ODE-based Learning to Optimize

Zhonglin Xie [1] Wotao Yin [2] Zaiwen Wen [1]

[1]Beijing International Center for Mathematical Research, Peking University [2]Alibaba US, DAMO Academy

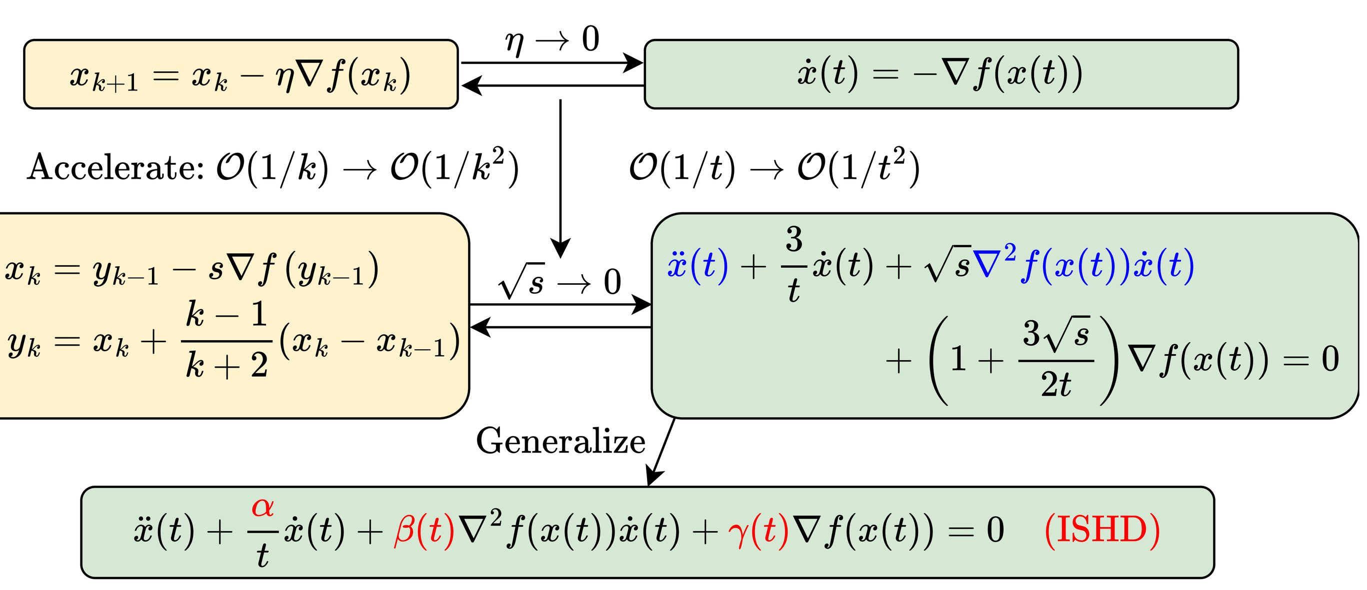
Two important problems

1. Translate the convergence property of ODEs to algorithms:

Combine error analysis in ODE and complexity analysis in OPT

2. Select the best coefficients for (ISHD):

A learning to optimize framework with theoretical guarantee



An enhanced convergence condition for (ISHD)

Given $\kappa \in (0,1], \lambda \in (0, \alpha-1]$, we define

$$\delta(t) = t^2(\gamma(t) - \kappa\dot{\beta}(t) - \kappa\beta(t)/t) + (\kappa(\alpha-1-\lambda) - \lambda(1-\kappa))t\beta(t),$$
$$w(t) = \gamma(t) - \dot{\beta}(t) - \beta(t)/t.$$

Theorem 1: Suppose the following conditions hold true under some mild assumptions:

$$\delta(t) > 0, \quad \text{and} \quad \dot{\delta}(t) \le \lambda t w(t). \qquad \text{(CVG-CDT)}$$

