UNDERGRAD: A Universal Black-Box Optimization Method With Almost Dimension-Free Convergence Rate Guarantees

Kimon Antonakopoulos
kimon.antonakopoulos@epfl.ch

joint work with

Dong Quan Vu (Safran) Kfir Y. Levy (Technion) Volkan Cevher (EPFL)
Panayotis Mertikopoulos (UGA/Ciriteo AI)

Laboratory for Information and Inference Systems (LIONS)
École Polytechnique Fédérale de Lausanne (EPFL)

lions@epfl swisscom E4S Enterprise for Society EPFL

July 16, 2022

The convex minimization formulation

$$f^\star = \min_{x : x \in \mathcal{X}} f(x) \quad (\text{argmin} \to x^\star)$$

[1] Y. Nesterov, "Introductory lectures on convex optimization: A basic course," Springer Science, 2013.

[2] Lan, Guanghui. First-order and Stochastic Optimization Methods for Machine Learning. Springer Nature, 2020.

The convex minimization formulation

$$f^\star = \min_{x : x \in \mathcal{X}} f(x) \quad (\text{argmin} \to x^\star)$$

- In the sequel,
  - the set $\mathcal{X}$ is convex and compact subset of $\mathbb{R}^d$
  - the objective function f is convex
  - The solution set $\mathcal{S}^\star := \{\mathbf{x}^\star \in \operatorname{dom}(f) \cap \mathcal{X} : f(\mathbf{x}^\star) = f^\star\}$ is non-empty.

[1] Y. Nesterov, "Introductory lectures on convex optimization: A basic course," Springer Science, 2013.
[2] Lan, Guanghui. First-order and Stochastic Optimization Methods for Machine Learning. Springer Nature, 2020.

What are the basic solution methods ?

First-Order Methods

First-order methods: iterative methods using first-order information (gradient queries)

First-Order Methods

First-order methods: iterative methods using first-order information (gradient queries)

The optimizer has access to:

$$\mathbf{x}^k \to g_t = \widetilde{\nabla} f(\mathbf{x}^k) + U_k \tag{SFO}$$

First-Order Methods

First-order methods: iterative methods using first-order information (gradient queries)

The optimizer has access to:

$$\mathbf{x}^k \to g_t = \widetilde{\nabla} f(\mathbf{x}^k) + U_k \tag{SFO}$$

The "noise" term U_t satisfies,

- Zero-Mean:

$$\mathbb{E}[U_k | \mathcal{F}_k] = 0$$

- Finite variance almost surely:

$$\|U_k\|_*^2 \leq \sigma^2 \text{ almost surely}$$

First-Order Methods

First-order methods: iterative methods using first-order information (gradient queries)

The optimizer has access to:

$$\mathbf{x}^k \to g_t = \widetilde{\nabla} f(\mathbf{x}^k) + U_k \tag{SFO}$$

The "noise" term U_t satisfies,

- Zero-Mean:

$$\mathbb{E}[U_k | \mathcal{F}_k] = 0$$

- Finite variance almost surely:

$$\|U_k\|_*^2 \leq \sigma^2 \text{ almost surely}$$

Examples:

- Perfect gradient: $U_k = 0$
- Stochastic gradients: $U_k = \nabla F(X_t; \omega_k) - \nabla f(X_k)$ (minibatch etc..)

What is achievable with first-order methods?

Lipschitz Regularity Conditions

Theoretical guarantees require some degree of regularity

Lipschitz Regularity Conditions

Theoretical guarantees require some degree of regularity

- **Bounded gradients / operators:**

$$\|\nabla f(x)\|_* \leq G$$

Lipschitz Regularity Conditions

Theoretical guarantees require some degree of regularity

- **Bounded gradients / operators:**

$$\|\nabla f(x)\|_* \leq G$$

- **Lipschitz continuity of the gradients / operators:**

$$\|\nabla f(x) - \nabla f(x')\|_* \leq L\|x - x'\|$$

Worst-case iteration complexities of first-order methods[1,2]

$f(x)$	gradient oracle	L-smooth	Stationarity measure	**GD/SGD**	**Accelerated GD/SGD**
Convex	stochastic	yes	$f(x^k) - f^\star =$	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$
Convex	deterministic	yes	$f(x^k) - f^\star =$	$\mathcal{O}\left(\frac{1}{k}\right)$	$\mathcal{O}\left(\frac{1}{k^2}\right)$
Convex	stochastic/deterministic	no	$f(x^k) - f^\star =$	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$	$\mathcal{O}\left(\frac{1}{\sqrt{k}}\right)$

