UniXGrad: A Universal, Adaptive Algorithm with Optimal Guarantees for Constrained Optimization

Ali Kavis\textsuperscript{1}  
EPFL

Kfir Y. Levy\textsuperscript{1}  
Technion

Francis Bach  
INRIA

Volkan Cevher  
EPFL

NeurIPS 2019

December 10, 2019

\textsuperscript{1}Equal contribution
Introduction

Problem Definition

\[ \min_{x \in \mathcal{K}} f(x) \]  

\( f : \mathcal{K} \rightarrow \mathbb{R} \) is a convex function

\( \mathcal{K} \subset \mathbb{R}^d \) is a \textbf{compact}, convex set
Motivation

Aim of this work

• **Universal:** Optimal for smooth and non-smooth problems

• **Adaptive:** No knowledge of Lipschitz constant and variance

• **Constrained:** Extend existing results to constrained problems
Our algorithm: UniXGrad

**Algorithm 1** UniXGrad

**Input:** # of iterations $T$, $y_0 \in K$, weight $\alpha_t = t$, learning rate $\{\eta_t\}_{t \in [T]}

1: for $t = 1, \ldots, T$ do

2: $x_t = \text{arg min}_{x \in K} \alpha_t \langle x, M_t \rangle + \frac{1}{\eta_t} D_R(x, y_{t-1})$, $M_t = \nabla f(\tilde{z}_t)$

3: $y_t = \text{arg min}_{y \in K} \alpha_t \langle y, g_t \rangle + \frac{1}{\eta_t} D_R(y, y_{t-1})$, $g_t = \nabla f(\bar{x}_t)$

4: end for

\[ \bar{x}_t = \frac{\alpha_t x_t + \sum_{i=1}^{t-1} \alpha_i x_i}{\sum_{i=1}^{t} \alpha_i} \quad \tilde{z}_t = \frac{\alpha_t y_{t-1} + \sum_{i=1}^{t-1} \alpha_i x_i}{\sum_{i=1}^{t} \alpha_i} \] (2)

\[ \eta_t = \frac{2D}{\sqrt{1 + \sum_{i=1}^{t-1} \alpha_i^2 \|g_i - M_i\|^2}} \] (3)
Conversion Scheme and Adaptive Bounds

Weighted Regret:

\[ R_T(x_*) = \sum_{t=1}^{T} \alpha_t \langle \nabla f(\bar{x}_t), x_t - x_* \rangle \]

Adaptive bound

\[ R_T(x_*) \leq \frac{7}{2} D \sqrt{1 + \sum_{t=1}^{T} \alpha_t^2 \|g_t - M_t\|_x^2} - \frac{D}{2} \] (4)

Lemma (Regret ⇒ Rate)

\[ f(\bar{x}_T) - f(x_*) \leq \frac{2R_T(x_*)}{T^2}. \] (5)
Convergence in Non-smooth Setting

Theorem

If $f$ is $G$-Lipschitz, Algorithm 1 guarantees

$$E \left[ f(\bar{x}_T) \right] - \min_{x \in K} f(x) \leq \frac{6D}{T^2} + \frac{14GD}{\sqrt{T}}.$$  \hspace{1cm} (6)

Remark

- Regret analysis is agnostic to definitions of $g_t$ and $M_t$.
- Conversion scheme requires $g_t = \nabla f(\bar{x}_t)$. 
Convergence in Smooth Setting

**Theorem**

If $f$ is $L$-smooth and oracle is **deterministic**, Algorithm 1 ensures

$$f(\bar{x}_T) - \min_{x \in K} f(x) \leq \frac{20\sqrt{7}D^2L}{T^2}. \quad (7)$$

If oracle is **stochastic**, 

$$\mathbb{E}[f(\bar{x}_T)] - \min_{x \in K} f(x) \leq \frac{224\sqrt{14}D^2L}{T^2} + \frac{14\sqrt{2}\sigma D}{\sqrt{T}}. \quad (8)$$

**Remark**

$f$ is $L$-smooth $\Rightarrow$ bounded gradients are not required.
Convergence Behavior

Least-squares with $\ell_2$ norm ball constraint:

$$\min_{\|x\|_2 < r} \frac{1}{2n} \|Ax - b\|^2_2,$$

(a) Average Iterate

(b) Last Iterate

Figure 1: Convergence in the stochastic oracle setting, $x_* \in \text{Boundary}(\mathcal{K})$
SVM Classification

- SVM with squared Hinge loss and $\ell_2$ regularization
- breast-cancer data from libsvm dataset, 80/20 train/test ratio
- Training batch size: 5, number of runs: 5.

![Graph showing convergence and test accuracy](image)

(a) Convergence w.r.t. training data  
(b) Test Accuracy

Figure 2: SVM classification using breast-cancer data (Chang and Lin, 2011)