Then, the solution trajectory of (ISHD), $x(t)$, is bounded and the following inequalities can be derived:

$$f(x(t)) - f_\star \le \mathcal{O}\left(\frac{1}{\delta(t)}\right), \quad \int_{t_0}^{\infty} t(\alpha-1-\lambda)\|\dot{x}(t)\|^2 \mathrm{d}t \le \infty,$$
$$\|\nabla f(x(t))\| \le \mathcal{O}\left(\frac{1}{t\beta(t)}\right), \quad \int_{t_0}^{\infty} t^2\beta(t)w(t)\|\nabla f(x)\|^2 \mathrm{d}t \le \infty.$$

Directly applying 4-th Runge-Kutta to (ISHD) diverges

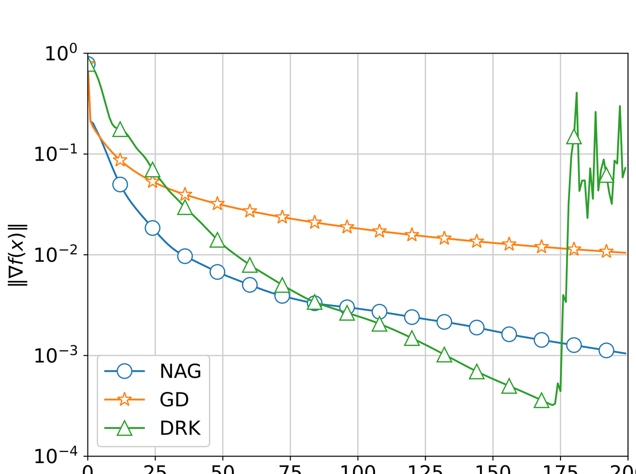
Using $\nabla^2 f(x(t))\dot{x}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\nabla f(x(t))$, (ISHD) equals to

$$\begin{pmatrix} \dot{x}(t) \\ \dot{v}(t) \end{pmatrix} = \underbrace{\begin{pmatrix} v(t) - \beta(t)\nabla f(x(t)) \\ -\frac{\alpha}{t}(v(t) - \beta(t)\nabla f(x(t))) + (\dot{\beta}(t) - \gamma(t))\nabla f(x(t)) \end{pmatrix}}_{\psi_\Xi(x(t), v(t), t), \text{ where } \Xi = (\alpha, \beta(\cdot), \gamma(\cdot))} \qquad \text{(FRT-ODR)}$$

Given $h > 0$, the forward Euler scheme of (FRT-ODR) writes

$$\begin{cases} \frac{x_{k+1} - x_k}{h} = v_k - \beta(t_k)\nabla f(x_k) \\ \frac{v_{k+1} - v_k}{h} = -\frac{\alpha}{t}(v_k - \beta(t_k)\nabla f(x_k)) + (\dot{\beta}(t_k) - \gamma(t_k))\nabla f(x_k) \end{cases} \qquad \text{(EIGAC)}$$

where $t_k = t_0 + kh$, and $v(t_0) = x(t_0) + \beta(t_0)\nabla f(x(t_0))$



- Consider
$$\min_{x \in \mathbb{R}^n} f(x) = \frac{1}{N}\sum_{i=1}^{N} \log(1 + \exp(-b_i\langle a_i, w\rangle))$$
- Set $p = 5, \alpha = 2p+1, \beta(t) \equiv 0$ and $\gamma(t) = p^2 t^{p-2}$ in (ISHD)
- Then, $\kappa = 1, \lambda = \alpha - 1, \delta(t) = p^2 t^p$, and $w(t) = p^2 t^{p-2}$
- (CVG-CDT) holds and $f(x(t)) - f_\star \le \mathcal{O}(1/t^p)$

