Universal Gradient Methods for Stochastic Convex Optimization

Anton Rodomanov[1] Ali Kavis[2] Yongtao Wu[3] Kimon Antonakopoulos[3] Volkan Cevher[3]

[1]CISPA Helmholtz Center for Information Security, Germany [2]Institute for Foundations of Machine Learning (IFML), UT Austin, USA

[3]Laboratory for Information and Inference Systems (LIONS), EPFL, Switzerland

Paper (PDF)

Problem Formulation

Consider the composite optimization problem:

$$F^* := \min_{x \in \operatorname{dom}\psi} \big[F(x) := f(x) + \psi(x) \big], \qquad \text{(P)}$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ and $\psi\colon \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ are convex, ψ is simple.

Assumptions: ($\|\cdot\|$ is a Euclidean norm, $\nu \in [0,1]$)

1. Hölder smoothness: $\|\nabla f(x) - \nabla f(y)\|_* \le L_\nu \|x - y\|^\nu$, $\forall x, y \in \operatorname{dom}\psi$.
2. Unbiased stochastic oracle: $\mathbb{E}_\xi[g(x,\xi)] = \nabla f(x)$, $\forall x \in \operatorname{dom}\psi$.
3. Bounded variance: $\mathbb{E}_\xi[\|g(x,\xi) - \nabla f(x)\|_*^2] \le \sigma^2$, $\forall x \in \operatorname{dom}\psi$.

Goal: Develop methods that can solve (P) without knowing ν, L_ν and σ.

We do so **assuming additionally** $\operatorname{dom}\psi$ is bounded with known diameter:

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 replacing (P) with $F^* = \min_x [f(x) + \psi_D(x)]$, where $\psi_D = \psi + \operatorname{Ind}_{B_0}$ with $B_0 = \{x : \|x - x_0\| \le R_0\}$.

Classical Universal Gradient Methods (UGMs)

UGM (Nesterov 2015): $x_{k+1} = \operatorname{argmin}_x \{\langle \nabla f(x_k), x\rangle + \psi(x) + \frac{H_k}{2}\|x - x_k\|^2\}$,

where H_k is found by line search to satisfy the following condition:

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

Efficiency bound: $O\big(\inf_{\nu \in [0,1]} [\frac{L_\nu}{\epsilon}]^{\frac{2}{1+\nu}} R_0^2\big)$ iterations to reach $F(x_k^*) - F^* \le \epsilon$,

where $R_0 = \|x_0 - x^*\|$ and x_k^* is the iterate with the smallest value of F.

Accelerated version (Nesterov 2015): $O\big(\inf_{\nu\in[0,1]} [\frac{L_\nu R_0^{1+\nu}}{\epsilon}]^{\frac{2}{1+3\nu}} R_0^2\big)$.

Main problem: UGMs do not work properly with the stochastic oracle.

AdaGrad Methods

Suppose that $\psi = \operatorname{Ind}_Q$ for a simple convex set Q.

AdaGrad (McMahan and Streeter 2010; Duchi et al. 2011): ($g_k = g(x_k, \xi_k)$)

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

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

$$\mathbb{E}[f(\bar{x}_k)] - f^* \le O\bigg(\min\Big\{\frac{M_0 D}{\sqrt{k}}, \frac{L_1 D^2}{k}\Big\} + \frac{\sigma D}{\sqrt{k}}\bigg),$$

where M_0 and L_1 are the Lipschitz constants of f and ∇f, respectively.

UniXGrad (Kavis et al. 2019): Accelerated version of AdaGrad accumulating $\|g_{i+1} - g_i\|_*^2$ instead of $\|g_i\|_*^2$. Convergence rate:

$$O\bigg(\min\Big\{\frac{M_0 D}{\sqrt{k}}, \frac{L_1 D^2}{k^2}\Big\} + \frac{\sigma D}{\sqrt{k}}\bigg).$$

Question: Do AdaGrad methods work for the entire Hölder class?

