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)

28 August 2024
ALGOPT 2024 Workshop on Algorithmic Optimization
Louvain-la-Neuve, Belgium

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

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

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

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:

- $\nu = 1$: $\|\nabla f(x) - \nabla f(y)\|_* \le H_1 \|x - y\|$ (Lipschitz gradient).
- $\nu = 0$: $\|\nabla f(x) - \nabla f(y)\|_* \le H_0$ (contains Lipschitz functions). This class is better than $\|\nabla f(x)\|_* \le L_0$.
- If $H_{\nu_1}, H_{\nu_2} < +\infty$ for some $\nu_1 \le \nu_2$, 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: $O\left(\inf_{\nu \in [0,1]} \left(\frac{H_\nu}{\epsilon}\right)^{\frac{2}{1+\nu}} D^2\right)$ iters to $f(x_k^*) - f^* \le \epsilon$.

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

Universal gradient methods for stochastic optimization?

Related Work: AdaGrad Methods

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

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

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:

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

Our work: Fully-Universal AdaGrad methods.

Main Algorithms and Results for Uniformly Bounded Variance

Approximate Smoothness

A function $f\colon \mathbb{R}^d \to \mathbb{R}$ is called approximately smooth if 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 $x, y \in \mathbb{R}^d$,

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

NB: $(\bar{f}, \bar{g})$ is a (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

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

Assumptions:

1. f is (δ_f, L_f)-approximately smooth with components $(\bar{f}, \bar{g})$.
2. f can be accessed only via a stochastic oracle $\hat{g}$ such that $\mathbb{E}_\xi[g(x,\xi)] = \bar{g}(x)$.
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: $\|x - y\| \le D$, $\forall x, y \in \mathrm{dom}\,\psi$.

Note: In general, $\hat{g}$ may be biased: $\mathbb{E}_\xi[g(x,\xi)] = \bar{g}(x) \neq \nabla f(x)$.

Note: Asm. 4 can always be ensured with $D = 2R_0$ whenever we know $R_0 \ge \|x_0 - x^*\|$ by considering $F^* = \min_{x \in \mathrm{dom}\,\psi_D} [F_D(x) = f(x) + \psi_D(x)]$, 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: $\text{Prox}_\psi(x, g, M) = \operatorname*{argmin}_{y \in \text{dom}\,\psi} \{\langle g, y\rangle + \psi(y) + \frac{M}{2}\|y - x\|^2\}$.

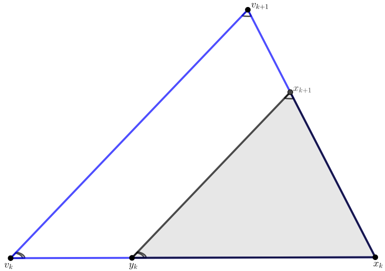
Output point: $\bar{x}_k = \frac{1}{k}\sum_{i=1}^{k} x_i$.

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

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



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

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

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

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

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

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. Stochastic oracle $\hat{g}$: $\mathbb{E}_\xi[g(x, \xi)] = \bar{g}(x)$.

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

Approximate Smoothness of Variance

New assumption on variance

$\mathrm{Var}_{\hat{g}}(x, y) := \mathbb{E}_\xi[\|[g(x,\xi) - g(y,\xi)] - [\bar{g}(x) - \bar{g}(y)]\|_*^2]$ satisfies

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

c.f.: $\|\nabla f(x) - \nabla f(y)\|_*^2 \leq 2L[f(y) - f(x) - \langle \nabla f(x), y - x \rangle]$.

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 $L_{\hat{g}} = L_{\max}$, $\delta_{\hat{g}} = \mathbb{E}_\xi[\delta_\xi]$, where $L_{\max} := \sup_\xi L_\xi$.

Note: 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 $L_{\hat{g}_b} = \frac{1}{b} L_{\hat{g}}$, $\delta_{\hat{g}_b} = \delta_{\hat{g}}$.

Another example: σ^2-bounded variance $\implies$ $L_{\hat{g}} = \frac{2\sigma^2}{\delta_{\hat{g}}}$ for any $\delta_{\hat{g}} > 0$.

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.

Note: When $\hat{g}$ has σ^2-bounded variance, we get

$$\min_{\delta_{\hat{g}}>0}\left[\frac{L_{\hat{g}} D^2}{k} + \delta_{\hat{g}}\right] = \min_{\delta_{\hat{g}}>0}\left[\frac{2\sigma^2 D^2}{k\delta_{\hat{g}}} + \delta_{\hat{g}}\right] = \frac{2\sqrt{2}\sigma D}{\sqrt{k}}.$$

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	$\left(\frac{H_f(\nu) D^{1+\nu}}{\epsilon}\right)^{\frac{2}{1+3\nu}} + \frac{1}{b} \min\left\{ \frac{\sigma^2 D^2}{\epsilon^2}, \left(\frac{H_{\max}(\nu)}{\epsilon}\right)^{\frac{2}{1+\nu}} D^2 + \frac{\sigma_*^2 D^2}{\epsilon^2} \right\}$

