Universality of AdaGrad Stepsizes for Stochastic Optimization: Inexact Oracle, Acceleration and Variance Reduction

Anton Rodomanov (CISPA)

Joint work with Xiaowen Jiang (CISPA) and Sebastian Stich (CISPA)

1 July 2024
EURO Conference on Operational Research
Copenhagen, Denmark

Motivation

Stochastic Convex Optimization

Problem:

$$f^* = \min_{x \in Q} f(x),$$

where $f : \mathbb{R}^d \to \mathbb{R}$ is a convex function, $Q \subseteq \mathbb{R}^d$ is a simple convex set.

Stochastic gradient oracle: Random vector $g(x, \xi) \in \mathbb{R}^d$ (ξ is a r.v.),

$$\mathbb{E}_\xi[g(x, \xi)] = \nabla f(x).$$

Main example: $f(x) = \mathbb{E}_\xi[f(x, \xi)]$. Then, $g(x, \xi) = \nabla_x f(x, \xi)$.

E.g.: $f(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \implies g(x, \xi) = \frac{1}{b} \sum_{j=1}^{b} \nabla f_{\xi_j}(x)$, where
$\xi = (\xi_1, \ldots, \xi_b)$ with i.i.d. components from $\mathrm{Unif}(\{1, \ldots, n\})$.

Stochastic Gradient Method (SGD)

Problem: $f^* = \min_{x \in Q} f(x)$.

Stochastic Gradient Method (SGD):

$$x_{k+1} = \pi_Q(x_k - h_k g_k), \quad g_k \cong \hat{g}(x_k),$$

where $\pi_Q(x) = \operatorname{argmin}_{y \in Q} \|x - y\|$ is the Euclidean projection onto Q.

Main questions:

- How to choose step sizes h_k?
- What is the rate of convergence?

Convergence Guarantees for SGD

Assume that:

- Q is bounded: $\|x-y\| \leq D, \ \forall x, y \in Q$.
- Variance of $\hat{g}$ is bounded: $\mathbb{E}_\xi[\|g(x,\xi) - \nabla f(x)\|_*^2] \leq \sigma^2, \ \forall x \in Q$.

Nonsmooth optimization: $\|\nabla f(x)\|_* \leq L_0$, $\forall x \in Q$.

$$h_k = \frac{D}{\sqrt{(L_0^2 + \sigma^2)(k+1)}} \quad \Longrightarrow \quad \mathbb{E}[f(\bar{x}_k)] - f^* \leq O\Big(\frac{(L_0 + \sigma)D}{\sqrt{k}}\Big),$$

where $\bar{x}_k = \frac{1}{k}\sum_{i=0}^{k-1} x_i$.

Smooth optimization: $\|\nabla f(x) - \nabla f(y)\|_* \leq L_1\|x-y\|$, $\forall x, y \in Q$.

$$h_k = \min\Big\{\frac{1}{2L_1}, \frac{D}{\sigma\sqrt{k+1}}\Big\} \quad \Longrightarrow \quad \mathbb{E}[f(\bar{x}_k)] - f^* \leq O\Big(\frac{L_1 D^2}{k} + \frac{\sigma D}{\sqrt{k}}\Big).$$

Discussion

- What we saw previously is the standard approach in Optimization:
  1. Fix a certain Problem class $\mathcal{P}$.
  2. Develop a "good" method tailored to $\mathcal{P}$.
- However:
  - A specific problem may belong to multiple problem classes.
  - Different problems may belong to different problem classes.
- Ideally, we would like to have universal algorithms suitable for multiple problem classes at the same time.

Universal Gradient Methods [Nesterov 2015]

Problem: $\min_{x \in Q} f(x)$.

Hölder constants: $H_\nu := \sup\limits_{x,y \in Q; x \neq y} \frac{\|\nabla f(x) - \nabla f(y)\|_*}{\|x-y\|^\nu}, \; \nu \in [0, 1]$.

Note:

- $\nu = 1$: $\|\nabla f(x) - \nabla f(y)\|_* \leq H_1 \|x - y\|$ (Lipschitz gradient).
- $\nu = 0$: $\|\nabla f(x) - \nabla f(y)\|_* \leq H_0$ (contains Lipschitz functions). This class is better than $\|\nabla f(x)\|_* \leq L_0$.
- If $H_{\nu_1}, H_{\nu_2} < +\infty$ for some $\nu_1 \leq \nu_2$, then $H_\nu < +\infty, \forall \nu \in [\nu_1, \nu_2]$.

Main assumption: There exists $\nu \in [0, 1]$ such that $H_\nu < +\infty$.

