Universal Gradient Methods for Stochastic Convex Optimization

Anton Rodomanov (CISPA)

15 April 2024
Seminar at CORE,
Louvain-la-Neuve, Belgium

Part I: Motivation

Stochastic Convex Optimization

Problem:

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

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

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

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

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

E.g.: $f(x) = \frac{1}{m} \sum_{i=1}^m f_i(x) \implies g(x, \xi) = \nabla f_\xi(x)$, $\xi \sim \mathrm{Unif}(\{1, \ldots, m\})$.

Stochastic Gradient Method (SGD)

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

Stochastic Gradient Method (SGD):

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

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

Nonsmooth optimization: $\|\nabla f(x)\| \le M$, $\forall x \in Q$.

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

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

Smooth optimization: $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$, $\forall x, y \in Q$.

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

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

Note:

- $\nu = 1$: $\|\nabla f(x) - \nabla f(y)\| \le L_1 \|x - y\|$ (Lipschitz gradient).
- $\nu = 0$: $\|\nabla f(x) - \nabla f(y)\| \le L_0$ (contains Lipschitz functions). This class is better than $\|\nabla f(x)\| \le M$.
- If $L_{\nu_1}, L_{\nu_2} < +\infty$ for some $\nu_1 \le \nu_2$, then $L_\nu < +\infty, \forall \nu \in [\nu_1, \nu_2]$.

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

Universal Gradient Methods – II

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

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

Efficiency bound: $O\left(\inf_{\nu \in [0,1]} \left(\frac{L_\nu}{\epsilon}\right)^{2/(1+\nu)} D^2\right)$ iters to $f(x_k^*) - f^* \leq \epsilon$

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

Great methods but don't work with stochastic oracle!

AdaGrad-type Methods

AdaGrad algorithm [Duchi et al. 2011]: $(g_k \sim \hat{g}(x_k))$

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

Foundation of nowadays popular Adam, RMSProp, ...

Convergence rate: Assuming $\|\nabla f(x)\| \leq M, \ \forall x$, we get

$$\mathbb{E}[f(\bar{x}_k)] - f^* \leq \frac{(M + \sigma)D}{\sqrt{k}},$$

where σ is the variance of gradient oracle.

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

$$\mathbb{E}[f(x_k)] - f^* \leq O\left(\min\left\{\frac{MD}{\sqrt{k}}, \frac{LD^2}{k^2}\right\} + \frac{\sigma D}{\sqrt{k}}\right).$$

(M and L are Lipschitz constants for f and ∇f.)

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 [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).
- Most recent work [Li and Lan 2023] Line-search-free accelerated gradient method, similar to ours step-size formula, but only for deterministic optimization.

Part II: Main Algorithms and Results

Problem Formulation

Composite optimization problem:

$$F^* = \min_{x \in \operatorname{dom} \psi} \{F(x) = f(x) + \psi(x)\},$$

where f and ψ are convex functions, ψ is simple.

Assumptions:

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

Discussion:

- Most important example: ψ is $\{0, +\infty\}$ indicator of set Q.
- Our methods require D and automatically adapt to ν, L_ν and σ.

Universal Stochastic Gradient Method

Method: Choose $x_0 \in \operatorname{dom} \psi$, set $H_0 = 0$ and iterate:

$$x_{k+1} = \operatorname*{argmin}_{x \in \operatorname{dom} \psi} \Big\{ \langle g_k, x \rangle + \psi(x) + \frac{H_k}{2} \|x - x_k\|^2 \Big\}, \qquad g_k \sim \hat{g}(x_k),$$

$$H_{k+1} = H_k + \frac{[\hat{\beta}_{k+1} - \frac{H_k}{2} r_{k+1}^2]_+}{D^2 + \frac{1}{2} r_{k+1}^2}, \quad \text{where} \quad \begin{aligned} r_{k+1} &= \|x_{k+1} - x_k\|, \\ \hat{\beta}_{k+1} &= \langle g_{k+1} - g_k, x_{k+1} - x_k \rangle \end{aligned}$$

