DADA: Dual Averaging with Distance Adaptation

Mohammad Moshtaghifar [1] Anton Rodomanov [2] Daniil Vankov [3] Sebastian U. Stich [2]

[1]Sharif University of Technology, Iran [2]CISPA Helmholtz Center for Information Security, Germany [3]Arizona State University, USA

Problem Formulation

Consider the convex optimization problem:

$$f^* \coloneqq \min_{x \in Q} f(x), \tag{1}$$

where $f\colon \mathbb{R}^d \to \mathbb{R}$ a convex function on Q and $Q \subseteq \mathbb{R}^d$ is a simple and nonempty closed convex set.

Notation. In this text, we work in the space $\mathbb{R}^d$ equipped with the standard inner product $\langle \cdot, \cdot \rangle$ and the general Euclidean (Mahalanobis) norm:

$$\|x\| \coloneqq \langle Bx, x \rangle^{1/2}, \qquad x \in \mathbb{R}^d,$$

where B is a fixed symmetric positive definite matrix.

$$\|s\|_* \coloneqq \max_{\|x\|=1} \langle s, x \rangle = \langle s, B^{-1}s \rangle^{1/2}, \qquad s \in \mathbb{R}^d.$$

Goal. Develop a universal, parameter-free method for solving (1).

- Existing parameter-free methods often require assumptions about f, (e.g., f must be nonsmooth Lipschitz or, in some cases, Lipschitz-smooth).
- Can we design a method that works across a broader class of convex functions without such assumptions?

Measuring The Quality of Solution

We focus on bounding the distance from x^* to the hyperplane $\{y : \langle \nabla f(x), x - y \rangle = 0\}$. Specifically, we define:

$$v(x) \coloneqq \frac{\langle \nabla f(x), x - x^* \rangle}{\|\nabla f(x)\|_*}.$$

Minimizing $v(x)$ also reduces the corresponding function residual, $f(x) - f^*$,

$$f(x) - f^* \leq \omega(v(x)), \tag{2}$$

where

$$\omega(t) \coloneqq \max_x \{ f(x) - f^* : \|x - x^*\| \leq t \},$$

measures the local growth of f around the solution x^*.

Classical Scheme of Dual Averaging

Algorithm General Scheme of DA

Require: $x_0 \in Q$, $T \geq 1$, coefficients $(a_k)_{k=0}^{T-1}$, $(\beta_k)_{k=1}^{T}$ with nondecreasing β_k
  for $k = 1, \ldots, T$ **do**
    Compute arbitrary $g_k \in \partial f(x_k)$
    $x_k = \operatorname{argmin}_{x \in Q} \{ \psi_k(x) = \sum_{i=0}^{k-1} a_i \langle g_i, x - x_i \rangle + \frac{\beta_k}{2} \|x - x_0\|^2 \}$
  end for
Ensure: $x_T^* = \operatorname{argmin}_{x \in \{x_0, \ldots, x_T\}} f(x)$

Theorem (Nesterov 2005): For all $1 \leq k \leq T$, it holds that

$$\sum_{i=0}^{k-1} a_i v_i \|g_i\|_* + \frac{\beta_k}{2} \|x_k - x^*\|^2 \leq \frac{\beta_k}{2} D_0^2 + \sum_{i=0}^{k-1} \frac{a_i^2}{2\beta_i} \|g_i\|_*^2,$$

where $D_0 = \|x_0 - x^*\|$ be the initial gap.

The classical DA method has two primary variants.

- Simple DA: uses a constant coefficient, $a_i = \hat{D}_0$.
- Weighted DA: adjusts the coefficients dynamically, $a_i = \frac{\hat{D}_0}{\|g_i\|_*}$.

Additional Price: multiplying the correct complexity by ρ^2, where

$$\rho = \max\{\frac{\hat{D}_0}{D_0}, \frac{D_0}{\hat{D}_0}\}$$

Main Problem: The cost is significantly high. Could this multiplicative factor be reduced to a logarithmic factor?