Basic Method

Algorithm Universal Stochastic Gradient Method (USGM)

Initialize: $x_0 \in \operatorname{dom}\psi$, $D > 0$, $H_0 = 0$, $g_0 = g(x_0, \xi_0)$.

for $k = 0, 1, \ldots$ **do**

$x_{k+1} = \operatorname{argmin}_x \{\langle g_k, x\rangle + \psi(x) + \frac{H_k}{2}\|x - x_k\|^2\}$, $g_{k+1} = g(x_{k+1}, \xi_{k+1})$.

$H_{k+1} = H_k + \frac{[\hat\beta_{k+1} - \frac{1}{2} H_k r_{k+1}^2]_+}{D^2 + \frac{1}{2} r_{k+1}^2}$ with $\begin{cases} r_{k+1} = \|x_{k+1} - x_k\|, \\ \hat\beta_{k+1} = \langle g_{k+1} - g_k, x_{k+1} - x_k\rangle. \end{cases}$

Theorem: For any $k \ge 1$ and $\bar{x}_k = \frac{1}{k}\sum_{i=1}^k x_i$, we have

$$\mathbb{E}[F(\bar{x}_k)] - F^* \le \inf_{\nu \in [0,1]} \frac{8 L_\nu D^{1+\nu}}{k^{\frac{1+\nu}{2}}} + \frac{4\sigma D}{\sqrt{k}}.$$

It suffices to make $O\big(\inf_{\nu\in[0,1]} [\frac{L_\nu}{\epsilon}]^{\frac{2}{1+\nu}} D^2 + \frac{\sigma^2 D^2}{\epsilon^2}\big)$ oracle calls to reach ϵ-accuracy.

Main Idea and Outline of Analysis

- Opt. condition for x_{k+1} gives (for $d_k = \|x_k - x^*\|$, $r_{k+1} = \|x_{k+1} - x_k\|$)

$$f(x_k) + \langle g_k, x_{k+1} - x_k\rangle + \psi(x_{k+1}) + \tfrac{H_k}{2} r_{k+1}^2 + \tfrac{H_k}{2} d_{k+1}^2 \le f(x_k) + \langle g_k, x^* - x_k\rangle + \psi(x^*) + \tfrac{H_k}{2} d_k^2.$$

- Use $\mathbb{E}_{\xi_k}[f(x_k) + \langle g_k, x^* - x_k\rangle] = f(x_k) + \langle \nabla f(x_k), x^* - x_k\rangle \le f(x^*)$ to get

$$\mathbb{E}\big[F_{k+1} + \tfrac{H_k}{2} d_{k+1}^2\big] \le \mathbb{E}\big[\tfrac{H_k}{2} d_k^2 + \beta_{k+1} - \tfrac{H_k}{2} r_{k+1}^2\big],$$

where $F_{k+1} = F(x_{k+1}) - F^*$, $\beta_{k+1} = f(x_{k+1}) - f(x_k) - \langle g_k, x_{k+1} - x_k\rangle$.

- To make d_k-terms telescope, require that $H_k \le H_{k+1}$ and estimate

$$\mathbb{E}\big[F_{k+1} + \tfrac{H_{k+1}}{2} d_{k+1}^2\big] \le \mathbb{E}\big[\tfrac{H_k}{2} d_k^2 + \beta_{k+1} - \tfrac{H_k}{2} r_{k+1}^2 + \tfrac{H_{k+1} - H_k}{2} d_{k+1}^2\big]$$

$$\le \mathbb{E}\big[\tfrac{H_k}{2} d_k^2 + \beta_{k+1} - \tfrac{H_{k+1}}{2} r_{k+1}^2 + (H_{k+1} - H_k) D^2\big].$$

- Main idea: balance the two error terms by choosing H_{k+1} from equation

$$(H_{k+1} - H_k) D^2 = \big[\hat\beta_{k+1} - \tfrac{H_{k+1}}{2} r_{k+1}^2\big]_+, \qquad (*)$$

where $\hat\beta_{k+1}$ is such that $\mathbb{E}[\beta_{k+1}] \le \mathbb{E}[\hat\beta_{k+1}]$ (see Alg. for explicit solution and note that $\beta_{k+1} \le \langle \nabla f(x_{k+1}) - g_k, x_{k+1} - x_k\rangle = \mathbb{E}_{\xi_{k+1}}[\hat\beta_{k+1}]$).

