# ODE-based Learning to Optimize

**Zhonglin Xie**

Beijing International Center for Mathematical Research
Peking University

Joint work with **Wotao Yin**, **Zaiwen Wen**

**On the Celebration of the 70th Anniversary of Chongqing Normal University**

June 19, 2024

# Outline

## A continuous-time viewpoint of acceleration methods: $\min_x f(x)$

- Gradient descent (GD) method corresponds to gradient flow

$$x_{k+1} = x_k - \sqrt{s}\nabla f(x_k) \quad \Leftrightarrow \quad \dot{x}(t) = -\nabla f(x(t))$$

- Nesterov accelerated gradient (NAG) method corresponds to

$$\begin{cases} x_k = y_{k-1} - s\nabla f(y_{k-1}) \\ y_k = x_k + \dfrac{k-1}{k+2}(x_k - x_{k-1}) \end{cases} \Leftrightarrow \quad \begin{aligned} &\ddot{x}(t) + \frac{3}{t}\dot{x}(t) + \sqrt{s}\nabla^2 f(x(t))\dot{x}(t) \\ &\quad + \left(1 + \frac{3\sqrt{s}}{2t}\right)\nabla f(x(t)) = 0 \end{aligned}$$

- Inertial system with Hessian-driven damping

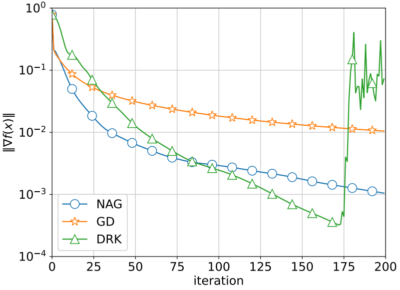
$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta(t)\nabla^2 f(x(t))\dot{x}(t) + \gamma(t)\nabla f(x(t)) = 0 \qquad \text{(ISHD)}$$

# Explicit discretization with fixed stepsize is unstable

Let $w(t) = \gamma(t) - \dot\beta(t) - \beta(t)/t$. Convergence condition for (ISHD) writes

$$\gamma(t) > \dot\beta(t) + \frac{\beta(t)}{t}, \quad t\dot w(t) \leqslant (\alpha - 3)w(t), \quad \text{for all } t \geqslant t_0 \qquad \text{(ISHD-CVG)}$$

Convergence rate: $f(x(t)) - f_\star = \mathcal{O}\left(1/(t^2 w(t))\right)$




- Consider

$$\min_{x\in\mathbb{R}^n} f(x) = \frac{1}{N}\sum_{i=1}^{N} \log(1 + \exp(-b_i\langle a_i, w\rangle))$$

  where the data pairs $\{a_, b_i\} \in \mathbb{R}^n \times \{0,1\}, i \in [N]$

- Set $p = 5, \alpha = 2p+1, \beta(t) \equiv 0$ and $\gamma(t) = p^2 t^{p-2}$
- (ISHD-CVG) holds and $f(x(t)) - f_\star \leq \mathcal{O}(1/t^p)$
- Directly applying 4-th Runge-Kutta diverges!

# Two important questions

- How to translate the fast convergence properties of ODEs to algorithms?

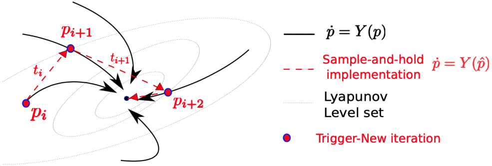
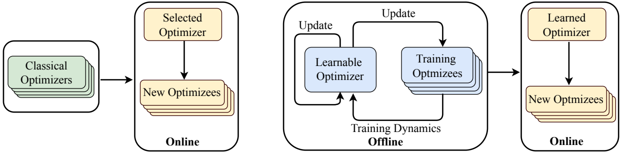



Figure: Rate-matching discretization

Combine error analysis in ODE and complexity analysis in optimization

- How to select the best coefficients for (ISHD)?

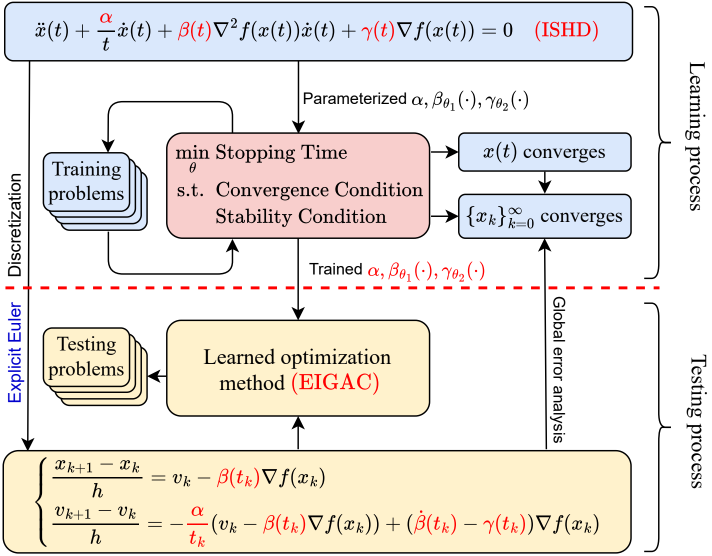



(a) Classic Optimizer (b) Learning to Optimize

Figure: Learning to optimize

A learning to optimize framework with theoretical guarantee

# Our training and testing framework



# Outline

## A fundamental result: an enhanced convergence condition for ISHD

### Theorem 1

*Given* $\kappa \in (0,1], \lambda \in (0, \alpha-1]$, $f$ *is twice differentiable convex,*