[1] Y. Nesterov, "Introductory lectures on convex optimization: A basic course," Springer Science, 2013.
[2] Y. Carmon, J.C. Duchi, O. Hinder, and A. Sidford, "Lower bounds for finding stationary points I–II." Mathematical Programming, 2019.
[3] D. Davis and D. Drusvyatskiy, "Stochastic model-based minimization of weakly convex functions," SIOPT, 2019.
[4] S. Ghadimi and G. Lan, "Accelerated gradient methods for nonconvex nonlinear and stochastic programming," MathProg, 2016.
[5] J. Zhang, et al., "On complexity of finding stationary points of nonsmooth nonconvex functions," arXiv:2002.04130, 2020.
[6] O. Shamir, "Can We Find Near-Approximately-Stationary Points of Nonsmooth Nonconvex Functions?" arXiv:2002.11962, 2020.
[7] V. Cevher and B.C. Vu, "On the linear convergence of the stochastic gradient method with constant step-size," Optimization Letters, 2019.

Can we achieve optimal performance with a single method?

Universal methods

Universal methods: Achieve optimal rates without knowing the regularity in advance

Universal methods

Universal methods: Achieve optimal rates without knowing the regularity in advance

First approach:

- Nesterov [4]

Universal methods

Universal methods: Achieve optimal rates without knowing the regularity in advance

First approach:

- Nesterov [4]
  - Uses line-search

Universal methods

Universal methods: Achieve optimal rates without knowing the regularity in advance

First approach:

- Nesterov [4]
  - Uses line-search
  - not appropriate for noise adaptivity

Universal methods

Universal methods: Achieve optimal rates without knowing the regularity in advance

First approach:

- Nesterov [4]
  - Uses line-search
  - not appropriate for noise adaptivity
- [UniXGrad] [3]
  - updates its step-size policy "on the fly"

Universal methods

Universal methods: Achieve optimal rates without knowing the regularity in advance

First approach:

- Nesterov [4]
  - Uses line-search
  - not appropriate for noise adaptivity
- [UniXGrad] [3]
  - updates its step-size policy "on the fly"
  - adaptive to noise

SOTA: The UniXGrad Method

◦ For brevity, let us denote stochastic gradient of f at x as $\widetilde{\nabla} f(x)$.

UniXGrad [3]

1. Choose $\mathbf{x}^0 \in \mathcal{X}$ arbitrarily as a starting point. Set $\alpha_k = k$.
2. For $k = 1, 2, \ldots,$ iterate

$$\begin{cases} \mathbf{x}^{k+1/2} & = \arg\min_{\mathbf{x}\in\mathcal{X}} \alpha_k \langle \widetilde{\nabla} f(\tilde{\mathbf{x}}^k), \mathbf{x}\rangle + \frac{1}{\gamma_k} D_h(\mathbf{x}, \mathbf{x}^k) \\ \mathbf{x}^{k+1} & = \arg\min_{\mathbf{x}\in\mathcal{X}} \alpha_k \langle \widetilde{\nabla} f(\bar{\mathbf{x}}^{k+1/2}), \mathbf{x}\rangle + \frac{1}{\gamma_k} D_h(\mathbf{x}, \mathbf{x}^k) \\ \gamma_{k+1} & = \frac{D}{\sqrt{G^2 + \sum_{s=1}^{k} \alpha_s^2 \|\widetilde{\nabla} f(\bar{\mathbf{x}}^{s+1/2}) - \widetilde{\nabla} f(\tilde{\mathbf{x}}^s)\|^2}} \end{cases}$$

3. Output $\bar{\mathbf{x}}^{k+1/2}$

◦ Bregman divergence w.r.t 1-strongly convex function h: $D_h(\mathbf{x}, \mathbf{y}) = h(\mathbf{x}) - h(\mathbf{y}) - \langle \nabla h(\mathbf{y}), \mathbf{x} - \mathbf{y}\rangle$

◦ Optimistic average iterates [5, 2]:

$$\bar{\mathbf{x}}^{k+1/2} = \frac{\alpha_k \mathbf{x}^{k+1/2} + \sum_{s=1}^{k-1} \alpha_s \mathbf{x}^{s+1/2}}{\sum_{s=1}^{k} \alpha_s} \qquad \tilde{\mathbf{x}}^k = \frac{\alpha_k \mathbf{x}^{k} + \sum_{s=1}^{k-1} \alpha_s \mathbf{x}^{s+1/2}}{\sum_{s=1}^{k} \alpha_s}$$