Universal Gradient Methods – II

Method: $x_{k+1} = \pi_Q(x_k - \frac{1}{M_k}\nabla f(x_k))$, where M_k is found by line search to satisfy the following condition:

$$f(x_{k+1}) \le f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{M_k}{2}\|x_{k+1} - x_k\|^2 + \frac{\epsilon}{2}.$$

Efficiency bound: $O\Big(\inf\limits_{\nu \in [0,1]} \Big(\frac{H_\nu}{\epsilon}\Big)^{\frac{2}{1+\nu}} D^2\Big)$ iters to $f(x_k^*) - f^* \le \epsilon$.

Universal Fast Gradient Method: $O\Big(\inf\limits_{\nu \in [0,1]} \Big(\frac{H_\nu D^{1+\nu}}{\epsilon}\Big)^{\frac{2}{1+3\nu}}\Big)$.

Great methods but they do not work with stochastic oracle!

AdaGrad Methods

AdaGrad [McMahan and Streeter 2010; Duchi et al. 2011]: $(g_k \cong \hat{g}(x_k))$

$$x_{k+1} = \pi_Q(x_k - h_k g_k), \qquad h_k = \frac{D}{\sqrt{\sum_{i=0}^{k} \|g_i\|_*^2}}.$$

Foundation of nowadays popular Adam, RMSProp, ...

Convergence rate [Levy et al. 2018]: If $\nabla f(x^*) = 0$, then

$$\mathbb{E}[f(\bar{x}_k)] - f^* \leq O\left(\min\left\{\frac{L_0 D}{\sqrt{k}}, \frac{L_1 D^2}{k}\right\} + \frac{\sigma D}{\sqrt{k}}\right),$$

(L_0, L_1 are the Lipschitz constants of f, ∇f; σ is the variance.)

UniXGrad [Kavis et al. 2019]: Accelerated gradient method with AdaGrad step sizes based on difference of gradients. Convergence rate:

$$O\left(\min\left\{\frac{L_0 D}{\sqrt{k}}, \frac{L_1 D^2}{k^2}\right\} + \frac{\sigma D}{\sqrt{k}}\right).$$

Motivation and Related Work

Develop "fully universal" gradient methods that automatically adjust to the right Hölder class and oracle's variance.

Related work:

- Universal methods with line search [Nesterov 2015; Grapiglia and Nesterov 2017; Grapiglia and Nesterov 2020; Doikov and Nesterov 2021; Doikov, Mishchenko, et al. 2024]. Only for deterministic optimization.
- Adaptive methods for stochastic optimization [McMahan and Streeter 2010; Duchi et al. 2011; Levy et al. 2018; Kavis et al. 2019; Ene et al. 2021] No specific guarantees for Hölder class.
- Parameter-free methods [Orabona 2014; Cutkosky and Boahen 2017; Cutkosky and Orabona 2018; Jacobsen and Cutkosky 2023; Carmon and Hinder 2022; Defazio and Mishchenko 2023] Slightly different focus, also no specific guarantees for Hölder class (with stochastic oracle).

Universal Stochastic Gradient Methods [Rodomanov et al. 2024]

Problem: $\min_{x \in \operatorname{dom} \psi} [F(x) = f(x) + \psi(x)]$, f and ψ are convex, ψ is simple.

Assumptions:

1. Hölder gradient: $\|\nabla f(x) - \nabla f(y)\|_* \leq H_\nu \|x - y\|^\nu$, $\nu \in [0, 1]$.
2. Bounded domain: $\|x - y\| \leq D, \quad \forall x, y \in \operatorname{dom} \psi$.
3. Stochastic oracle: $\mathbb{E}_\xi[g(x, \xi)] = \nabla f(x)$, $\mathbb{E}_\xi[\|g(x, \xi) - \nabla f(x)\|_*^2] \leq \sigma^2$.

Methods using (modified) AdaGrad stepsizes and needing to know only D:

- Basic method: $O\left(\inf_{\nu \in [0,1]} \frac{H_\nu D^{1+\nu}}{k^{(1+\nu)/2}} + \frac{\sigma D}{\sqrt{k}} \right)$.
- Accelerated method: $O\left(\inf_{\nu \in [0,1]} \frac{H_\nu D^{1+\nu}}{k^{(1+3\nu)/2}} + \frac{\sigma D}{\sqrt{k}} \right)$.

This work: Show that AdaGrad stepsizes are even more universal.

Main Algorithms and Results for Uniformly Bounded Variance

Problem Formulation – I: Approximate Smoothness

Problem: $\min\limits_{x \in \operatorname{dom}\psi} [F(x) = f(x) + \psi(x)]$, f and ψ are convex, ψ is simple.