$$\begin{aligned} \delta(t) &= t^2(\gamma(t) - \kappa\dot{\beta}(t) - \kappa\beta(t)/t) + (\kappa(\alpha-1-\lambda) - \lambda(1-\kappa))t\beta(t), \\ w(t) &= \gamma(t) - \dot{\beta}(t) - \beta(t)/t, \quad \delta(t) > 0, \quad \text{and} \quad \dot{\delta}(t) \le \lambda t w(t), \end{aligned} \tag{CVG-CDT}$$

*where* $\alpha \geqslant 3, t_0 > 0, \varepsilon > 0$ *are real numbers,* $\beta$ *and* $\gamma$ *are nonnegative continuously differentiable functions defined on* $[t_0, +\infty)$*. Then* $x(t)$ *is bounded and*

$$f(x(t)) - f_\star \le \mathcal{O}\left(\frac{1}{\delta(t)}\right), \ \|\nabla f(x(t))\| \le \mathcal{O}\left(\frac{1}{t\beta(t)}\right), \ \|\dot{x}(t)\| \le \mathcal{O}\left(\frac{1}{t}\right),$$

$$\int_{t_0}^{\infty} (\lambda t w(t) - \dot{\delta}(t))(f(x(t)) - f_\star)\,\mathrm{d}t \le \infty, \quad \int_{t_0}^{\infty} t(\alpha-1-\lambda)\|\dot{x}(t)\|^2\,\mathrm{d}t \le \infty,$$

$$\int_{t_0}^{\infty} t^2\beta(t)w(t)\|\nabla f(x)\|^2\,\mathrm{d}t \le \infty, \quad \int_{t_0}^{\infty} t^2\beta(t)\langle\nabla^2 f(x(t))\dot{x}(t), \dot{x}(t)\rangle\,\mathrm{d}t \le \infty.$$

# Proof: Lyapunov function and term cancelling

- Construct the Lyapunov function

$$
\begin{aligned}
E(t) =& \delta(t)\left(f(x(t)) - f_\star\right) + \frac{1}{2}\|\lambda(x(t) - x_\star) + t(\dot{x}(t) + \kappa\beta(t)\nabla f(x(t)))\|^2 \\
&+ \lambda(1-\kappa)t\beta(t)\langle \nabla f(x(t)), x(t) - x_\star\rangle + \frac{\kappa(1-\kappa)}{2}\|t\beta(t)\nabla f(x)\|^2 \\
&+ \frac{\lambda(\alpha - 1 - \lambda)}{2}\|x(t) - x_\star\|^2
\end{aligned}
$$

- Differentiating through $t$, we set the term with brown color to 0:

$$
\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}t}E(t) =& \dot{\delta}(t)(f(x(t)) - f_\star) - \lambda t w(t)\langle \nabla f(x(t)), x(t) - x_\star\rangle - (\alpha - 1 - \lambda)t\|\dot{x}(t)\|^2 \\
&+ \Big(\delta(t) - (t^2 u(t) + (\kappa(\alpha - 1 - \lambda) - \lambda(1-\kappa))t\beta(t))\Big)\langle \nabla f(x(t)), \dot{x}(t)\rangle \\
&- \kappa t^2\beta(t)w(t)\|\nabla f(x(t))\|^2 - (1-\kappa)t^2\beta(t)\langle \nabla^2 f(x(t))\dot{x}(t), \dot{x}(t)\rangle \le 0
\end{aligned}
$$

- Integrating the inequality above from $t_0$ to $t$ gives $E(t) \le E(t_0)$

# Applying forward Euler scheme to (ISHD)

- Let $v(t_0) = x(t_0) + \beta(t_0)\nabla f(x(t_0))$ and

$$\psi_{\Xi}(x(t), v(t), t) = \begin{pmatrix} v(t) - \beta(t)\nabla f(x(t)) \\ -\frac{\alpha}{t}(v(t) - \beta(t)\nabla f(x(t))) + (\dot{\beta}(t) - \gamma(t))\nabla f(x(t)) \end{pmatrix} \quad (1)$$

- The equation (ISHD) can be reformulated as the first-order system

$$\begin{pmatrix} \dot{x}(t) \\ \dot{v}(t) \end{pmatrix} = \psi_{\Xi}(x(t), v(t), t), \text{ notice that } \nabla^2 f(x(t))\dot{x}(t) = \frac{\mathrm{d}}{\mathrm{d}t}\nabla f(x(t))$$