Convergence rates of UniXGrad And The Curse of Dimensionality

Oracle	$f(\cdot)$	Assumptions	Convergence rate
Stochastic	L-smooth	$\mathbb{E}\left[\widetilde{\nabla} f(\mathbf{x}) \vert \mathbf{x}\right] = \nabla f(\mathbf{x})$ $\mathbb{E}\left[\Vert\widetilde{\nabla} f(\mathbf{x}) - \nabla f(\mathbf{x})\Vert^2 \vert \mathbf{x}\right] \leq \sigma^2$	$\mathcal{O}\left(\frac{\sigma D}{\sqrt{k}} + \frac{L D^2}{k^2}\right)$
Stochastic	non-smooth	$\mathbb{E}\left[\widetilde{\nabla} f(\mathbf{x}) \vert \mathbf{x}\right] = \nabla f(\mathbf{x})$ $\Vert\widetilde{\nabla} f(\mathbf{x})\Vert^2 \leq G^2$	$\mathcal{O}\left(\frac{G D}{\sqrt{k}}\right)$

Main drawback These convergence rates crucially rely on the uniform boundedness of the underlying Bregman divergence!

Why do we care?

Dimension scalability of universal methods

- *Bounded Bregman diameter does not provide almost-dimension free rates under favourable geometry !*
  - Simplex/trace constrained convex/semidefinite Programs
  - Combinatorial bandits

Dimension scalability of universal methods

- *Bounded Bregman diameter does not provide almost-dimension free rates under favourable geometry !*
  - Simplex/trace constrained convex/semidefinite Programs
  - Combinatorial bandits

	Domain ($\mathcal{X}$)	Bregman Diameter (B_h)	Range ($R_h = h(x) - \min h$)	Shape (χ)	Rate ($L = \infty$)	Rate ($L < \infty$, $\sigma = 0$)
Euclidean	any below	$\mathcal{O}(1)$	$\mathcal{O}(1)$	$\sqrt{d}$	$\mathcal{O}(\sqrt{d/T})$	$\mathcal{O}(d/T)$
Entropic	simplex	∞	$\log d$	1	$\mathcal{O}(\sqrt{\log d/T})$	$\mathcal{O}(\log d/T)$
Von Neumann	spectrahedron	∞	$\log d$	1	$\mathcal{O}(\sqrt{\log d/T})$	$\mathcal{O}(\log d/T)$
ComBand	$\mathrm{conv}(\mathcal{A})$	∞	$\mathcal{O}(\log d)$	1	$\mathcal{O}(\sqrt{\log d/T})$	$\mathcal{O}(\log d/T)$

Can we achieve almost dimension-free rates & order-optimal dependence ?

The UnderGrad method

Dual Extrapolation

- Leading state via a prox-step (primal update)

$$\mathbf{x}^{k+1/2} = P_{\mathbf{x}^t}(-\gamma_k \widetilde{\nabla} f(\mathbf{x}^k))$$

with $P_{\mathbf{x}}(\mathbf{y}) = \arg\min_{\mathbf{x}' \in \mathcal{X}} \{\langle \mathbf{y}, \mathbf{x} - \mathbf{x}' \rangle + D(\mathbf{x}', \mathbf{x})\}$

- Oracle's feedback at $\mathbf{x}^{t+1/2}$, $\widetilde{\nabla} f(\mathbf{x}^{t+1/2})$
- Gradients aggregation (dual update)

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \widetilde{\nabla} f(\mathbf{x}^{t+1/2})$$

- Update via a mirror step (primal-dual update)

$$\mathbf{x}^{k+1} = \mathcal{Q}(\gamma_{k+1} \mathbf{y}^{k+1})$$

with $\mathcal{Q}(v) = \arg\max_{\mathbf{x} \in \mathcal{X}} \{\langle v, \mathbf{x} \rangle - h(\mathbf{x})\}$

UnderGrad [1]

- Leading state via a prox-step (primal update)

$$\mathbf{y}^{k+1/2} = \mathbf{y}^k - \alpha^k \widetilde{\nabla} f(\bar{\mathbf{x}}^k)$$
$$\mathbf{x}^{k+1/2} = \mathcal{Q}(\gamma_k \mathbf{y}^{k+1/2})$$

