space of functions with bounded first-order weak derivatives and compactly supported in Ω) and integrate over Ω :

$$\langle\mathcal{A}[u, \gamma], \varphi\rangle:=\int_{\Omega} \mathcal{A}[u, \gamma](x) \varphi(x) \mathrm{d} x=0$$

Apply integration by parts to transfer certain gradient operator(s) in $\mathcal{A}[u,\gamma]$ to φ, such that the requirement on the regularity of u (and γ if applicable) can be reduced. For example, in the case of inverse conductivity problem (IP 1), the integration by parts and the fact that $\varphi = 0$ on $\partial\Omega$ together yield

$$\langle \mathcal{A}[u,\gamma], \varphi \rangle = \int_\Omega (\gamma \nabla u \cdot \nabla \varphi - f\varphi) \mathrm{d}x = 0,$$

where $\gamma \nabla u$ is not necessarily differentiable.

We consider the weak formulation of the $\mathrm{PDE}\, \mathcal{A}[u,\gamma] = 0$. To cope with the unknown solution u and parameter γ of the PDE in an IP, we parameterize both u and γ as deep neural networks, and consider $\mathcal{A}[u,\gamma] : H_0^1(\Omega) \to \mathbb{R}$ as a linear functional such that $\mathcal{A}[u,\gamma](\varphi) := \langle \mathcal{A}[u,\gamma], \varphi \rangle$. We define the norm of $\mathcal{A}[u,\gamma]$ induced by the H_1 norm as

$$\|\mathcal{A}[u,\gamma]\|_{\mathrm{op}} := \sup_{\varphi \in H_0^1, \varphi \neq 0} \frac{\langle \mathcal{A}[u,\gamma], \varphi \rangle}{\|\varphi\|_{H^1}}$$

where the H^1-norm of φ is given by $\|\varphi\|_{H^1(\Omega)}^2 = \int_\Omega \left(|\varphi(x)|^2 + |\nabla \varphi(x)|^2 \right) \mathrm{d}x$. Therefore, (u,γ) is a weak solution if and only if $\|\mathcal{A}[u,\gamma]\|_{\mathrm{op}} = 0$ and $\mathcal{B}[u,\gamma] = 0$ on $\partial\Omega$. As $\|\mathcal{A}[u,\gamma]\|_{\mathrm{op}} \geqslant 0$, we know that a weak solution (u,γ) thus solves the following problem:

$$\underset{u,\gamma}{\mathrm{minimize}} \, \|\mathcal{A}[u,\gamma]\|_{\mathrm{op}}^2 = \underset{u,\gamma}{\mathrm{minimize}} \sup_{\varphi \in H_0^1, \varphi \neq 0} \frac{|\langle \mathcal{A}[u,\gamma], \varphi \rangle|^2}{\|\varphi\|_{H^1}^2}$$

among all $(u,\gamma) \in H^1(\Omega) \times L^2(\Omega)$, and attains minimal value 0.

Theorem 1. Suppose that (u^*,γ^*) satisfies the boundary condition $\mathcal{B}\left[u^*,\gamma^*\right] = 0$, then (u^*,γ^*) is a weak solutio if and only if $\left\|\mathcal{A}\left[u^*,\gamma^*\right]\right\|_{\mathrm{op}} = 0$.

Therefore, we can solve (u^*,γ^*) from the following minimization problem which is equivalent:

$$\underset{u,\gamma}{\mathrm{minimize}} \, I(u,\gamma) = \|\mathcal{A}[u,\gamma]\|_{\mathrm{op}}^2 + \beta \|\mathcal{B}[u,\gamma]\|_{L^2(\partial\Omega)}^2.$$