- Let $h$ be the step size, $t_k = t_0 + kh, k \geq 0$. The forward Euler scheme of the (ISHD) is Explicit Inertial Gradient Algorithm with Correction (EIGAC)

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

# Conditions for stable discretization

**Theorem 2**

*Suppose the assumptions in Theorem 1 and* (CVG-CDT) *hold,* $0 \le C_1$, $0 < C_2 \le 1/h - 1/t_0$, *and* $0 < C_3$ *fulfill* $|\dot\beta(t)| \le C_1\beta(t), |\dot\gamma(t) - \ddot\beta(t)| \le C_2(\gamma(t) - \dot\beta(t)), \beta(t) \le C_3 w(t)$. *Given* $t_0$, $s_0$, *and* $h$, *the sequence* $\{x_k\}_{k=0}^{\infty}$ *is generated by* (EIGAC) *and* $\bar{x}(t)$ *is defined 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:*

$$\Lambda(x, f) \ge \|\nabla^2 f(x)\|, \quad \alpha\beta(t)/t \le \gamma(t) - \dot\beta(t) \le \beta(t)/h, \tag{STB-CDT}$$

$$\sqrt{\int_0^1 \Lambda((1-\tau)X(t, \Xi, f) + \tau\bar{x}(t), f)\, \mathrm{d}\tau} \le \frac{\sqrt{\gamma(t) - \dot\beta(t)} + \sqrt{\gamma(t) - \dot\beta(t) - \frac{\alpha}{t}\beta(t)}}{\beta(t)}.$$

# Key technique: error decomposition

- Local truncated error:

$$\varphi(t) = \begin{pmatrix} x(t+h) - x(t) \\ v(t+h) - v(t) \end{pmatrix} - h \begin{pmatrix} v(t) - \beta(t)\nabla f(x(t)) \\ -\frac{\alpha}{t} v(t) + \left(\frac{\alpha}{t}\beta(t) + \dot{\beta}(t) - \gamma(t)\right)\nabla f(x(t)) \end{pmatrix}$$

- Global error: $r_k = x(t_k) - x_k, s_k = v(t_k) - v_k$, and $e_k = (r_k, s_k)$

- We only need to control $e_{k+1}$, which has two resources

$$e_{k+1} = \begin{pmatrix} r_{k+1} \\ s_{k+1} \end{pmatrix} = \begin{pmatrix} x(t_k) \\ v(t_k) \end{pmatrix} + \begin{pmatrix} x(t_k+h) - x(t_k) \\ v(t_k+h) - v(t_k) \end{pmatrix} - \begin{pmatrix} x_k \\ v_k \end{pmatrix} - h\psi(t_k)$$

$$= \underbrace{\begin{pmatrix} I - h\beta(t_k)G(t_k) & hI \\ (\alpha\beta(t_k)/t_k + \dot{\beta}(t_k) - \gamma(t_k))G(t_k) & (1 - \alpha h/t_k)I \end{pmatrix}}_{W(t_k, G(t_k))} \underbrace{\begin{pmatrix} r_k \\ s_k \end{pmatrix}}_{e_k} + \varphi_{\Xi}(t_k)$$

where $G(t_k) = \int_0^1 \nabla^2 f(x(t_k) + \tau r_k)\,\mathrm{d}\tau$. Abbreviate $W_k = W(t_k, G(t_k))$

# Proof: bound the product of contraction factor

- We estimate $\|e_{n+1}\|$ using

$$\|e_{n+1}\| \le \|W_n\|\|e_n\| + h\|\varphi(t_n)\| \le \underbrace{\prod_{k=0}^{n}\|W_k\|\|e_0\|}_{0} + \|\varphi(t_n)\| + \sum_{k=0}^{n-1}\prod_{l=k+1}^{n}\|W_l\|\|\varphi(t_l)\|$$

- Matrix analysis and (STB-CDT) ensure that $\|W_k\| = \rho(t_k) \le 1 - \alpha h/(2t_k)$

- Define the contraction factor $\rho(t) = \|W(t, G(t))\|$. For $k \le n$, we have

$$\prod_{l=k}^{n}\|W_l\| = \prod_{l=k}^{n}\rho(t_l) = \exp\left(\sum_{l=k}^{n}\ln\left(\rho_l - 1 + 1\right)\right) \le \exp\left(\sum_{l=k}^{n}(\rho_l - 1)\right)$$
$$\le \exp\left(-\sum_{l=k}^{n}\frac{\alpha h}{2t_l}\right) \le \exp\left(-\frac{\alpha}{2}\int_{t_k}^{t_{n+1}}\frac{1}{t}\,\mathrm{d}t\right) = \left(\frac{t_k}{t_{n+1}}\right)^{\alpha/2}$$

# Proof: bound the summation of local truncated errors

- Set $M_1 = \max\{1 + (\alpha+1)/t_0, C_2/h + (1+\alpha/t_0)(1/h + C_1), \alpha/t_0 + 1/h + 1\}$. We have

$$\|\varphi(t)\| \leq M_1 \int_t^{t+h} \left( \frac{\alpha}{t} \|\dot{x}(\tau)\| + \|\beta(\tau)\nabla f(x(\tau))\| + \beta(\tau)\|\nabla^2 f(x(\tau))\dot{x}(\tau)\| \right) \mathrm{d}\tau$$