- $\hat{\beta}_{k+1}$ is a stoch. estimate of symmetrized Bregman distance:

$$\hat{\beta}_f(x, y) = \langle \nabla f(y) - \nabla f(x), y - x \rangle = \beta_f(x, y) + \beta_f(y, x),$$

  where $\beta_f(x, y) = f(y) - f(x) - \langle \nabla f(x), y - x \rangle$.
- Convergence rate for $\bar{x}_k = \frac{1}{k} \sum_{i=1}^k x_i$:

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

Universal Stochastic Fast Gradient Method

Set $v_0 = x_0$, $H_0 = A_0 = 0$, $a_k = k$, $A_k = \sum_{i=1}^k a_i = \frac{1}{2}k(k+1)$ and iterate

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

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

$$x_{k+1} = \frac{A_k x_k + a_{k+1} v_{k+1}}{A_{k+1}},$$

$$H_{k+1} = H_k + \frac{[A_{k+1}\hat{\beta}_{k+1} - \frac{H_k}{2} r_{k+1}^2]_+}{D^2 + \frac{1}{2} r_{k+1}^2}, \quad \begin{aligned} r_{k+1} &= \|v_{k+1} - v_k\|, \\ \hat{\beta}_{k+1} &= \langle g_{k+1}^x - g_{k+1}^y, x_{k+1} - y_k \rangle, \\ g_{k+1}^x &\sim \hat{g}(x_{k+1}). \end{aligned}$$

Convergence rate:

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

Part III: Main Ideas and Outline of Analysis

Starting Recurrence

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

- Central inequality (for $d_k = \|x_k - x^*\|$, $r_{k+1} = \|x_{k+1} - x_k\|$):

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

(Cf: $\phi(x) \geq \phi(\bar{x}) + \frac{\mu}{2} \|x - \bar{x}\|^2$ for μ-strongly cvx ϕ with minimizer $\bar{x}$.)

- Estimating $f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*)$ and rearranging gives

$$F(x_{k+1}) - F^* + \frac{H_k}{2} d_{k+1}^2 \leq \frac{H_k}{2} d_k^2 + \beta_{k+1} - \frac{H_k}{2} r_{k+1}^2, \qquad (*)$$

where $\beta_{k+1} = f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k \rangle \equiv \beta_f(x_k, x_{k+1})$.

Universal Gradient Method with Line Search – I

Recall: For $\beta_{k+1} = \beta_f(x_k, x_{k+1})$, $r_{k+1} = \|x_{k+1} - x_k\|$, we have

$$F(x_{k+1}) - F^* + \frac{H_k}{2} d_{k+1}^2 \le \frac{H_k}{2} d_k^2 + \beta_{k+1} - \frac{H_k}{2} r_{k+1}^2. \qquad (*)$$

Line-Search Approach: Choose H_k such that $\boxed{\beta_{k+1} - \frac{H_k}{2} r_{k+1}^2 \le \frac{\epsilon}{2}}$ (#),

and divide $(*)$ by H_k to make d_k^2-terms telescopic:

$$\frac{1}{H_k}[F(x_{k+1}) - F^*] + \frac{1}{2} d_{k+1}^2 \le \frac{1}{2} d_k^2 + \frac{\epsilon}{2H_k}.$$

Telescoping and diving by $S_k = \sum_{i=0}^{k-1} \frac{1}{H_i}$, we get (for $H_k^* = \max\limits_{0 \le i \le k-1} H_i$)

$$F(x_k^*) - F^* \le \frac{d_0^2}{2S_k} + \frac{\epsilon}{2} \le \frac{H_k^* d_0^2}{2k} + \frac{\epsilon}{2}. \qquad (**)$$

It remains to upper bound H_k^*.

Universal Gradient Method with Line Search – II

Recall: H_k needs to satisfy $\Delta_k := \beta_{k+1} - \frac{H_k}{2} r_{k+1}^2 \leq \frac{\epsilon}{2}$ (#).