- Oracle's feedback at $\overline{\mathbf{x}}^{t+1/2}$, i.e., $\widetilde{\nabla} f(\overline{\mathbf{x}}^{k+1/2})$
- Gradients aggregation (dual update)

$$\mathbf{y}^{k+1} = \mathbf{y}^k - \alpha^k \widetilde{\nabla} f(\bar{\mathbf{x}}^{k+1/2})$$

- Update via a mirror step (primal-dual update)

$$\mathbf{x}^{k+1} = \mathcal{Q}(\gamma_{k+1} \mathbf{y}^{k+1})$$

Averaged iterates:

$$\overline{\mathbf{x}}^k = \frac{\alpha^k \mathbf{x}^k + \sum_{j=1}^{k-1} \alpha_j^2 \mathbf{x}^{k+1/2}}{\sum_{j=1}^{k} \alpha^k} \qquad \overline{\mathbf{x}}^{k+1/2} = \frac{\alpha^k \mathbf{x}^{k+1/2} + \sum_{j=1}^{k-1} \alpha_j^2 \mathbf{x}^{k+1/2}}{\sum_{j=1}^{k} \alpha^k}$$

Almost-dimension free convergence guarantees

Theorem

Assume that $\mathbf{x}^k, k = 1/2, 1\ldots$ *are the UnderGrad run with:*

$$\gamma_k = \frac{b}{\sqrt{\alpha^2 + \sum_{s=1}^{k-1} \alpha_s^2 \|\widetilde{\nabla} f(\bar{\mathbf{x}}^{s+1/2}) - \widetilde{\nabla} f(\bar{\mathbf{x}}^s)\|_*^2}}$$

with $\alpha_k = k$, $\alpha = \sqrt{K_h}$, $b = C_h\sqrt{K_h}$, $C_h = \sqrt{R_h + K_h\|\mathcal{X}\|^2}$, $R_h = \max h - \min h$ *and* $\overline{\mathbf{x}}_{k+1/2} = (\sum_{s=1}^{T} \alpha_s)^{-1} \sum_{s=1}^{k} \alpha_s \mathbf{x}^{s+1/2}$, *then the following hold:*

a) *If f has bounded gradients, then*

$$\mathbb{E}[f(\overline{\mathbf{x}}_{k+1/2}) - \min f] \leq 2C_h \sqrt{\frac{K_h + 8(G^2 + \sigma^2)}{K_h k}} \tag{1a}$$

b) *If f has Lipschitz continuous gradient, then*

$$\mathbb{E}[f(\overline{\mathbf{x}}_{k+1/2}) - \min f] \leq \frac{32\sqrt{2}C_h^2 L}{K_h k^2} + \frac{8\sqrt{2}C_h \sigma}{\sqrt{K_h} k} \tag{1b}$$

Conclusions

Design first order methods which exhibit optimal performance both in iterations and in dimensional dependence for convex programming!

Conclusions

Design first order methods which exhibit optimal performance both in iterations and in dimensional dependence for convex programming!

- What about higher order methods?

Conclusions

Design first order methods which exhibit optimal performance both in iterations and in dimensional dependence for convex programming!

- What about higher order methods?
- What about VI's /min-max?

Conclusions

Design first order methods which exhibit optimal performance both in iterations and in dimensional dependence for convex programming!

- What about higher order methods?
- What about VI's /min-max?

Thank you!

References I

[1] K. Antonakopoulos, D.-Q. Vu, V. Cevher, K.-Y Levy, and P. Mertikopoulos.
Undergrad: A universal black-box optimization method with almost dimension-free convergence rate guarantees.
In *International Conference in Machine Learning*, 2022.
(Cited on page 27.)

[2] Jelena Diakonikolas and Lorenzo Orecchia.
Accelerated extra-gradient descent: A novel accelerated first-order method.
In *ITCS*, 2018.
(Cited on page 21.)

[3] Ali Kavis, Kfir Y. Levy, Francis Bach, and Volkan Cevher.
Unixgrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization.
In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 6260–6269. Curran Associates, Inc., 2019.
(Cited on pages 15, 16, 17, 18, 19, 20, and 21.)

[4] Yurii Nesterov.
Universal gradient methods for convex optimization problems.
Mathematical Programming, 152(1-2):381–404, 2015.
(Cited on pages 15, 16, 17, 18, 19, and 20.)

References II

[5] Jun-Kun Wang and Jacob D Abernethy.
Acceleration through optimistic no-regret dynamics.
In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 3824–3834. Curran Associates, Inc., 2018.
(Cited on page 21.)