- Using Cauchy inequality and Theorem 1, for certain $M_3$, we have

$$\|\varphi(t)\| \leq o(1/t) \quad \text{and} \quad \sum_{k=0}^{n} t_k^{\alpha/2} \|\varphi(t_k)\| \leq M_3 t_n^{\alpha/2 - 1/2}$$

- Combining these results, we have

$$\|e_{n+1}\| \leq \|\varphi(t_n)\| + \sum_{k=0}^{n-1} \prod_{l=k+1}^{n} \|W_l\| \|\varphi(t_l)\| \leq \sum_{k=0}^{n} \left( \frac{t_{k+1}}{t_{n+1}} \right)^{\alpha/2} \|\varphi(t_k)\| \leq M_3 \frac{1}{\sqrt{t_{n+1}}}$$

# Proof: derive the function value minimization rate

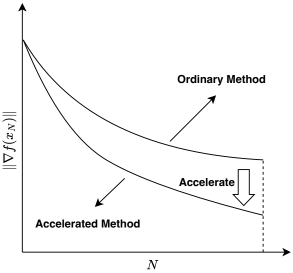
- The function value can be decomposed as

$$
\begin{aligned}
f(x_k) - f_\star \leq & |f(x_k) - f(x(t_k))| + |f(x(t_k)) - f_\star| \\
\leq & \underbrace{\|\nabla f(x(t_k))\|}_{\mathcal{O}(1/(t_k\beta(t_k)))} \|e_k\| + \frac{1}{2} \underbrace{\left\| \int_0^1 \nabla^2 f(x(t_k) + \tau e_k) \,\mathrm{d}\tau \right\|}_{\mathcal{O}(1/\beta(t_k))} \underbrace{\|e_k\|^2}_{\mathcal{O}(1/t_k)} + \frac{E(t_0)}{t_k^2 w(t_k)}
\end{aligned}
$$

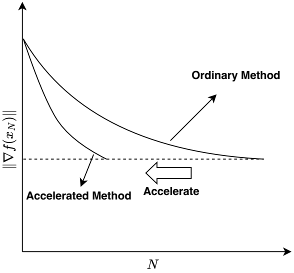
- The rate is at least $\mathcal{O}(1/k)$, while the dominate term comes from the global error

# Outline

# Stopping time: a differentiable continuous-time complexity




(a) Measure-based




(b) Complexity-based

**Definition 3 (Stopping Time)**

Given the initial time $t_0$, the initial value $x_0$, the initial velocity $\dot{x}(t_0)$, the trajectory $X(\Xi, t, f)$ of the system (ISHD), and a tolerance $\varepsilon$, the stopping time of the criterion $\|\nabla f(x)\| \leq \varepsilon$ is

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

# Tackle the point-wise contraints using integration

- With $w(t), \delta(t)$ defined in (CVG-CDT), we introduce

$$p(x, \bar{x}, \Xi, t, f) = \Bigg[\beta(t)\sqrt{\int_0^1 \Lambda((1-\tau)x + \tau\bar{x}, f)\,\mathrm{d}\tau} - \sqrt{\gamma(t) - \dot{\beta}(t)} - \sqrt{\gamma(t) - \dot{\beta}(t) - \frac{\alpha}{t}\beta(t)}\Bigg]_+$$

$$q(\Xi, t) = \left[\gamma(t) - \dot{\beta}(t) - \beta(t)/h\right]_+ + \left[\dot{\beta}(t) + \alpha\beta(t)/t - \gamma(t)\right]_+ + \left[\dot{\delta}(t) - \lambda t w(t)\right]_+ + [-\delta(t)]_+$$

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

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

# A L2O framework for selecting the best coefficients

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

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

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

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

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

# Solving the L2O problem using exact penalty method

Given the penalty parameter $\rho$, the $\ell_1$ exact penalty problem writes

$$\begin{aligned}\min_{\theta} \Upsilon(\theta) &= \mathbb{E}_f[T(\theta, f)] + \rho\left(\mathbb{E}_f[P(\theta, f)] + \mathbb{E}_f[Q(\theta, f)]\right) \\ &= \mathbb{E}_f\left[T(\theta, f) + \rho\left(P(\theta, f) + Q(\theta, f)\right)\right]\end{aligned}$$

**Algorithm** Stochastic Penalty Method (StoPM) for L2O problem

**Input:** initial weight $\theta_0$, penalty coefficient $\rho$, training dataset $\mathcal{F}$
**while** Not(`Stopping Condition`) **do**
  Sample a function: $f_k \in \mathcal{F}$
  Computing the gradients $J_T, J_P$ and $J_Q$ correspond to $T, P$ and $Q$
  Update variable: $\theta_{k+1} \leftarrow \theta_k - \eta(D_T + D_P + D_Q)$
  Update index: $k \leftarrow k + 1$
**end while**
**Output:** the trained weight $\theta_\star$