A stability condition for applying the forward Euler scheme

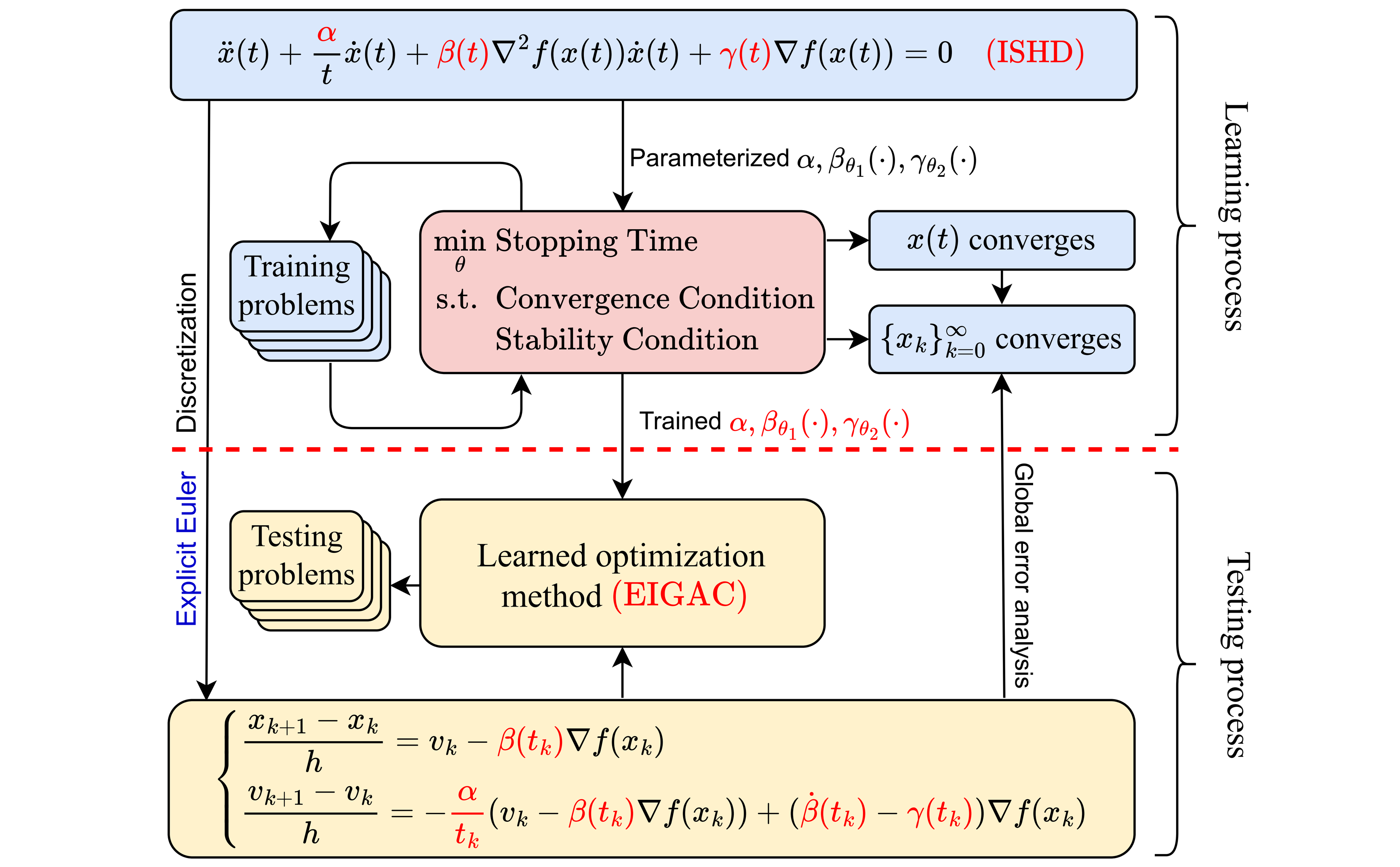
Suppose (CVG-CDT) holds. Given t_0, s_0, h, the sequence $\{x_k\}_{k=0}^{\infty}$ generated by (EIGAC), we denote the continuous time interpolation $\bar{x}(t)$ as

$$\bar{x}(t) = x_k + \frac{x_{k+1} - x_k}{h}(t - t_k), \qquad t \in [t_k, t_{k+1}).$$