- We thus get $\mathbb{E}[F_{k+1} + \frac{H_{k+1}}{2} d_{k+1}^2] \le \mathbb{E}[\frac{H_k}{2} d_{k+1}^2 + 2(H_{k+1} - H_k) D^2]$, and so

$$\mathbb{E}[F(\bar{x}_k)] - F^* \le \mathbb{E}\big[\tfrac{1}{k}\textstyle\sum_{i=1}^k F_i\big] \le \frac{2\,\mathbb{E}[H_k] D^2}{k}.$$

- To estimate growth rate of H_k, we first estimate

$$\hat\beta_{k+1} \equiv \langle \nabla f(x_{k+1}) - \nabla f(x_k) + \Delta_{k+1}, x_{k+1} - x_k\rangle \le L_\nu r_{k+1}^{1+\nu} + \sigma_{k+1} r_{k+1},$$

where $\Delta_k = \delta_{k+1} - \delta_k$, $\delta_k = g_k - \nabla f(x_k)$, $\sigma_k = \|\Delta_k\|_*$ (note: $\mathbb{E}[\sigma_k^2] \le 2\sigma^2$).

Substituting this into (*) gives the following recurrence:

$$(H_{k+1} - H_k) D^2 \lesssim \frac{(1-\nu) L_\nu^{\frac{2}{1-\nu}}}{H_{k+1}^{\frac{1+\nu}{1-\nu}}} + \frac{\sigma_{k+1}^2}{H_{k+1}}.$$

Its solution is $H_k \le O\big(\frac{L_\nu}{D^{1-\nu}} k^{\frac{1-\nu}{2}} + \frac{1}{D}(\sum_{i=1}^k \sigma_i^2)^{\frac{1}{2}}\big)$, so

$$\mathbb{E}[H_k] \le O\Big(\frac{L_\nu}{D^{1-\nu}} k^{\frac{1-\nu}{2}} + \frac{\sigma}{D}\sqrt{k}\Big).$$

Accelerated Algorithm

Algorithm Universal Stochastic Fast Gradient Method (USFGM)

Initialize: $x_0 = v_0 \in \operatorname{dom}\psi$, $D > 0$, $H_0 = A_0 = 0$.

for $k = 0, 1, \ldots$ **do**

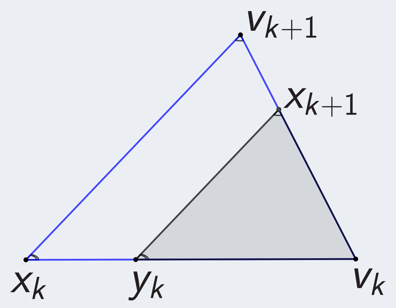
$a_{k+1} = k + 1$, $A_{k+1} = A_k + a_{k+1}$.

$y_k = \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} v_k$, $g_k^y = g(y_k, \xi_k^y)$.

$v_{k+1} = \operatorname{argmin}_x \{a_{k+1}[\langle g_k^y, x\rangle + \psi(x)] + \frac{H_k}{2}\|x - v_k\|^2\}$.

$x_{k+1} = \frac{A_k}{A_{k+1}} x_k + \frac{a_{k+1}}{A_{k+1}} v_{k+1}$, $g_{k+1}^x = g(x_{k+1}, \xi_{k+1}^x)$.

$H_{k+1} = H_k + \frac{[A_{k+1}\hat\beta_{k+1} - \frac{1}{2} H_k r_{k+1}^2]_+}{D^2 + \frac{1}{2} r_{k+1}^2}$ with $\begin{cases} r_{k+1} = \|v_{k+1} - v_k\|, \\ \hat\beta_{k+1} = \langle g_{k+1}^x - g_k^y, x_{k+1} - y_k\rangle. \end{cases}$