# Outline

# Conservative gradient

- When parameterize $\alpha, \beta, \gamma$ using neural networks, they may be nonsmooth
- The output of *auto differentiation* in nonsmooth functions may not be Clarke subdifferentials, but they are certainly conservative gradients
- Consider the example:

$$f(s) = ([-s]_+ + s) - [s]_+ \equiv 0 \quad \underset{\text{autograd using TensorFlow}}{\Longrightarrow} \quad g(s) = \begin{cases} 0 & (s \neq 0) \\ 1 & (s = 0) \end{cases}$$

  $g$ is not the Clarke subdifferential of $f$ but a conservative gradient
- Conservative gradient generalizes subdifferentials while preserving chain rule
- $\Psi$ is termed the conservative Jacobian (gradient if $m = 1$) of $\pi$ if and only if

$$\frac{\mathrm{d}}{\mathrm{d}\iota}\pi(r(\iota)) = A\dot{r}(\iota), \quad \text{for all } A \in \Psi(r(\iota)), \text{ for almost all } \iota \in [0,1]$$

  for any absolutely continuous curve $r : [0,1] \to \mathbb{R}^d$

# Differentiate through the ODE flow of (ISHD): $\partial X / \partial \theta$

- Reformulate (ISHD) as a first-order system (1) with a parameterized right-hand-side term $\psi$:

$$\psi \colon \mathbb{R}^{2n+1+p} \to \mathbb{R}^{2n}, \quad (x, v, t, \theta) \mapsto \psi_\theta(x, v, t).$$

  Denote the flow of (1) with parameterized $\psi$ as $X(x_0, v_0, \theta, t)$

- Denote $D^\psi \colon \mathbb{R}^{2n+1+p} \rightrightarrows \mathbb{R}^{2n \times (2n+1+p)}$ as a conservative Jacobian of $\psi$ with respect to $(x, v, t, \theta)$. The coordinate projection (partial derivative) writes $D^\psi_{x,v} = \Pi_{x,v} D^\psi$, $D^\psi_t = \Pi_t D^\psi$ and $D^\psi_\theta = \Pi_\theta D^\psi$

- Applying the general result to the first-order system (1): $\theta \mapsto A(t_0)$ is a conservative Jacobian of $\theta \to X(x_0, v_0, \theta, t_1)$ (smooth version: $\partial X / \partial \theta$)

$$\dot{A}(t) = D^\psi_{x,v}(t) A(t) + D^\psi_\theta(t), \quad A(t_1) = 0_{2n \times p} \quad \text{for all } t \in [t_0, t_1]$$

- Smooth version:

$$\frac{\partial X}{\partial \theta} = \int_{t_0}^{t_1} \frac{\partial \psi_\theta}{\partial x} \frac{\mathrm{d} X}{\mathrm{d} \theta} + \frac{\partial \psi_\theta}{\partial \theta} \, \mathrm{d} t$$

# Evaluate the derivative of stopping time: $\nabla_\theta T(\theta, f)$

- Take limit by continuity: $\|\nabla f(X(T(\theta, f), f, \theta))\|^2 - \varepsilon^2 \equiv 0$
- Implicit function theorem (valid in nonsmooth case):

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

where $T = T(\theta, f), X = X(T(\theta, f), f, \theta)$

- Invoking the first-order form of (ISHD):

$$\frac{\partial X}{\partial t}\bigg|_{t=T} = \dot{x}(T) = v(T) - x(T) - \beta(T)\nabla f(x(T))$$

where $x(t) = X(t, f, \theta)$

- The derivative:

$$\nabla_\theta T(\theta, f) = \left( \nabla f(X)^\top \nabla^2 f(X) \left( v(T) - X - \beta(T)\nabla f(X) \right) \right)^{-1} \nabla f(X)^\top \nabla^2 f(X) \frac{\partial X}{\partial \theta}$$

# Conservative gradient of the constraints

- Recap:

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

- Applying the chain rule gives

$$\frac{\mathrm{d}P(\theta,f)}{\mathrm{d}\theta} = \int_{t_0}^{T(\theta,f)} \frac{\partial \psi_\theta(s(t),t,f)}{\partial\theta} w(t) + \frac{\partial p(x(t),\bar{x}(t),\theta,t,f)}{\partial\theta}\,\mathrm{d}t + p(x(T),\bar{x}(T),\theta,T,f)\frac{\mathrm{d}T(\theta,f)}{\mathrm{d}\theta}$$

$$\frac{\mathrm{d}Q(\theta,f)}{\mathrm{d}\theta} = \int_{t_0}^{T(\theta,f)} \frac{\mathrm{d}q(\theta,t)}{\mathrm{d}\theta}\,\mathrm{d}t + q(\theta,T)\frac{\mathrm{d}T(\theta,f)}{\mathrm{d}\theta}$$

where $\bar{x}(\cdot)$ is the interpolation of $\{x_k\}$ defined in Theorem 2, $w(\cdot)$ is the solution of

$$-\begin{pmatrix} \frac{\partial p(x(t),\bar{x}(t),\theta,t,f)}{\partial x} \\ 0_{n\times 1} \end{pmatrix} - \frac{\partial \psi_\theta(s(t),t,f)}{\partial s} w(t) = \dot{w}(t), \quad w(T(\theta,f)) = 0_{2n\times 1}$$