Main assumption: f is approximately smooth: there exist $L_f, \delta_f \geq 0$ and $\bar{f} \colon \mathbb{R}^d \to \mathbb{R}$, $\bar{g} \colon \mathbb{R}^d \to \mathbb{R}^d$ such that, for any $x, y \in \mathbb{R}^d$, we have

$$0 \leq \left[\beta_{f,\bar{f},\bar{g}}(x, y) := f(y) - \bar{f}(x) - \langle \bar{g}(x), y - x \rangle\right] \leq \frac{L_f}{2}\|y - x\|^2 + \delta_f.$$

NB: This is the (δ, L)-oracle introduced by [Devolder et al. 2013].

Examples:

- f is L-smooth $\iff$ $(\bar{f}, \bar{g}) = (f, \nabla f)$ with $L_f = L$, $\delta_f = 0$
- f is (ν, H_ν)-Hölder smooth $\implies$ $(\bar{f}, \bar{g}) = (f, \nabla f)$ with $L_f = [\frac{1-\nu}{2(1+\nu)\delta_f}]^{\frac{1-\nu}{1+\nu}} H_\nu^{\frac{2}{1+\nu}}$ and any $\delta_f > 0$.
- $\phi(x) \leq f(x) \leq \phi(x) + \delta$, $\forall x$, with L-smooth ϕ $\implies$ $(\bar{f}, \bar{g}) = (\phi, \nabla\phi)$ with $L_f = L$, $\delta_f = \delta$.
- $f(x) = \max_u \Psi(x, u)$ with str. concave Ψ, $\bar{u}(x) \approx_\delta \operatorname{argmax}_u \Psi(x, u)$ $\implies$ $\bar{f}(x) = \Psi(x, \bar{u}(x))$, $\bar{g}(x) = \nabla_u \Psi(x, \bar{u}(x))$ with $\delta_f = \delta$.

Problem Formulation – II

Problem: $F^* = \min_{x \in \operatorname{dom} \psi} [F(x) = f(x) + \psi(x)]$.

Assumptions:

1. f is (δ_f, L_f)-approximately smooth with components $(\bar{f}, \bar{g})$.
2. f can be accessed only via unbiased stochastic oracle $\hat{g}$ for $\bar{g}$: $\mathbb{E}_\xi[g(x,\xi)] = \bar{g}(x)$.
3. Uniformly bounded variance: $\mathrm{Var}_{\hat{g}}(x) := \mathbb{E}_\xi[\|g(x,\xi) - \bar{g}(x)\|_*^2] \le \sigma^2$.
4. Bounded domain: $\|x - y\| \le D$, $\forall x, y \in \operatorname{dom} \psi$.

Note: Asm. 4 can always be ensured with $D = 2R_0$ whenever we know $R_0 \ge \|x_0 - x^*\|$ by considering $F^* = \min_{x \in \operatorname{dom} \psi_D} [F_D(x) = f(x) + \psi_D(x)]$, where $\psi_D = \psi + \mathrm{Ind}_{B_0}$ with $B_0 = \{x : \|x - x_0\| \le R_0\}$.

Basic Universal Gradient Method

Algorithm 1 $\text{UniSgd}_{\hat{g},\psi}(x_0; D)$

```
$g_0 \cong \hat{g}(x_0)$.
for $k = 0, 1 \dots$ do
    $x_{k+1} = \text{Prox}_\psi(x_k, g_k, M_k)$,  $g_{k+1} \cong \hat{g}(x_{k+1})$.
    $M_{k+1} = \sqrt{M_k^2 + \frac{1}{D^2}\|g_{k+1} - g_k\|_*^2}$.
```

Prox-mapping: $\text{Prox}_\psi(x, g, M) = \operatorname*{argmin}_{y \in \text{dom}\,\psi} \{\langle g, y\rangle + \psi(y) + \frac{M}{2}\|y - x\|^2\}$.

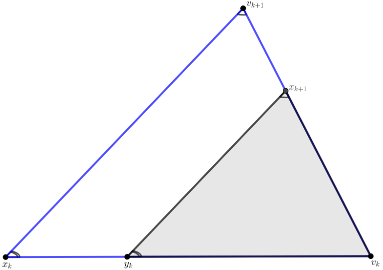
Output point: $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$.

Convergence rate: $\mathbb{E}[F(\bar{x}_k)] - F^* \leq O\left(\frac{L_f D^2}{k} + \frac{\sigma D}{\sqrt{k}} + \delta_f\right)$.

