Proceedings of Machine Learning Research vol 247:1–68, 2024 37th Annual Conference on Learning Theory

Accelerated Parameter-Free Stochastic Optimization

Itai Kreisler KREISLER@MAIL.TAU.AC.IL
Maor Ivgi MAOR.IVGI@CS.TAU.AC.IL
Oliver Hinder OHINDER@PITT.EDU
Yair Carmon YCARMON@TAUEX.TAU.AC.IL

Editors: Shipra Agrawal and Aaron Roth

Abstract

We propose a method that achieves near-optimal rates for *smooth* stochastic convex optimization and requires essentially no prior knowledge of problem parameters. This improves on prior work which requires knowing at least the initial distance to optimality d_0. Our method, U-DoG, combines UniXGrad (Kavis et al. [30]) and DoG (Ivgi et al. [27]) with novel iterate stabilization techniques. It requires only loose bounds on d_0 and the noise magnitude, provides high probability guarantees under sub-Gaussian noise, and is also near-optimal in the non-smooth case. Our experiments show consistent, strong performance on convex problems and mixed results on neural network training.

Keywords: Parameter-free, Adaptive, Stochastic convex optimization, Smooth optimization.

1. Introduction

We consider the problem of minimizing a smooth convex function using access to an unbiased stochastic gradient oracle. This is a fundamental problem in machine learning, including many important special cases such as logistic and linear regression. Moreover, the smoothness assumption is crucial for developing one of the most widely used improvements for the classical gradient method: Nesterov acceleration [44].

Nesterov acceleration obtains the optimal rate of convergence for this problem but is strongly reliant on knowing the problem parameters. Specifically, Lan [35], who first demonstrated the theoretical value of Nesterov acceleration on smooth *stochastic* convex functions, requires knowledge of the smoothness parameter β, the distance d_0 from the initial point to the optimum, and a value σ for which the noise is σ-sub-Gaussian. Accelerated adaptive methods [14, 30] do not require knowledge of β and σ, but assume knowledge of d_0. For *non-smooth* stochastic convex optimization, *parameter-free methods* [e.g., 7, 9, 16, 27, 28, 41, 49] require only loose knowledge of problem parameters to obtain near-optimal rates. Finding such parameter-free methods for *smooth* stochastic optimization is a longstanding open problem.

Our contribution. We solve this open problem, designing an accelerated parameter-free method which we call UniXGrad-DoG, or U-DoG for short. U-DoG combines the "universal extra-gradient" (UniXGrad) framework [30] with the "distance over gradient" (DoG) technique [27]. More specifically, we replace the domain diameter D in the UniXGrad step size numerator with the maximum distance from the initial point, similar to the DoG step size numerator. Furthermore, we use this maximum distance to automatically tune the "momentum" parameter α_t of UniXGrad.

Algorithm name	Unbounded domain?	Insensitive to... d_0/D	β	σ	Rate of convergence	High probability?
U-DoG (this work)	✓	✓	✓	✓	$\widetilde{O}\left(\frac{\beta d_0^2}{T^2} + \frac{\sigma d_0}{\sqrt{T}} + \frac{\hat{\sigma} d_0}{T}\right)$	✓
	✗	✓	✓	✓	$\widetilde{O}\left(\frac{\beta D^2}{T^2} + \frac{\sigma D}{\sqrt{T}}\right)$	✓
UNIXGRAD [30]	✗	✗	✓	✓	$O\left(\frac{\beta D^2}{T^2} + \frac{\sigma D}{\sqrt{T}}\right)$	✗
Cutkosky [14]	✓	✗	✓	✓	$\widetilde{O}\left(\frac{\beta d_0^2}{T^2} + \frac{\sigma d_0}{\sqrt{T}}\right)$	✗
Lan [35]	✓	✗	✗	✗	$O\left(\frac{\beta d_0^2}{T^2} + \frac{\sigma d_0}{\sqrt{T}}\right)$	✓
DoG [27] / CO [16]	✓	✓	✓	✗	$\widetilde{O}\left(\frac{\beta d_0^2}{T} + \frac{\sigma d_0}{\sqrt{T}} + \frac{\hat{L} d_0}{T}\right)$	✓ / ✗

Table 1: Comparison of U-DoG and prior work on β-smooth stochastic optimization with σ-sub-Gaussian noise. "Unbounded domain" indicates if the algorithm is defined over the whole Euclidean space or a bounded subspace. In the former case we express rates in terms of the initial distance to optimality d_0 and in the latter case we use the domain diameter D. Under "Insensitive to..." we mark ✗ if the suboptimality bound grows polynomially with error in the parameter, ✓ if it only affects logarithmic factors or low order terms, and ✓ if there is no dependence on the parameter at all. The marker ✗ indicates algorithms that require an upper bound $\hat{L}$ on gradient norm, which may be much larger the the upper bound $\hat{\sigma}$ on the noise. The notation $\widetilde{O}(\cdot)$ hides polylogarithmic factors.

Finally, we modify the UNIXGRAD step size denominator to ensure the stability of the iterate sequence. U-DoG only requires a loose upper bound $\hat{\sigma}$ on σ and lower bound r_ϵ on D.[1] As long as $\hat{\sigma}$ is loose by at most a $\sqrt{T}$ factor and r_ϵ is loose by any $\mathrm{poly}(T)$ factor, we obtain a near-optimal, high-probability rate of convergence; Table 1 states U-DoG's guarantees and compares it to prior work. Moreover, U-DoG simultaneously enjoys a near-optimal, parameter-free rate of convergence for *non-smooth* problems.