# Clarke subdifferential of the point-wise maximal function

- Let $\{f_\eta : \mathbb{R}^n \to (-\infty, +\infty]\}_{\eta \in A}$ be a family of *proper convex* functions and
$$f(x) = \sup_{\eta \in A} f_\eta(x)$$
- If $x_0 \in \bigcap_{\eta \in A} \text{int dom} f_\eta$, and $I(x_0) = \{\eta \in A \mid f_\eta(x_0) = f(x_0)\}$, then
$$\text{conv}\left( \bigcup_{\eta \in I(x_0)} \partial f_\eta(x_0) \right) = \partial f(x_0)$$
- $\lambda_{\max}(A) = \sup_{\|u\|=1} u^\top A u$. Set $z = \arg\max_{\|u\|=1} u^\top \nabla^2 f(x) u$. We have

$$\frac{\partial \Lambda(f,x)}{\partial x} = \left\{ \left\langle \frac{\partial \lambda_{\max}(A)}{\partial A}\bigg|_{A=\nabla^2 f(x)}, \frac{\partial \nabla^2 f(x)}{\partial x_k} \right\rangle \right\}_k = \left\{ \sum_{i,j} \partial_{ijk} f(x) z_i z_j \right\}_k$$
$$= \frac{\mathrm{d}}{\mathrm{d}\eta_2} \left( \frac{\mathrm{d}}{\mathrm{d}\eta_1} \nabla f(x + \eta_1 z + \eta_2 z)\bigg|_{\eta_1=0} \right)\bigg|_{\eta_2=0} = D^3 f(x)[z,z]$$

when $\Lambda(f,x) = \lambda_{\max}(\nabla^2 f(x))$. This enables the evaluation of $\partial p/\partial\theta, \partial p/\partial x$

# Outline

# Criteria for *D*-stationarity using directional derivative

- Let $\Upsilon$ be Lipschitz continuous near $\bar{\theta}$
- The *D*-directional derivative of $\Upsilon$ at $\bar{\theta}$ along a nonzero vector $\vartheta$:

$$\Upsilon^{\circ}(\bar{\theta};\vartheta) \triangleq \limsup_{\tau\downarrow 0} D^{\Upsilon}(\bar{\theta}+\tau\vartheta)$$

- *D*-stationarity: $0 \in \partial\Upsilon(\theta)$
- Since *D* has closed convex graph, $\theta$ is a *D*-stationary point of $\Upsilon$ if and only if $\Upsilon^{\circ}(\theta;\vartheta) \geq 0$ for all $\vartheta \in \mathbb{R}^{d_\theta}$

# Precludes infeasible stationary point using sufficient decrease condition

- Given the training dataset $\mathcal{F}$, we denote the residual function as
$$R(\theta) = \mathbb{E}_f[P(\theta, f) + Q(\theta, f)]$$
- This function measures the constraints violation. The feasible set is defined by
$$S = \{\theta \mid P(\theta, f) \leq 0, Q(\theta, f) \leq 0, \forall f \in \mathcal{F}\}$$

**Assumption 1 (Uniform sufficient decrease condition)**

*For each infeasible point $\theta$, i.e. $\theta \notin S$, there exists a nonzero vector $\vartheta$, such that $R^\circ(\theta; \vartheta) \leq -c\|\vartheta\|$. Here the constant $c$ is uniform for each $\theta$.*

**Theorem 4**

*Suppose $\mathbb{E}_f[T(\theta, f)]$ is globally Lipschitz continuous with Lipschitz constant $L_T$. Let Assumption 1 hold. Given the penalty parameter $\rho > L_T/c$, any infeasible point of the penalty function $\Upsilon$ can not be a D-stationary point.*

# Sufficient decrease condition precludes infeasible stationary point

- Consider the penalty function

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

- For any infeasible point $\theta$, using Assumption 1, there must exists a direction $\vartheta$, such that

$$\Upsilon^\circ(\theta; \vartheta) = \mathbb{E}_f[T(\cdot, f)]^\circ(\theta; \vartheta) + \rho R^\circ(\theta, \vartheta) \leq L_T\|\vartheta\| - c\rho\|\vartheta\| < 0$$

- The criteria of $D$-stationary point ensures that $\theta$ cannot be a $D$-stationary point of $\Upsilon$

# SGD converges with (nonsmooth) auto-differentiation: Assumptions

**Assumption 2 (Assumptions of the SGD)**

1. *The step sizes* $\{\eta_k\}_{k\geq 1}$ *satisfy*

$$\eta_k \geq 0, \quad \sum_{k=1}^{\infty} \eta_k = \infty, \quad \text{and} \quad \sum_{k=1}^{\infty} \eta_k^2 < \infty.$$