Accelerated Universal Gradient Method

Algorithm 2 UniFastSgd$_{\hat{g},\psi}(x_0; D)$

```
$v_0 = x_0,\ M_0 = A_0 = 0.$
for $k = 0, 1, \dots$ do
    $a_{k+1} = \frac{1}{2}(k+1),\ A_{k+1} = A_k + a_{k+1}$
    $y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} v_k,\quad g_{y_k} \cong \hat{g}(y_k).$
    $v_{k+1} = \mathrm{Prox}_\psi(v_k, g_{y_k}, M_k / a_{k+1}).$
    $x_{k+1} = \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} v_{k+1}.$
    $g_{x_{k+1}} \cong \hat{g}(x_{k+1}).$
    $M_{k+1} = \sqrt{M_k^2 + \frac{a_{k+1}^2}{D^2} \| g_{x_{k+1}} - g_{y_k} \|_*^2}.$
```



Convergence rate: $\mathbb{E}[F(x_k)] - F^* \le O\left(\frac{L_f D^2}{k^2} + \frac{\sigma D}{\sqrt{k}} + k\delta_f \right).$

Example: Hölder Smooth Functions

Suppose f is (ν, H_ν)-Hölder smooth. Then, f as approximately smooth with $(\bar{f}, \bar{g}) = (f, \nabla f)$, arbitrary $\delta_f > 0$ and $L_f \sim [\frac{1}{\delta_f}]^{\frac{1-\nu}{1+\nu}} H_\nu^{\frac{2}{1+\nu}}$.

For UniSgd, we get, for $F_k = \mathbb{E}[F(\bar{x}_k)] - F^*$,

$$F_k \lesssim \frac{L_f D^2}{k} + \frac{\sigma D}{\sqrt{k}} + \delta_f \sim \frac{H_\nu^{\frac{2}{1+\nu}} D^2}{k \delta_f^{\frac{1-\nu}{1+\nu}}} + \frac{\sigma D}{\sqrt{k}} + \delta_f.$$

Minimizing this expression in δ_f, we get

$$F_k \leq O\biggl(\frac{H_\nu D^{1+\nu}}{k^{\frac{1+\nu}{2}}} + \frac{\sigma D}{\sqrt{k}} \biggr) \leq \epsilon \quad \text{in} \quad O\biggl(\biggl[\frac{H_\nu D^{1+\nu}}{\epsilon} \biggr]^{\frac{2}{1+\nu}} + \frac{\sigma^2 D^2}{\epsilon^2} \biggr) \text{ orac. calls.}$$

Similar reasoning for UniFastSgd gives, for $F_k = \mathbb{E}[F(x_k)] - F^*$,

$$F_k \leq O\biggl(\frac{H_\nu D^{1+\nu}}{k^{\frac{1+3\nu}{2}}} + \frac{\sigma D}{\sqrt{k}} \biggr) \leq \epsilon \quad \text{in} \quad O\biggl(\biggl[\frac{H_\nu D^{1+\nu}}{\epsilon} \biggr]^{\frac{2}{1+3\nu}} + \frac{\sigma^2 D^2}{\epsilon^2} \biggr) \text{ orac. calls.}$$

Implicit Variance Reduction

Problem Formulation

Problem: $F^* = \min_{x \in \operatorname{dom} \psi} [F(x) = f(x) + \psi(x)]$.

Assumptions:

1. f is (δ_f, L_f)-approximately smooth with components $(\bar{f}, \bar{g})$.
2. Bounded domain: $\|x - y\| \leq D$, $\forall x, y \in \operatorname{dom} \psi$.
3. f can be accessed only via unbiased stochastic oracle $\hat{g}$ for $\bar{g}$: $\mathbb{E}_\xi[g(x, \xi)] = \bar{g}(x)$.

Goal: Express complexity bounds in terms of $\sigma_*^2 \coloneqq \mathrm{Var}_{\hat{g}}(x^*)$ instead of σ^2.

New assumption on variance

The variance $\mathrm{Var}_{\hat{g}}(x, y) \coloneqq \mathbb{E}_\xi[\|[g(x, \xi) - g(y, \xi)] - [\bar{g}(x) - \bar{g}(y)]\|_*^2]$ is approximately smooth w.r.t. f:

$$\mathrm{Var}_{\hat{g}}(x, y) \leq 2L_{\hat{g}}[\beta_{f, \bar{f}, \bar{g}}(x, y) + \delta_{\hat{g}}].$$

Approximate Smoothness of Variance

Condition: $\mathrm{Var}_{\hat{g}}(x, y) \le 2L_{\hat{g}}[\beta_{f,\bar{f},\bar{g}}(x, y) + \delta_{\hat{g}}]$, where $\mathrm{Var}_{\hat{g}}(x, y) := \mathbb{E}_\xi[\|[g(x, \xi) - g(y, \xi)] - [\bar{g}(x) - \bar{g}(y)]\|_*^2]$.

Note:

- $\mathrm{Var}_{\hat{g}}(x, y)$ is the usual variance of $g(x, \xi) - g(y, \xi)$.
- If $\hat{g}_b$ is the mini-batch version of $\hat{g}$ of size b, then $\mathrm{Var}_{\hat{g}_b}(x, y) = \frac{1}{b}\mathrm{Var}_{\hat{g}}(x, y)$, and hence $L_{\hat{g}_b} = \frac{1}{b}L_{\hat{g}}$, $\delta_{\hat{g}_b} = \delta_{\hat{g}}$.