Consider a simple multi-layer neural network u_θ as follows:

$$u_\theta(x) = w_K^\top l_{K-1} \circ \cdots \circ l_0(x) + b_K$$

where the k th layer $l_k : \mathbb{R}^{d_k} \to \mathbb{R}^{d_{k+1}}$ is given by $l_k(z) = \sigma_k\left(W_k z + b_k\right)$ with weight $W_k \in \mathbb{R}^{d_{k+1} \times d_k}$ and bias $b_k \in \mathbb{R}^{d_{k+1}}$ for $k = 0, 1, \ldots, K-1$, and the network parameters of all layers are collectively denoted by θ as follows,

$$\theta := (w_K, b_K, W_{K-1}, b_{K-1}, \ldots, W_0, b_0).$$

With the parameterized $(u_\theta, \gamma_\theta)$ and φ_η, we define

$$E(\theta,\eta) := \left| \langle \mathcal{A}\left[u_\theta, \gamma_\theta\right], \varphi_\eta \rangle \right|^2$$

Instead of normalizing $E(\theta, \eta)$ by $\|\varphi_\eta\|_{H_1}^2$ as in the original definition, we approximate (up to a constant scaling of) the squared operator norm by the following max-type function of θ :

$$L_{\text{int}}(\theta) := \max_{|\eta|^2 \leqslant 2B} E(\theta, \eta)$$

where $B > 0$ is a prescribed bound to constrain the magnitude of network parameter η. Here $|\eta|^2 = \sum_k \left(\sum_{ij} [W_k]_{ij}^2 + \sum_i [b_k]_i^2 \right)$, and $[M]_{ij} \in \mathbb{R}$ stands for the (i, j) th entry of a matrix M, and $[v]_i \in \mathbb{R}$ the i th component of a vector v. The constraint is introduced so that the integrals are bounded (the actual value of this bound can be arbitrary). Furthermore, we define the loss function associated with the boundary condition by

$$L_{\text{bdry}}(\theta) := \|\mathcal{B}[u_\theta, \gamma_\theta]\|_{L^2(\partial\Omega)}^2 = \int_{\partial\Omega} |\mathcal{B}[u_\theta, \gamma_\theta](x)|^2 \, \mathrm{d}S(x)$$

Finally, we define the total loss function $L(\theta)$, and solve the following minimization problem of its optimal θ^* :

$$\underset{\theta}{\text{minimize}}\, L(\theta), \quad \text{where } L(\theta) := L_{\text{int}}(\theta) + \beta L_{\text{bdry}}(\theta)$$

where we also constrain on the magnitude of the parameter θ such that $|\theta|^2 \leqslant 2B$ for the same B to simplify notation.

Algorithm 1. IP solver by weak adversarial network (IWAN).

Input: the domain Ω and data for the IP (1).
Initialize: $(u_\theta, \gamma_\theta)$, φ_η.
for $j = 1, \ldots, J$: **do**
  Sample $X_r = \{x_r^{(i)} : 1 \leqslant i \leqslant N_r\} \subset \Omega$ and $X_b = \{x_b^{(i)} : 1 \leqslant i \leqslant N_b\} \subset \partial\Omega$.
  $\eta \leftarrow \text{SGD}(-\nabla_\eta E(\theta, \eta), X_r, \eta, \tau_\eta, J_\eta)$.
  $\theta \leftarrow \text{SGD}(\partial_\theta E(\theta, \eta) + \beta \nabla_\theta L_{\text{bdry}}(\theta), (X_r, X_b), \theta, \tau_\theta, 1)$.
end for
Output: $(u_\theta, \gamma_\theta)$.

ParticleWNN(Particle Weak-form based Neural Networks)

The main idea is to compute the weak form in small origin to reduce the integration error.