- Since $\beta_{k+1} \equiv f(x_{k+1}) - f(x_k) - \langle \nabla f(x_k), x_{k+1} - x_k \rangle \leq \frac{L_\nu}{1+\nu} r_{k+1}^{1+\nu}$, we can estimate (maximizing in the expression in r_{k+1}):

$$\Delta_k \leq \frac{L_\nu}{1+\nu} r_{k+1}^{1+\nu} - \frac{H_k}{2} r_{k+1}^2 \leq \frac{(1-\nu) L_\nu^{2/(1-\nu)}}{2(1+\nu) H_k^{(1+\nu)/(1-\nu)}}.$$

- Hence, (#) is satisfied whenever $H_k \geq \bar{H}_\nu$, where

$$\bar{H}_\nu := L_\nu^{2/(1+\nu)} \left[\frac{1-\nu}{(1+\nu)\epsilon} \right]^{(1-\nu)/(1+\nu)}.$$

- Line search ensures that $H_k \leq 2\bar{H}_*$, where $\bar{H}_* := \inf_{\nu \in [0,1]} \bar{H}_\nu$.
- Substituting this bound into $(**)$, we get the final complexity of

$$O\left(\inf_{\nu \in [0,1]} \frac{\bar{H}_\nu d_0^2}{\epsilon} \right) = O\left(\inf_{\nu \in [0,1]} \left[\frac{L_\nu}{\epsilon} \right]^{2/(1+\nu)} d_0^2 \right)$$

iterations to reach $F(x_k^*) - F^* \leq \epsilon$.

Our Approach: How to Avoid Line Search

Recall: $F(x_{k+1}) - F^* + \frac{H_k}{2} d_{k+1}^2 \leq \frac{H_k}{2} d_k^2 + \beta_{k+1} - \frac{H_k}{2} r_{k+1}^2$. $(*)$

To make ($d_k \equiv \|x_k - x^*\|$)-terms telescope, require $H_k \leq H_{k+1}$ and rewrite

$$F(x_{k+1}) - F^* + \frac{H_{k+1}}{2} d_{k+1}^2 - \frac{H_k}{2} d_k^2 \leq \beta_{k+1} - \frac{H_k}{2} r_{k+1}^2 + \frac{1}{2}(H_{k+1} - H_k) d_{k+1}^2$$
$$\leq \beta_{k+1} - \frac{H_k}{2} r_{k+1}^2 + \frac{1}{2}(H_{k+1} - H_k) D^2.$$

Main idea: Choose H_{k+1}: $\boxed{\frac{1}{2}(H_{k+1} - H_k) D^2 = \left[\beta_{k+1} - \frac{H_k}{2} r_{k+1}^2\right]_+}$ $(\#)$

Then, we get easy-to-telescope recurrence:

$$F(x_{k+1}) - F^* + \frac{H_{k+1}}{2} d_{k+1}^2 \leq \frac{H_k}{2} d_k^2 + (H_{k+1} - H_k) D^2,$$

which gives us, after telescoping,

$$F(x_k^*) - F^* \leq \frac{1}{k}\left[\frac{H_0}{2} d_0^2 + (H_k - H_0) D^2\right] \leq \frac{H_k D^2}{k}.$$

Our Approach: Estimating growth rate of H_k

- To estimate growth of H_k, use (#) and Hölder smoothness:

$$\frac{1}{2}(H_{k+1} - H_k)D^2 = \Big[\beta_{k+1} - \frac{H_k}{2} r_{k+1}^2\Big]_+ \leq \frac{(1-\nu)L_\nu^{2/(1-\nu)}}{2(1+\nu)H_k^{(1+\nu)/(1-\nu)}}.$$