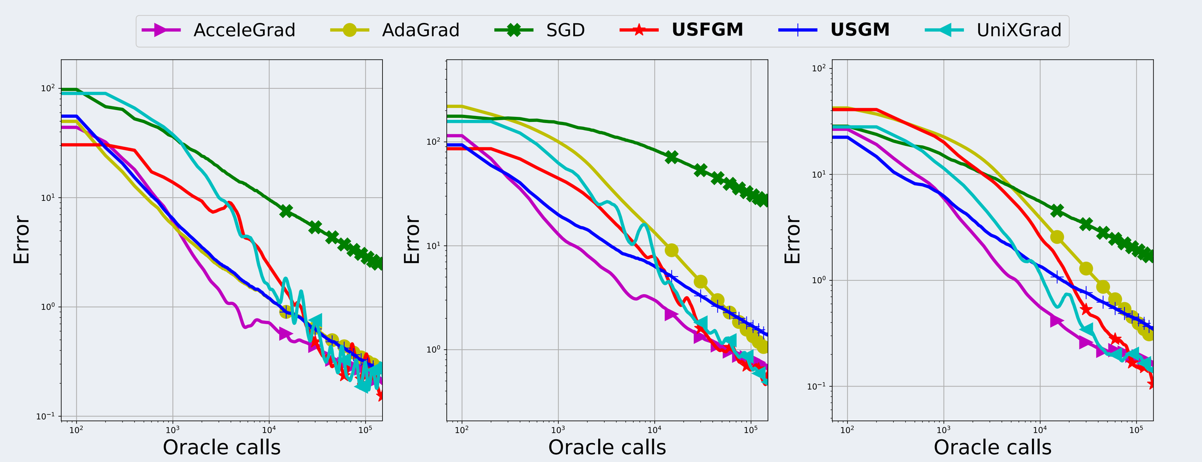
Theorem: For any $k \ge 1$, it holds that

$$\mathbb{E}[F(x_k)] - F^* \le \inf_{\nu\in[0,1]} \frac{32 L_\nu D^{1+\nu}}{k^{\frac{1+3\nu}{2}}} + \frac{8\sigma D}{\sqrt{3k}}.$$

It suffices to make $O\big(\inf_{\nu\in[0,1]} [\frac{L_\nu D^{1+\nu}}{\epsilon}]^{\frac{2}{1+3\nu}} + \frac{\sigma^2 D^2}{\epsilon^2}\big)$ oracle calls to reach ϵ-accuracy.

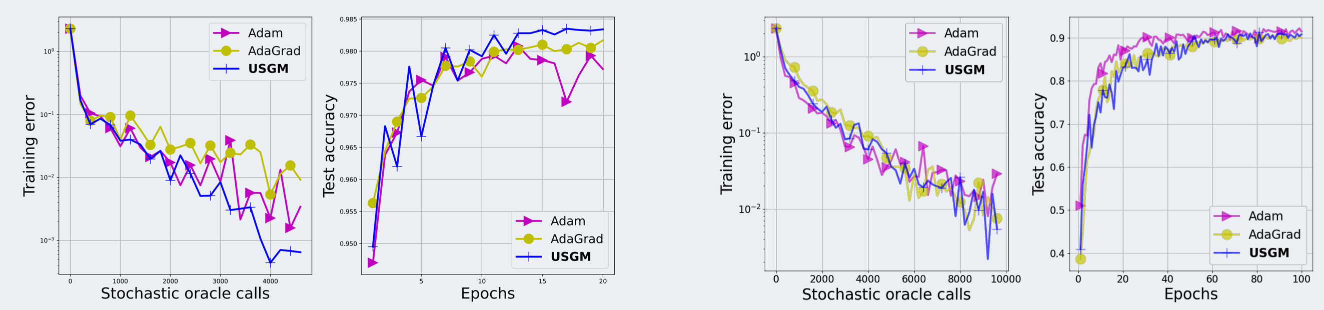
Experiments

Least squares: $\min_{\|x\|\le 1} \frac{1}{2}\|Ax - b\|^2$. Logistic regression: $\min_{\|x\|\le 1} \sum_{i=1}^m \ln(1 + e^{-b_i\langle a_i, x\rangle})$.



(a) Least-square, diabetes. (b) Logistic regression, a1a. (c) Logistic regression, ionosphere.

Neural network training:



3-layer fully connected on MNIST ResNet18 on CIFAR-10

References

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.

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.