To illustrate the proposed method, we consider the classical Poisson equation with the Dirichlet boundary condition:

$$\begin{cases} -\Delta u(x) = f(x), & \text{in } \Omega \subset \mathbb{R}^d, \\ u(x) = g(x), & \text{on } \partial\Omega \end{cases}$$

The weak formulation of Poisson's equation involves finding a function in $\{u \in H^1(\Omega) | u|_{\partial\Omega} = g\}$ such that, for all test functions $\varphi \in H_0^1(\Omega)$, the following equation holds:

$$\int_\Omega \nabla u \cdot \nabla \varphi dx = \int_\Omega f \varphi dx.$$

Generally, weak-form DNN-based methods) approximate the function u with a neural network $u_{NN}(x;\theta)$:

$$u_{NN}(x;\theta) = T^{(l+1)} \circ T^l \circ T^{(l-1)} \circ \cdots \circ T^{(1)}(x)$$

Here, the linear mapping $T^{(l+1)} : \mathbb{R}^{\mathcal{N}_l} \to \mathbb{R}$ indicates the output layer, and $T^{(i)}(\cdot) = \sigma\left(\boldsymbol{W}_i \cdot + \boldsymbol{b}_i\right), i = 1, \cdots l$ are nonlinear mappings with weights $\boldsymbol{W}_i \in \mathbb{R}^{\mathcal{N}_1 \times \mathcal{N}_{i-1}}$ and biases $b_i \in \mathbb{R}^{\mathcal{N}_i}$. The network parameters are collected in $\theta = \{\boldsymbol{W}_i, \boldsymbol{b}_i\}_{i=1}^{l+1}$. We denote $\mathcal{R}$ as the weak-form residual:

$$\mathcal{R}\left(u_{NN};\varphi\right) = \int_\Omega \nabla u_{NN} \cdot \nabla \varphi dx - \int_\Omega f \varphi dx$$

Different choices of the test functions vary in different weak-form methods. We can use the CSRBFs defined in $B\left(x^c, R\right)$ have the following form:

$$\varphi(r) = \begin{cases} \varphi_+(r), & r(x) \le 1 \\ 0, & r(x) > 1 \end{cases}$$

where $r(x) = \|x - x^c\|_2 / R$. To improve training efficiency, we use multiple test functions to formulate the loss function. We generate N_p particles $\{\boldsymbol{x}_i^c\}_i^{N_p}$ and the corresponding $\{R_i\}_i^{N_p}$ randomly or with predefined rules in the domain, and then define N_p CSRBFs $\{\varphi_i\}_i^{N_p}$ in each small neighbourhood $B\left(\boldsymbol{x}_i^c, R_i\right)$. Therefore, we obtain the MSE of the weak-form residuals:

$$\mathcal{L}_\mathcal{R} = \frac{1}{N_p} \sum_{i=1}^{N_p} \left|\mathcal{R}\left(u_{NN};\varphi_i\right)\right|^2.$$

For the boundary condition (and/or initial condition), we can treat it as a penalty term:

$$\mathcal{L}_\mathcal{B} = \frac{1}{N_{bd}} \sum_{j=1}^{N_{bd}} \left|\mathcal{B}\left[u_{NN}\left(x_j\right)\right] - g\left(x_j\right)\right|^2$$

where $\{x_j\}_{j=1}^{N_{bd}}$ are sampled points on the $\partial\Omega$. Finally, we formulate our loss function as:

$$\mathcal{L}(\theta) = \lambda_\mathcal{R} \mathcal{L}_\mathcal{R} + \lambda_\mathcal{B} \mathcal{L}_\mathcal{B}$$

where $\lambda_\mathcal{R}$ and $\lambda_\mathcal{B}$ are weight coefficients in the loss function.

Typically, we can use the following Wendland's type CSRBFs:

$$\phi_{d,2}(r) = \begin{cases} \frac{(1-r)^{l+2}}{3}\left[\left(l^2+4l+3\right)r^2 + (3l+6)r + 3\right], & r < 1, \\ 0, & r \ge 1, \end{cases}$$

where $l = \lfloor d/2 \rfloor + 3$, d is the dimension of the domain, and $\lfloor \cdot \rfloor$ indicates the flooring function.