Estimating D_0

Using the distance between x_i and the initial point x_0 to estimate D_0, is an an idea that has been explored,

Definition: for all $1 \leq k \leq T$,

$$\bar{r}_k = \max\{\max_{1 \leq t \leq k} r_t, \bar{r}\}, \; r_t = \|x_0 - x_t\|,$$

and $\bar{r}$ is a certain user-specified parameter.

Here, it is the recent method that use $(\bar{r}_i)_{i=1}^T$ to estimate D_0.

- **DoG** (Ivgi et al. 2023):

  $$x_{k+1} = \operatorname{Proj}(x_k - \eta_k g_k),$$

  where

  $$\eta_k = \frac{\bar{r}_k}{\sqrt{\sum_{i=0}^k \|g_i\|^2}}.$$

  Main Differences:
  - Using Dual Averaging instead of Gradient Descent.
  - Normalizing by g_k instead of accumulated norms.

DADA Method

To address the limitation of classical Dual Averaging, we propose the following coefficients,

$$a_k = \frac{\bar{r}_k}{\|g_k\|_*}, \quad \beta_k = 2\sqrt{k+1} \tag{3}$$

Theorem: Consider Dual Averaging for solving problem (1) using the coefficients from eq. (3). Then, using $v_T^* = \min_{0 \leq t \leq T} v_t$,

$$f(\bar{x}_T^*) - f^* \leq \omega(v_T^*),$$

where

$$v_T^* \leq \frac{6R}{\sqrt{T}} \left(\frac{8R}{\bar{r}}\right)^{\frac{1}{T}} \log \frac{8eR}{\bar{r}},$$

and $R = \max\{\|x_0 - x^*\|, \bar{r}\}$. Further, for a given $\delta > 0$, it holds that $v_T^* \leq \delta$ whenever

$$T \geq \max\left\{\log \frac{8R}{\bar{r}}, \frac{36e^2R^2}{\delta^2} \log^2 \frac{8eR}{\bar{r}}\right\}$$

Universality of DADA: Examples of Applications

If $\nabla f(x^*) = 0$, then DADA requires, at most, the following iterations to achieve ϵ-accuracy:

Hölder-smooth functions

$\|\nabla f(x) - \nabla f(y)\|_* \leq H_\nu \|x - y\|^\nu$ for all $x, y \in Q$.

$$O\left(\left[\frac{H_\nu}{\epsilon}\right]^{\frac{2}{1+\nu}} R^2 \log_+^2 \frac{R}{\bar{r}}\right)$$

Functions with Lipschitz high-order derivative

$\|\nabla^p f(x) - \nabla^p f(y)\|_* \leq L_p \|x - y\|$ for all $x, y \in \mathbb{R}^d$ and a $p \geq 1$. For example, softmax is in this class.

$$O\left(\max\left\{\max_{2 \leq i \leq p} \left[\frac{p}{i!} \frac{\|\nabla^i f(x^*)\|_*}{\epsilon}\right]^{\frac{2}{i}}, \left[\frac{p}{(p+1)!} \frac{L_p}{\epsilon}\right]^{\frac{2}{p+1}}\right\} R^2 \log_+^2 \frac{R}{\bar{r}}\right),$$

Quasi-self-concordant (QSC) functions

$\nabla^3 f(x)[u, u, v] \leq M \langle \nabla^2 f(x) u, u \rangle \|v\|$ for all $x \in \mathbb{R}^d$ and arbitrary directions $u, v \in \mathbb{R}^d$. For example, exponential function is in this class.

$$O\left(\frac{\|\nabla^2 f(x^*)\|_* R^2}{\epsilon} \log_+^2 \frac{R}{\bar{r}} + (MR)^2 \log_+^2 \frac{R}{\bar{r}} + \log_+ \frac{R}{\bar{r}}\right).$$