Main example: $f(x) = \mathbb{E}_\xi[f_\xi(x)]$, where each f_ξ is convex and (δ_ξ, L_ξ)-approx. smooth with components $(\bar{f}_\xi, \bar{g}_\xi)$. Then, $g(x, \xi) = \bar{g}_\xi(x)$ satisfies the variance condition with $\bar{f}(x) = \mathbb{E}_\xi[\bar{f}_\xi(x)]$, $\bar{g}(x) = \mathbb{E}_\xi[\bar{g}_\xi(x)]$, and $L_{\hat{g}} = L_{\max}$, $\delta_{\hat{g}} = \frac{1}{L_{\max}}\mathbb{E}_\xi[L_\xi \delta_\xi]$ $(\le \mathbb{E}_\xi[\delta_\xi])$, where $L_{\max} := \sup_\xi L_\xi$.

Explanation:

$$\begin{aligned}\mathrm{Var}_{\hat{g}}(x, y) &\le \mathbb{E}_\xi[\|\bar{g}_\xi(x) - \bar{g}_\xi(y)\|_*^2] \le \mathbb{E}_\xi\left[2L_\xi\left(\beta_{f_\xi,\bar{f}_\xi,\bar{g}_\xi}(x, y) + \delta_\xi\right)\right] \\ &\le 2L_{\max}\left(\mathbb{E}_\xi[\beta_{f_\xi,\bar{f}_\xi,\bar{g}_\xi}(x, y)] + \delta_{\hat{g}}\right) = 2L_{\max}[\beta_{f,\bar{f},\bar{g}}(x, y) + \delta_{\hat{g}}].\end{aligned}$$

Efficiency Bounds

NB: Consider the same methods as before (no modifications).

UniSgd: $O\left(\frac{(L_f + L_{\hat{g}})D^2}{k} + \frac{\sigma_* D}{\sqrt{k}} + \delta_f + \delta_{\hat{g}}\right).$

- When $\delta_f = \delta_{\hat{g}} = 0$, we recover the well-known rates for SGD with predefined stepsizes based on the knowledge of all the constants.

UniFastSgd: $O\left(\frac{L_f D^2}{k^2} + \frac{L_{\hat{g}} D^2}{k} + \frac{\sigma_* D}{\sqrt{k}} + k\delta_f + \delta_{\hat{g}}\right).$

- Different rates for L_f and $L_{\hat{g}}$ terms are unavoidable [Woodworth and Srebro 2021].
- For the special case $\delta_f = \delta_{\hat{g}} = 0$, similar results were obtained in [Woodworth and Srebro 2021; Ilandarideva et al. 2023] assuming that all constants are known.

Example: Problem with Hölder Smooth Components

Problem: $f(x) = \mathbb{E}_\xi[f_\xi(x)]$ with convex and $(\nu, H_\xi(\nu))$-Hölder smooth f_ξ.

Standard mini-batch oracle: $g_b(x, \xi_{[b]}) = \frac{1}{b} \sum_{j=1}^{b} \nabla f_{\xi_j}(x)$.

Method	Stochastic-Oracle (SO) Complexity
UniSgd	$\bigl(\frac{H_f(\nu) D^{1+\nu}}{\epsilon}\bigr)^{\frac{2}{1+\nu}} + \frac{1}{b} \min\bigl\{ \frac{\sigma^2 D^2}{\epsilon^2}, \bigl(\frac{H_{\max}(\nu)}{\epsilon}\bigr)^{\frac{2}{1+\nu}} D^2 + \frac{\sigma_*^2 D^2}{\epsilon^2} \bigr\}$
UniFastSgd	$\bigl(\frac{H_f(\nu) D^{1+\nu}}{\epsilon}\bigr)^{\frac{2}{1+3\nu}} + \frac{1}{b} \min\bigl\{ \frac{\sigma^2 D^2}{\epsilon^2}, \bigl(\frac{H_{\max}(\nu)}{\epsilon}\bigr)^{\frac{2}{1+\nu}} D^2 + \frac{\sigma_*^2 D^2}{\epsilon^2} \bigr\}$

Notation: $\sigma^2 = \sup_{x \in \operatorname{dom} \psi} \operatorname{Var}_{\hat{g}_1}(x) \equiv \sup_{x \in \operatorname{dom} \psi} \mathbb{E}_\xi[\|\nabla f_\xi(x) - \nabla f(x)\|_*^2]$, $\sigma_*^2 = \operatorname{Var}_{\hat{g}_1}(x^*) \equiv \mathbb{E}_\xi[\|\nabla f_\xi(x^*) - \nabla f(x^*)\|_*^2]$, $H_f(\nu)$ is the Hölder constant of degree ν for f.