- Suppose we have H_{k+1} instead of H_k in the right-hand side. This is $C \geq M_{k+1}^{p-1}(M_{k+1} - M_k) \geq \int_{M_k}^{M_{k+1}} t^{p-1} dt = \frac{1}{p}(M_{k+1}^p - M_k^p)$, which means that $M_k \leq (pCk)^{1/p}$ (provided that $M_0 = 0$). Thus,

$$H_k \lesssim \inf_{\nu \in [0,1]} \frac{L_\nu}{D^{1-\nu}} k^{(1-\nu)/2},$$

and $F(x_k^*) - F^* \leq \frac{H_k D^2}{k} \lesssim \inf_{\nu \in [0,1]} \frac{L_\nu D^{1+\nu}}{k^{(1+\nu)/2}} \leq \epsilon$ in

$$O\bigg(\inf_{\nu \in [0,1]} \Big[\frac{L_\nu}{\epsilon}\Big]^{2/(1+\nu)} D^2\bigg) \quad \text{iterations.}$$

Our Approach: Final Comments

To replace H_k with H_{k+1}, we go back to $(*)$, rewrite

$$F(x_{k+1}) - F^* + \frac{H_{k+1}}{2} d_{k+1}^2 - \frac{H_k}{2} d_k^2 \leq \beta_{k+1} - \frac{H_k}{2} r_{k+1}^2 + \frac{1}{2}(H_{k+1} - H_k) D^2$$
$$\leq \beta_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2 + (H_{k+1} - H_k) D^2,$$

and choose H_{k+1} from $\boxed{(H_{k+1} - H_k) D^2 = \Big[\beta_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2\Big]_+}$ $(\#')$.

The explicit solution is $H_{k+1} = H_k + \dfrac{[\hat{\beta}_{k+1} - \frac{H_k}{2} r_{k+1}^2]_+}{D^2 + \frac{1}{2} r_{k+1}^2}$.

Proceed as before: $F(x_{k+1}) - F^* + \frac{H_{k+1}}{2} d_{k+1}^2 \leq \frac{H_k}{2} d_k^2 + 2(H_{k+1} - H_k) D^2$, to get

$$F(x_k^*) - F^* \leq \frac{2 H_k D^2}{k} \leq \inf_{\nu \in [0,1]} \frac{2 L_\nu D^{1+\nu}}{k^{(1+\nu)/2}}.$$

Stochastic Oracle: Outline of Analysis

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

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

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

Using $\mathbb{E}_{\xi_k}[f(x_k) + \langle g_k, x^* - x_k \rangle] = f(x_k) + \langle \nabla f(x_k), x^* - x_k \rangle \leq f(x^*)$ (assuming that $g_k \equiv g(x_k, \xi_k)$) and rearranging as before, we get

$$\mathbb{E}\Big[F(x_{k+1}) - F^* + \frac{H_{k+1}}{2} d_{k+1}^2 - \frac{H_k}{2} d_k^2 \Big] \\ \leq \mathbb{E}\Big[\beta_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2 + (H_{k+1} - H_k) D^2 \Big],$$

where $\beta_{k+1} \coloneqq f(x_{k+1}) - f(x_k) - \langle g_k, x_{k+1} - x_k \rangle$.

Stochastic Oracle: Outline of Analysis – II

Our recurrence:

$$\mathbb{E}\Big[F(x_{k+1}) - F^* + \frac{H_{k+1}}{2} d_{k+1}^2 - \frac{H_k}{2} d_k^2\Big] \le \mathbb{E}\Big[\beta_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2 + (H_{k+1} - H_k) D^2\Big],$$

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

Note: Cannot compute β_{k+1}!