Then, it holds $f(x_k) - f_\star \le \mathcal{O}(1/k)$ under the following stability condition:

$$\Lambda(x, f) \ge \|\nabla^2 f(x)\|, \quad \alpha\beta(t)/t \le \gamma(t) - \dot{\beta}(t) \le \beta(t)/h,$$
$$\sqrt{\int_0^1 \Lambda((1-\tau)X(t, \Xi, f) + \tau\bar{x}(t), f)\,\mathrm{d}\tau} \le \frac{\sqrt{\gamma(t) - \dot{\beta}(t)} + \sqrt{\gamma(t) - \dot{\beta}(t) - \frac{\alpha}{t}\beta(t)}}{\beta(t)}. \qquad \text{(STB-CDT)}$$

Selecting the best ODE using a complexity-inspired model



L2O Framework: minimize the expectation of stopping time under conditions of convergence and stable discretization

$$\min_\Xi \quad \mathbb{E}_f[T(\Xi, f)]$$
$$\text{s.t.} \quad \mathbb{E}_f[P(\Xi, f)] \le 0 \quad \mathbb{E}_f[Q(\Xi, f)] \le 0$$

Setting $P, Q \le 0$ ensures (CVG-CDT) and (STB-CDT) hold for f

$$P(\Xi, f) = \int_{t_0}^{T(\Xi,f)} p(X(t, \Xi, f), \bar{x}(t), \Xi, t, f)\,\mathrm{d}t, \quad Q(\Xi, f) = \int_{t_0}^{T(\Xi,f)} q(\Xi, t)\,\mathrm{d}t$$

Induced Probability: Given a random variable $\xi \sim \mathbb{P}$. We say $\mathbb{P}$ is the induced probability of the parameterized function $f(\cdot;\xi)$

$$\mathbb{E}_f[T(\Xi, f)] = \int_\xi T(\Xi, f(\cdot;\xi))\,\mathrm{d}\mathbb{P}(\xi) = \mathbb{E}_\xi[T(\Xi, f(\cdot;\xi))]$$

Stopping Time: $X(\Xi, t, f)$ is the trajectory of (ISHD). Given $\varepsilon > 0$, the stopping time of the criterion $\|\nabla f(x)\| \le \varepsilon$ is

$$T(\Xi, f) = \inf\{t \mid \|\nabla f(X(\Xi, t, f))\| \le \varepsilon, t \ge t_0\}$$

Solve the L2O problem using stochastic penalty method

Parameterization: $\beta \to \beta_{\theta_1}, \gamma \to \gamma_{\theta_2}$. Set $\theta = (\alpha, \theta_1, \theta_2)$

Stochastic Penalty Method (StoPM): Apply SGD to ℓ_1 exact penalty function of the L2O problem

$$\min_\theta \Upsilon_\rho(\theta) = \mathbb{E}_f[T(\theta, f)] + \rho\left(\mathbb{E}_f[P(\theta, f)] + \mathbb{E}_f[Q(\theta, f)]\right)$$
$$= \mathbb{E}_f[T(\theta, f) + \rho(P(\theta, f) + Q(\theta, f))]$$

Gradient of stopping time: Take limit: $\|\nabla f(X(T(\theta, f), f, \theta))\|^2 - \varepsilon^2 \equiv 0$. implicit function theorem gives