Explicit Variance Reduction with SVRG

Universal SVRG

SVRG Oracle: $G(x,\xi) = g(x,\xi) - g(\tilde{x},\xi) + \bar{g}(\tilde{x})$.

Algorithm 3 $\mathrm{UniSvrg}_{\hat{g},\bar{g},\psi}(x_0; D)$

```
$\tilde{x}_0 = x_0,\ M_0 = 0.$
for $t = 0, 1, \dots$ do
    $\hat{G}_t = \mathrm{SvrgOrac}_{\hat{g},\bar{g}}(\tilde{x}_t)$.
    $(\tilde{x}_{t+1}, x_{t+1}, M_{t+1}) \cong \mathrm{UniSgd}_{\hat{G}_t,\psi}(x_t, M_t, 2^{t+1}; D)$.
```

Algorithm 4 $\mathrm{UniSgd}_{\hat{g},\psi}(x_0, M_0, N; D)$

```
$g_0 \cong \hat{g}(x_0)$.
for $k = 0, \dots, N-1$ do
    $x_{k+1} = \mathrm{Prox}_{\psi}(x_k, g_k, M_k)$,  $g_{k+1} \cong \hat{g}(x_{k+1})$.
    $M_{k+1} = \sqrt{M_k^2 + \frac{1}{D^2}\|g_{k+1} - g_k\|_*^2}$.
return $(\bar{x}_N, x_N, M_N)$, where $\bar{x}_N := \frac{1}{N}\sum_{i=1}^N x_i$.
```

Universal Fast SVRG

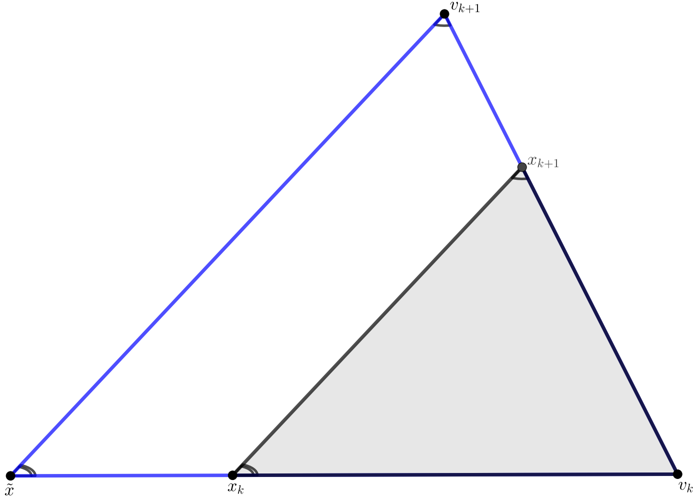
Algorithm 5 $\mathrm{UniFastSvrg}_{\hat{g},\bar{g},\psi}(x_0, N; D)$

```
$\tilde{x}_0 = \mathrm{argmin}_x\{\langle \bar{g}(x_0), x\rangle + \psi(x)\},\ v_0 = x_0,\ M_0 = 0,\ A_0 = \frac{1}{N}$.
for $t = 0, 1, \dots$ do
    $a_{t+1} = \sqrt{A_t},\ \ A_{t+1} = A_t + a_{t+1}$.
    $(\tilde{x}_{t+1}, v_{t+1}, M_{t+1}) \cong \mathrm{UniTriSvrgEpoch}_{\hat{g},\bar{g},\psi}(\tilde{x}_t, v_t, M_t, A_t, a_{t+1}, N; D)$.
```

Algorithm 6 $\mathrm{UniTriSvrgEpoch}_{\hat{g},\bar{g},\psi}(\tilde{x}, v_0, M_0, A, a, N; D)$

```
$A_+ = A + a,\ x_0 = \frac{A}{A_+}\tilde{x} + \frac{a}{A_+}v_0,\ \hat{G} = \mathrm{SvrgOrac}_{\hat{g},\bar{g}}(\tilde{x}),\ G_{x_0} \cong \hat{G}(x_0)$.
for $k = 0, \dots, N-1$ do
    $v_{k+1} = \mathrm{Prox}_\psi(v_k, G_{x_k}, M_k/a)$.
    $x_{k+1} = \frac{A}{A_+}\tilde{x} + \frac{a}{A_+}v_{k+1},\ \ G_{x_{k+1}} \cong \hat{G}(x_{k+1})$.
    $M_{k+1} = \sqrt{M_k^2 + \frac{a^2}{D^2}\|G_{x_{k+1}} - G_{x_k}\|_*^2}$.
return $(\bar{x}_N, v_N, M_N)$, where $\bar{x}_N := \frac{1}{N}\sum_{k=1}^N x_k$.
```