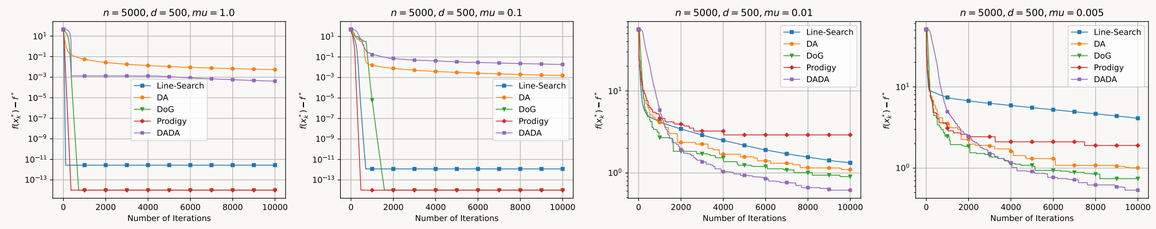
(L0, L1)-smooth functions

$\|\nabla^2 f(x)\|_* \leq L_0 + L_1 \|\nabla f(x)\|_*$ for all $x \in \mathbb{R}^n$. For example $f(x) = \|x\|^p$ is in this class

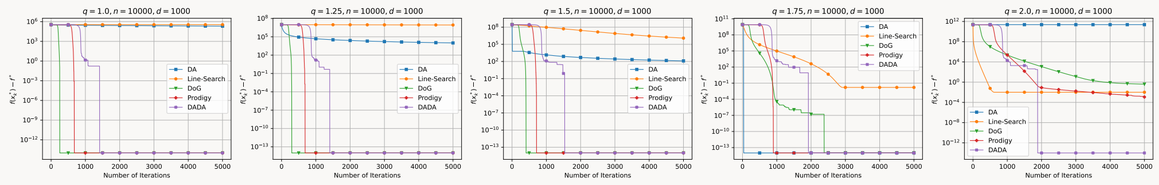
$$O\left(\frac{L_0 R^2}{\epsilon} \log_+^2 \frac{R}{\bar{r}} + (L_1 R)^2 \log_+^2 \frac{R}{\bar{r}} + \log_+ \frac{R}{\bar{r}}\right).$$

Numerical Results

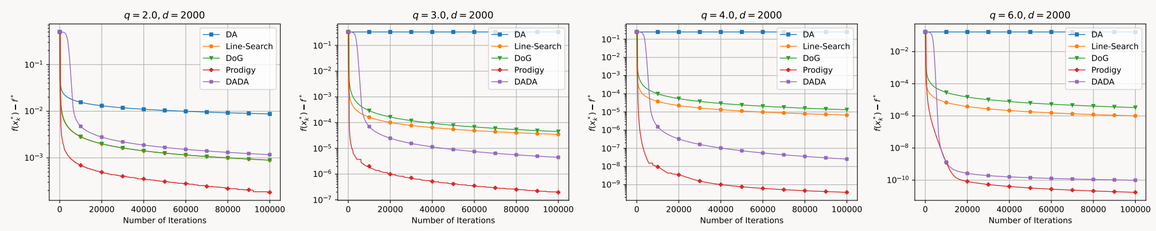
Softmax. $\min_{x \in \mathbb{R}^d} f(x) \coloneqq \mu \log\left(\sum_{i=1}^n \exp\left[\frac{\langle a_i, x \rangle - b_i}{\mu}\right]\right)$



Polyhedron Feasibility Problem. $\min_{x \in \mathbb{R}^d} f(x) \coloneqq \frac{1}{n} \sum_{i=1}^n [\langle a_i, x \rangle - b_i]_+^q$



Worst-case Function. $\min_{x \in \mathbb{R}^d} f(x) \coloneqq \frac{1}{q} \sum_{i=1}^n |x^{(i)} - x^{(i+1)}|^q + \frac{1}{q} |x^{(n)}|^q$



References

Ivgi, Maor, Oliver Hinder, and Yair Carmon (2023). "DoG is SGD's Best Friend: A Parameter-Free Dynamic Step Size Schedule". In: *Proceedings of the 40th International Conference on Machine Learning*, pp. 14465–14499.

Nesterov, Yurii (2005). "Primal-dual subgradient methods for convex problems". In: *Mathematical Programming* 120, pp. 221–259.

— (2018). *Lectures on Convex Optimization*. 2nd. Springer Publishing Company, Incorporated, pp. 221–259. ISBN: 3319915770.