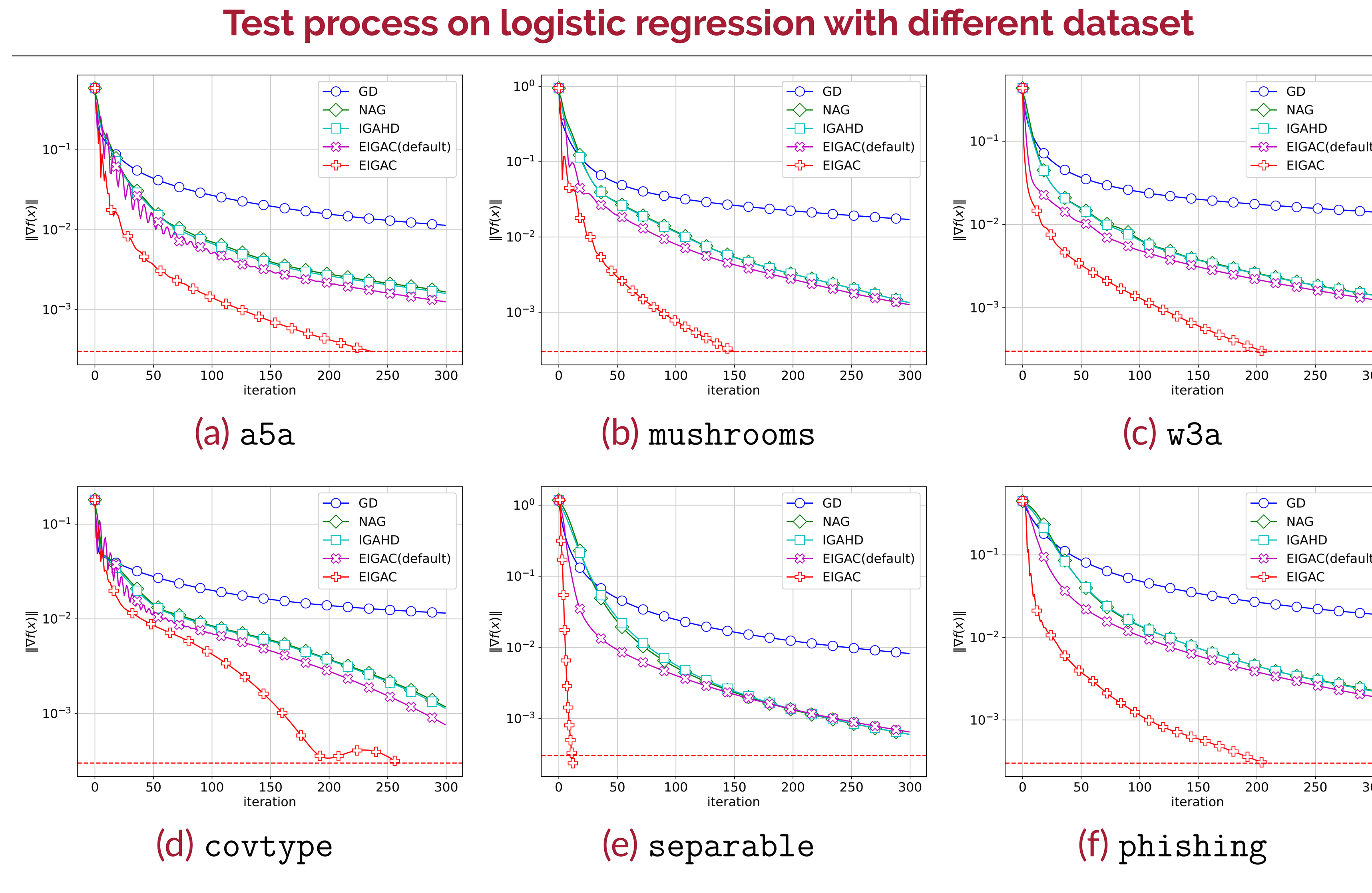
$$\nabla f(X)^\top \nabla^2 f(X)\left(\frac{\partial X}{\partial t}\bigg|_{t=T} \nabla_\theta T(\theta, f) + \frac{\partial X}{\partial \theta}\right) = 0$$

Gradient of P and Q: Combine chain rule and $\nabla_\theta T(\theta, f)$

Nonsmooth cases: Replace gradient with conservative gradient

Convergence: StoPM converges to a feasible stationary point under uniform sufficient decrease condition

Test process on logistic regression with different dataset



(a) a5a (b) mushrooms (c) w3a

(d) covtype (e) separable (f) phishing