Geometry of UniTriSvrgEpoch



Efficiency Guarantees

Method	Convergence rate	SO complexity
UniSgd	$\frac{L_f D^2}{k} + \delta_f + \min\left\{\frac{\sigma D}{\sqrt{k}}, \frac{\sigma_* D}{\sqrt{k}} + \frac{L_{\hat{g}} D^2}{k} + \delta_{\hat{g}}\right\}$	k
UniFastSgd	$\frac{L_f D^2}{k^2} + k\delta_f + \min\left\{\frac{\sigma D}{\sqrt{k}}, \frac{\sigma_* D}{\sqrt{k}} + \frac{L_{\hat{g}} D^2}{k} + \delta_{\hat{g}}\right\}$	k
UniSvrg	$\frac{(L_f + L_{\hat{g}}) D^2}{2^t} + \delta_f + \delta_{\hat{g}}$	$2^t + n \log t$
UniFastSvrg	$\frac{(L_f + L_{\hat{g}}) D^2}{n(t - \log\log n)^2} + t(\delta_f + \delta_{\hat{g}})$	nt

Note: Assuming that querying $\bar{g}$ is n times more expensive than $\hat{g}$.

Example: Problem with Hölder Smooth Components

Problem: $f(x) = \mathbb{E}_\xi[f_\xi(x)]$ with convex and $(\nu, H_\xi(\nu))$-Hölder smooth f_ξ.

Standard mini-batch oracle: $g_b(x, \xi_{[b]}) = \frac{1}{b} \sum_{j=1}^{b} \nabla f_{\xi_j}(x)$.

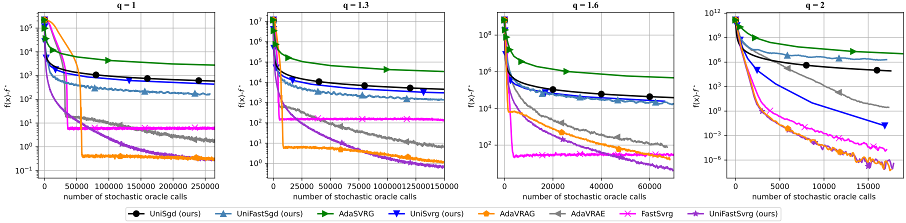
Method	Stochastic-Oracle (SO) Complexity
UniSgd	$\left(\frac{H_f(\nu) D^{1+\nu}}{\epsilon}\right)^{\frac{2}{1+\nu}} + \frac{1}{b} \min\left\{\frac{\sigma^2 D^2}{\epsilon^2}, \left(\frac{H_{\max}(\nu)}{\epsilon}\right)^{\frac{2}{1+\nu}} D^2 + \frac{\sigma_*^2 D^2}{\epsilon^2}\right\}$
UniFastSgd	$\left(\frac{H_f(\nu) D^{1+\nu}}{\epsilon}\right)^{\frac{2}{1+3\nu}} + \frac{1}{b} \min\left\{\frac{\sigma^2 D^2}{\epsilon^2}, \left(\frac{H_{\max}(\nu)}{\epsilon}\right)^{\frac{2}{1+\nu}} D^2 + \frac{\sigma_*^2 D^2}{\epsilon^2}\right\}$
UniSvrg	$\left[N_\nu(\epsilon) := \left(\frac{H_f(\nu) D^{1+\nu}}{\epsilon}\right)^{\frac{2}{1+\nu}} + \frac{1}{b}\left(\frac{H_{\max}(\nu)}{\epsilon}\right)^{\frac{2}{1+\nu}} D^2\right] + n_b \log_+ N_\nu(\epsilon)$
UniFastSvrg	$\left[\frac{n_b^\nu H_f(\nu) D^{1+\nu}}{\epsilon}\right]^{\frac{2}{1+3\nu}} + \left[\frac{n_b^\nu H_{\max}(\nu) D^{1+\nu}}{b^{(1+\nu)/2} \epsilon}\right]^{\frac{2}{1+3\nu}} + n_b \log \log n_b$

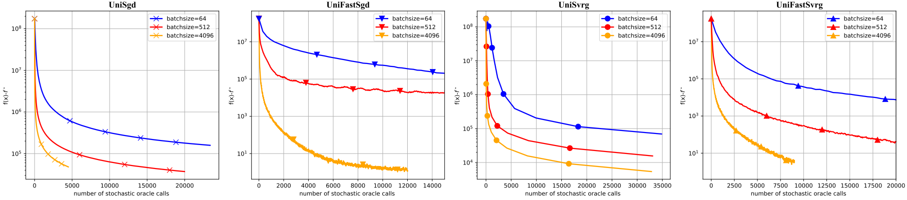
Note: Assuming that querying $\bar{g}$ is n_b times more expensive than $\hat{g}_b$.