2. *Almost surely, the iterates* $\{\theta_k\}_{k\geq 1}$ *are bounded, i.e.,* $\sup_{k\geq 1} \|\theta_k\| < \infty$.
3. $\{\xi_k\}_{k\geq 1}$ *is a uniformly bounded difference martingale sequence with respect to the increasing* $\sigma$*-fields*

$$\mathscr{F}_k = \sigma(\theta_j, \varrho_j, \xi_j : j \leq k).$$

*In other words, there exists a constant* $M_\xi > 0$ *such that*

$$\mathbb{E}[\xi_k \mid \mathscr{F}_k] = 0 \quad \text{and} \quad \mathbb{E}[\|\xi_k\|^2 \mid \mathscr{F}_k] \leq M_\xi \quad \text{for all} \quad k \geq 1.$$

# SGD converges with (nonsmooth) auto-differentiation

**Assumption 3**

*The complementary of* $\{\Upsilon(\theta) \mid 0 \in D^{\Upsilon}(\theta)\}$ *is dense in* $\mathbb{R}$.

**Theorem 5 (SGD converges using conservative gradient)**

*Suppose that Assumptions 2 and 3 hold. Then every limit point of* $\{\theta_k\}_{k\geq 1}$ *is stationary and the function values* $\{\Upsilon(\theta_k)\}_{k\geq 1}$ *converge.*

**Theorem 6 (Convergence guarantee for Algorithm 1)**

*Suppose Assumption 1, 2 and 3 hold,* $\{\theta_k\}_{k\geq 1}$ *is generated by Algorithm 1. Then almost surely, every limit point* $\theta_\star$ *of* $\{\theta_k\}_{k\geq 1}$ *satisfies* $\theta_\star \in S$, $0 \in D^{\Upsilon}(\theta_\star)$ *and the sequence* $\{\Upsilon(\theta_k)\}_{k\geq 1}$ *converges.*

# Outline

# Setting and datasets

- Consider the logistic regression problem defined by

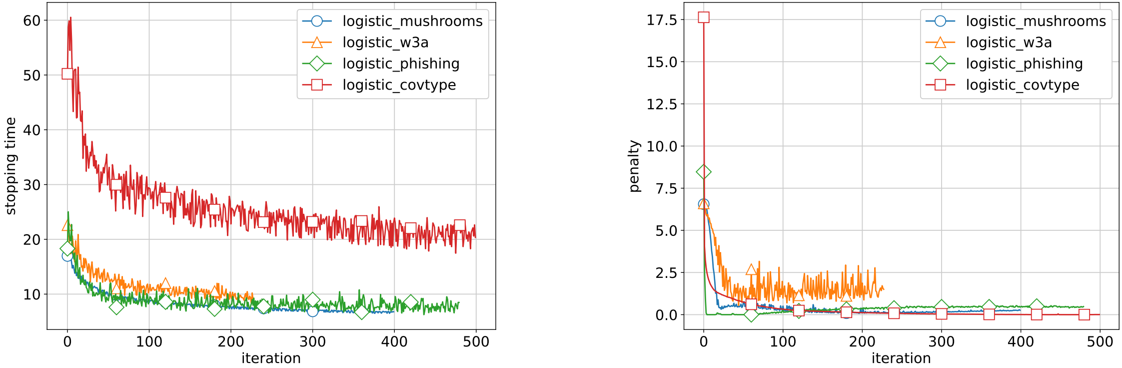
$$\min_{x \in \mathbb{R}^n} f_{\mathscr{D}}(x) = \frac{1}{|\mathscr{D}|} \sum_{(a_i, b_i) \in \mathscr{D}} \log(1 + \exp(-b_i \langle a_i, x \rangle)),$$

where $\mathscr{D}$ is a subset of a given dataset $\Sigma$ and $\{a_i, b_i\} \in \mathbb{R}^n \times \{0, 1\}, i \in [|\mathscr{D}|]$

- The datasets are listed as below

| Dataset | $n$ | $N_{\text{train}}$ | $N_{\text{test}}$ | Separable |
|---|---|---|---|---|
| `a5a` | 123 | 6,414 | 26,147 | No |
| `w3a` | 300 | 4,912 | 44,837 | No |
| `mushrooms` | 112 | 3,200 | 4,924 | Yes |
| `covtype` | 54 | 102,400 | 478,612 | No |
| `phishing` | 68 | 8,192 | 2,863 | No |
| `separable` | 101 | 20,480 | 20,480 | Yes |

# Training results




(a) Stopping time on logistic regression

(b) Penalty on logistic regression

Figure: The training process in different tasks.