Main idea: Estimate $\beta_{k+1} \le \langle \nabla f(x_{k+1}) - g_k, x_{k+1} - x_k \rangle = \mathbb{E}_{\xi_{k+1}}[\hat{\beta}_{k+1}]$, where $\hat{\beta}_{k+1} := \langle g_{k+1} - g_k, x_{k+1} - x_k \rangle$ can be computed, and choose H_{k+1} from equation

$$\boxed{(H_{k+1} - H_k) D^2 = \Big[\hat{\beta}_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2\Big]_+}$$

This gives us, as before,

$$\mathbb{E}[F(\bar{x}_k)] - F^* \le \frac{2\,\mathbb{E}[H_k] D^2}{k}.$$

Stochastic Oracle: Estimating growth of H_k

To estimate growth of H_k, we first estimate

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

where $\Delta_{k+1} \coloneqq \delta_{k+1} - \delta_k$ with $\delta_k \coloneqq g_k - \nabla f(x_k)$, and $\sigma_{k+1} \coloneqq \|\Delta_{k+1}\|$ (note: $\mathbb{E}[\sigma_{k+1}^2] \leq 2\sigma^2$).

This gives us

$$(H_{k+1} - H_k) D^2 = \left[\hat{\beta}_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2\right]_+ \lesssim \frac{(1-\nu) L_\nu^{2/(1-\nu)}}{(1+\nu) H_{k+1}^{(1+\nu)/(1-\nu)}} + \frac{\sigma_{k+1}^2}{H_{k+1}}.$$

Analyzing recurrence gives $H_k \leq O\big(\frac{L_\nu}{D^{1-\nu}} k^{(1-\nu)/2} + \frac{1}{D} (\sum_{i=1}^k \sigma_i^2)^{1/2}\big)$, so

$$\mathbb{E}[H_k] \leq O\Big(\inf_{\nu \in [0,1]} \frac{L_\nu}{D^{1-\nu}} k^{(1-\nu)/2} + \frac{\sigma}{D} \sqrt{k}\Big).$$

Comparison with AdaGrad-type Methods

Recall main recurrence: (for $\hat{\beta}_{k+1} := \langle g_{k+1} - g_k, x_{k+1} - x_k \rangle$)

$$\mathbb{E}\Big[F(x_{k+1}) - F^* + \frac{H_{k+1}}{2} d_{k+1}^2 - \frac{H_k}{2} d_k^2\Big] \le \mathbb{E}\Big[\hat{\beta}_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2 + (H_{k+1} - H_k) D^2\Big]$$

- Note that (for $\gamma_{k+1} := \|g_{k+1} - g_k\|$)

$$\hat{\beta}_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2 \le \gamma_{k+1} r_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2 \le \frac{\gamma_{k+1}^2}{2H_{k+1}}.$$

- So in our alg., $(H_{k+1} - H_k) D^2 = [\hat{\beta}_{k+1} - \frac{H_{k+1}}{2} r_{k+1}^2]_+ \le \frac{\gamma_{k+1}^2}{2H_{k+1}}$, i.e.,

$$H_k \le H_k' := \frac{1}{D}\Big(\sum_{i=1}^k \gamma_i^2\Big)^{1/2} \quad \text{(AdaGrad step-size coefficient)}$$

- Thus, our "step-size" $\frac{1}{H_k}$ is smaller than $\frac{1}{H_k'}$ of AdaGrad.
- AdaGrad corresponds to balance equation $(H_{k+1} - H_k) D^2 = \frac{\gamma_{k+1}^2}{2H_{k+1}}$.

Conclusions

Conclusions

- We presented Universal gradient methods for Stochastic Optimization.
- They only need to know diameter D of feasible set, and automatically adjust to smoothness class (ν, L_ν) and oracle's variance σ.
- These are standard methods which use a special rule for adjusting step-size coefficients based on the idea of balancing the two error terms arising in the convergence analysis.

Paper

Universal Gradient Methods for Stochastic Convex Optimization
arXiv:2402.03210

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

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.

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.

References III

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.

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.

References IV

K. Y. Levy, A. Yurtsever, and V. Cevher. Online Adaptive Methods, Universality and Acceleration. In **Neural and Information Processing Systems (NeurIPS)**, 2018.

T. Li and G. Lan. A simple uniformly optimal method without line search for convex optimization. 2023. arXiv: 2310.10082.

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

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.