Experiments & Conclusions

Experiments

Polyhderon feasibility problem: $\min_{\|x\| \le R} \left\{ f(x) := \frac{1}{n} \sum_{i=1}^{n} [\langle a_i, x \rangle - b_i]_+^q \right\}.$





Conclusions

- We showed that AdaGrad stepsizes can be applied, in a unified manner, in a large variety of situations, leading to universal methods suitable for multiple problem classes at the same time.
- The corresponding methods only need to know diameter D of feasible set, and automatically adapt to the best possible problem class described by various smoothness and variance assumptions.
- The universality is not for free: we need to know a good estimate of D. Adaptation to D could be addressed using the recently developed techniques from parameter-free methods. This is an important direction for future work.

Paper

Universality of AdaGrad Stepsizes for Stochastic Optimization: Inexact Oracle, Acceleration and Variance Reduction

arXiv:2406.06398

Thank you!

References I

- Y. Carmon and O. Hinder. Making SGD Parameter-Free. In **Proceedings of Thirty Fifth Conference on Learning Theory**, volume 178, pages 2360–2389, 2022.
- A. Cutkosky and K. A. Boahen. Online Learning Without Prior Information. In **Annual Conference Computational Learning Theory**, 2017.
- A. Cutkosky and F. Orabona. Black-Box Reductions for Parameter-free Online Learning in Banach Spaces. In **Annual Conference Computational Learning Theory**, 2018.
- A. Defazio and K. Mishchenko. Learning-Rate-Free Learning by D-Adaptation. In **Proceedings of the 40th International Conference on Machine Learning**, volume 202 of **Proceedings of Machine Learning Research**, pages 7449–7479, 2023.

References II

- O. Devolder, F. Glineur, and Y. Nesterov. First-order methods of smooth convex optimization with inexact oracle. **Mathematical Programming**, 146:37–75, 2013. DOI: 10.1007/s10107-013-0677-5.

- N. Doikov, K. Mishchenko, and Y. Nesterov. Super-universal regularized newton method. **SIAM Journal on Optimization**, 34(1):27–56, 2024.

- N. Doikov and Y. Nesterov. Minimizing uniformly convex functions by cubic regularization of newton method. **Journal of Optimization Theory and Applications**, 189(1):317–339, 2021.

- J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. **Journal of Machine Learning Research**, 12:2121–2159, 2011.

References III

- A. Ene, H. L. Nguyen, and A. Vladu. Adaptive Gradient Methods for Constrained Convex Optimization and Variational Inequalities. In **Thirty-Fifth AAAI Conference on Artificial Intelligence**, pages 7314–7321, 2021.

- G. N. Grapiglia and Y. Nesterov. Regularized Newton methods for minimizing functions with Hölder continuous Hessians. **SIAM Journal on Optimization**, 27(1):478–506, 2017.

- G. N. Grapiglia and Y. Nesterov. Tensor Methods for Minimizing Convex Functions with Hölder Continuous Higher-Order Derivatives. **SIAM Journal on Optimization**, 30(4):2750–2779, 2020.

- S. Ilandarideva, A. Juditsky, G. Lan, and T. Li. Accelerated stochastic approximation with state-dependent noise. **arXiv preprint arXiv:2307.01497**, 2023.

References IV

A. Jacobsen and A. Cutkosky. Unconstrained online learning with unbounded losses. In **Proceedings of the 40th International Conference on Machine Learning**, 2023.

A. Kavis, K. Y. Levy, F. Bach, and V. Cevher. UniXGrad: A Universal, Adaptive Algorithm with Optimal Guarantees for Constrained Optimization. In **Advances in Neural Information Processing Systems 32**, pages 6260–6269. 2019.

K. Y. Levy, A. Yurtsever, and V. Cevher. Online Adaptive Methods, Universality and Acceleration. **Advances in Neural Information Processing Systems**, 31, 2018.

H. B. McMahan and M. Streeter. Adaptive Bound Optimization for Online Convex Optimization. **arXiv preprint arXiv:1002.4908**, 2010.

Y. Nesterov. Universal gradient methods for convex optimization problems. **Math. Program.**, 152:381–404, 2015.

References V

F. Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In **Proceedings of the 27th International Conference on Neural Information Processing Systems**, pages 1116–1124, 2014.

A. Rodomanov, A. Kavis, Y. Wu, K. Antonakopoulos, and V. Cevher. Universal Gradient Methods for Stochastic Convex Optimization. **arXiv preprint arXiv:2402.03210**, 2024.

B. E. Woodworth and N. Srebro. An Even More Optimal Stochastic Optimization Algorithm: Minibatching and Interpolation Learning. **Advances in Neural Information Processing Systems**, 34:7333–7345, 2021.