# Testing: Compared methods

- **GD**. $x_{k+1} = x_k - h\nabla f(x_k)$. We set the stepsize as $h = 1/L$
- **NAG**. We choose $h = 1/L$ and employ the version for convex functions

$$y_{k+1} = x_k - h\nabla f(x_k), \quad x_{k+1} = y_{k+1} + \frac{k-1}{k+2}(y_{k+1} - y_k)$$

- **EIGAC**. Explicit inertial gradient algorithm with correction. We provide two versions of EIGAC: default coefficients

$$\alpha = 6, \quad \beta(t) = (4/h - 2\alpha/t)/L, \quad \textit{and} \quad \beta(t) = h\gamma(t)$$

and the coefficients learned by Algorithm 1

- **IGAHD**. Inertial gradient algorithm with Hessian-driven damping. This method is obtained by applying a NAG inspired time discretization of

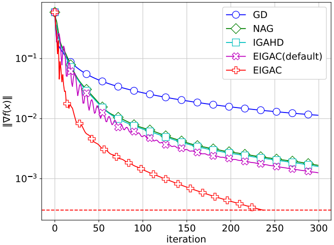
$$\ddot{x}(t) + \frac{\alpha}{t}\dot{x}(t) + \beta\nabla^2 f(x(t))\dot{x}(t) + \left(1 + \frac{\beta}{t}\right)\nabla f(x(t)) = 0$$

Let $s = 1/L$. In each iteration, setting $\alpha_k = 1 - \alpha/k$, the method performs

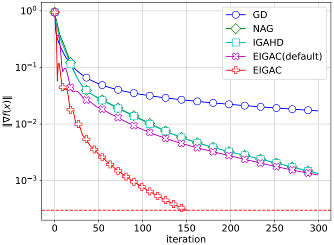
$$\begin{cases} y_k = x_k + \alpha_k\left(x_k - x_{k-1}\right) - \beta\sqrt{s}\left(\nabla f\left(x_k\right) - \nabla f\left(x_{k-1}\right)\right) - \dfrac{\beta\sqrt{s}}{k}\nabla f\left(x_{k-1}\right) \\ x_{k+1} = y_k - s\nabla f\left(y_k\right) \end{cases}$$

IGAHD owns $\mathcal{O}(1/k^2)$ convergence rate when $0 \leq \beta < 2/\sqrt{s}$ and $s \leq 1/L$

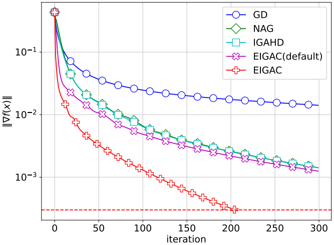
# Testing results



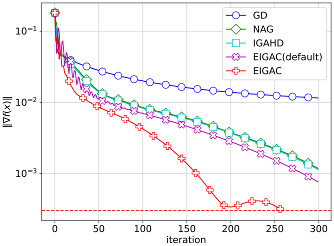

(a) a5a



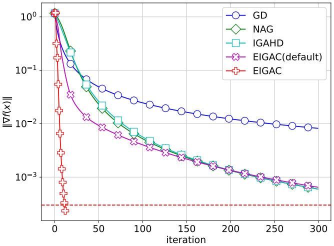

(b) mushrooms



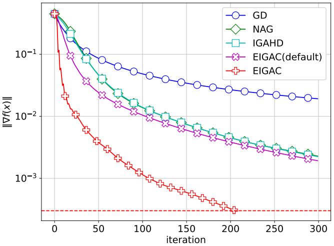

(c) w3a




(d) covtype




(e) separable




(f) phishing

# References I


[1] Hedy Attouch et al. "First-order optimization algorithms via inertial systems with Hessian-driven damping". In: *Math. Program.* 193.1 (2022), pp. 113–155. ISSN: 0025-5610,1436-4646.

[2] Ying Cui, Junyi Liu, and Jong-Shi Pang. "Nonconvex and nonsmooth approaches for affine chance-constrained stochastic programs". In: *Set-Valued Var. Anal.* 30.3 (2022), pp. 1149–1211. ISSN: 1877-0533,1877-0541.

[3] Damek Davis et al. "Stochastic subgradient method converges on tame functions". In: *Found. Comput. Math.* 20.1 (2020), pp. 119–154. ISSN: 1615-3375,1615-3383.

[4] Hao Luo and Long Chen. "From differential equation solvers to accelerated first-order methods for convex optimization". In: *Math. Program.* 195.1-2 (2022), pp. 735–781. ISSN: 0025-5610,1436-4646.

[5] Bin Shi et al. "Understanding the acceleration phenomenon via high-resolution differential equations". In: *Math. Program.* 195.1-2 (2022), pp. 79–148. ISSN: 0025-5610,1436-4646.

# References II

[6] Weijie Su, Stephen Boyd, and Emmanuel J. Candès. "A differential equation for modeling Nesterov's accelerated gradient method: theory and insights". In: *J. Mach. Learn. Res.* 17 (2016), Paper No. 153, 43. ISSN: 1532-4435,1533-7928.

[7] Andre Wibisono, Ashia C. Wilson, and Michael I. Jordan. "A variational perspective on accelerated methods in optimization". In: *Proc. Natl. Acad. Sci. USA* 113.47 (2016), E7351–E7358. ISSN: 0027-8424,1091-6490.

**Many Thanks For Your Attention!**