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 $t = 0, 1, \ldots$ **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 $k = 0, \ldots, N-1$ **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

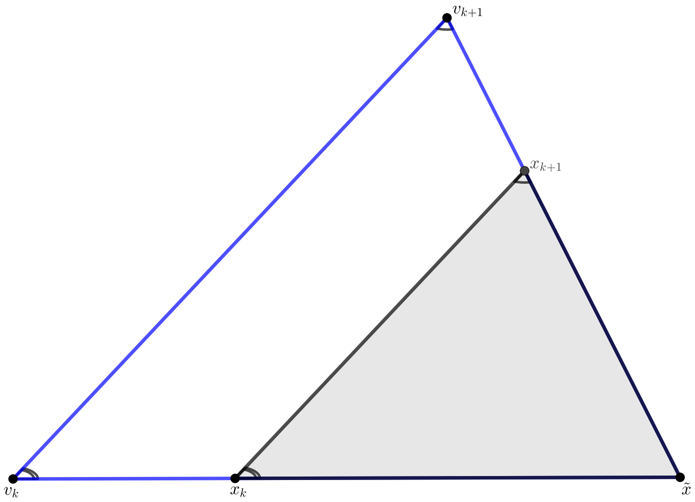
Algorithm 5 $\mathrm{UniFastSvrg}_{\hat{g},\bar{g},\psi}(x_0, N; D)$

$\tilde{x}_0 = \operatorname{argmin}_x\{\langle \bar{g}(x_0), x\rangle + \psi(x)\}$, $v_0 = x_0$, $M_0 = 0$, $A_0 = \frac{1}{N}$.
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	$\frac{L_f D^2}{k} + \delta_f + \min\{\frac{\sigma D}{\sqrt{k}}, \frac{\sigma_* D}{\sqrt{k}} + \frac{L_{\hat{g}} D^2}{k} + \delta_{\hat{g}}\}$	k
UniFastSgd	$\frac{L_f D^2}{k^2} + k\delta_f + \min\{\frac{\sigma D}{\sqrt{k}}, \frac{\sigma_* D}{\sqrt{k}} + \frac{L_{\hat{g}} D^2}{k} + \delta_{\hat{g}}\}$	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: $g_b(x, \xi_{[b]}) = \frac{1}{b} \sum_{j=1}^{b} \nabla f_{\xi_j}(x)$.

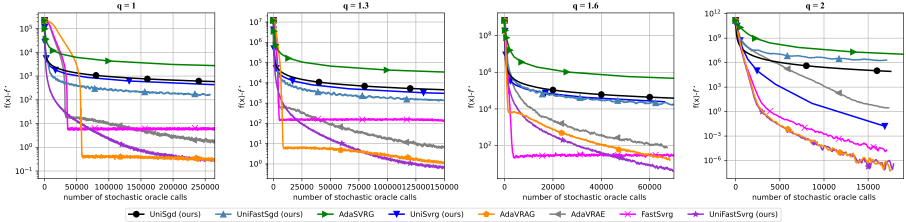
Method	Stochastic-Oracle (SO) Complexity
UniSgd	$\left(\frac{H_f(\nu) D^{1+\nu}}{\epsilon}\right)^{\frac{2}{1+\nu}} + \frac{1}{b} \min\left\{\frac{\sigma^2 D^2}{\epsilon^2}, \left(\frac{H_{\max}(\nu)}{\epsilon}\right)^{\frac{2}{1+\nu}} D^2 + \frac{\sigma_*^2 D^2}{\epsilon^2}\right\}$
UniFastSgd	$\left(\frac{H_f(\nu)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

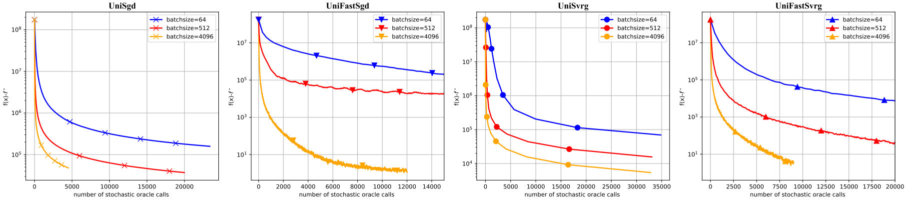
Note: Assuming that querying $\bar{g}$ is n_b times more expensive than $\hat{g}_b$.

Experiments & Conclusions

Experiments

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





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 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 D for $\|x_0 - x^*\|$. Adaptation to D is possible but at the expense of knowing smoothness parameters ("parameter-free" methods).

Paper

Universality of AdaGrad Stepsizes for Stochastic Optimization: Inexact Oracle, Acceleration and Variance Reduction (arXiv:2406.06398)

Thank you!

References I

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.

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.

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

References II

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.

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.