We conduct preliminary experiments with U-DoG as well as another algorithm, A-DoG, which combines ACCELEGRAD [36] and DoG. On convex optimization problems, both U-DoG and A-DoG often substantially improve over DoG, especially at large batch sizes, with A-DoG outperforming U-DoG, likely due to not requiring an extra-gradient computation at each step. On several problems, A-DoG matches the performance of carefully tuned SGD with Nesterov momentum. On neural network optimization problems, however, we observe that both U-DoG and A-DoG do not consistently improve over DoG.

1.1. Related work

Non-smooth stochastic optimization. The majority of tuning-insensitive stochastic optimization methods are developed for online convex optimization. Online regret bounds immediately translate to suboptimality guarantees for non-smooth stochastic optimization using online-to-batch conversion [48, Section 3]. Proposed methods divide roughly into *adaptive* algorithms such as adaptive SGD [22, 38], AdaGrad [21, 40] and variants [e.g., 33, 55, 58], and *parameter-free* methods [7, 15, 16, 28, 39, 41, 47, 49, 59]. Adaptive methods typically require no knowledge of the stochastic gradient bound but need to know the initial distance to optimality (or the domain diameter), while

1. In fact, we only require *local* upper bounds of the form $\hat{\sigma}(x)$ on the noise sub-Gaussianity.

parameter-free methods are robust to uncertainty in the distance but require some (loose) bound on the stochastic gradient norms.

Recent work [9, 27] develops parameter-free methods that hew closer to SGD and eschew online-to-batch conversion for high-probability guarantees in the stochastic setting; U-DoG continues this line. In particular, it extends the core mechanism of DoG [27] wherein iterate movement serves as a proxy for the distance to optimality. D-Adaptation [17], DoWG [32], and Prodigy [42] use a similar mechanism, but only provide guarantees for the non-stochastic setting. Ensuring the validity of the mechanism (i.e., that iterates never move too far away from the optimum) is a key challenge in its analysis. This challenge becomes greater in the smooth setting, where selecting too small of a step size nullifies the benefit of acceleration. Much of our algorithmic and analytical innovation addresses this challenge.

Non-stochastic smooth optimization. Without noise, Nesterov acceleration requires knowledge of the smoothness constant β but not the distance to optimality [44, 45]. The methods [30, 36] reverse this tradeoff, requiring the distance but not β. Line search techniques such as [6, 11] provide much stronger adaptivity, attaining the optimal gradient evaluation complexity up to an additive term that depends logarithmically on the uncertainty in β. However, line search can be challenging to employ efficiently in the stochastic setting as we can no longer accurately evaluate the function. Indeed, there are many works that analyze stochastic line search techniques [e.g., 50, 60] but none have obtained convergence guarantees close to that of Lan [35].

Smooth stochastic optimization. Several adaptive and parameter-free methods [9, 16, 22, 27, 32] converge faster on smooth functions. However, they do not improve all the way to the optimal rate (see Table 1) due to a missing "momentum" component. Cutkosky [14] gives an improved online-to-batch conversion framework that endows adaptive SGD with momentum and accelerated rates in the smooth case, but requires a bound on the distance to optimality. Kavis et al. [30] propose UNIXGRAD, combining ideas from [14] with the mirror-prox/extragradient algorithm [19, 43] and online learning [38, 54] to obtain optimal rates assuming bounded domains of known diameter D and assuming that d_0 is of the order of D. U-DoG modifies UNIXGRAD and removes both assumptions, yielding the first parameter-free accelerated method.

2. Preliminaries and algorithmic framework

In this section, we set up our notation and terminology, and use them to present the general U-DoG template (Algorithm 1) defining the algorithm up to the choice of adaptive step sizes, which we gradually develop in the following sections.

Basic notation and conventions. Throughout, $\|\cdot\|$ denotes the Euclidean norm, $\log$ is base e and $\log_+(x) \coloneqq 1 + \log(x)$. The function $\mathrm{Proj}_{\mathcal{X}}(\cdot)$ denotes Euclidean projection onto set $\mathcal{X}$. We say that $f : \mathcal{K} \to \mathbb{R}$ is β-smooth if ∇f is β-Lipschitz, i.e., $\|\nabla f(u) - \nabla f(v)\| \leq \beta\|u - v\|$ for all $u, v \in \mathcal{K}$. We write $[\cdot]_+ := \max\{\cdot, 0\}$.

In this work, we minimize an objective function f via queries to a stochastic gradient estimator $\mathcal{G}$. We make the following assumption in all of our theoretical analyses.

Assumption 1 (Made throughout) *The objective function $f : \mathcal{K} \to \mathbb{R}$ is convex, L-Lipschitz, β-smooth,[2] has closed convex domain $\mathcal{K}$, and its minimum is attained at some $x_\star \in \arg\min_{x\in\mathcal{K}} f(x)$. For all $x \in \mathcal{K}$, the gradient estimator $\mathcal{G}$ satisfies $\mathbb{E}\mathcal{G}(x) = \nabla f(x)$.*

Algorithm 1: U-DoG (UNIXGRAD-DoG) template

Input: Initial $x_0 \in \mathcal{K}$, iteration budget T, initial movement r_ϵ, step sizes $\{\eta_{x,t}, \eta_{y,t}\}$

Set $y_0 = x_0$

for $t = 0, 1, 2, \ldots, T-1$ **do**

  Set $\alpha_t = \sum_{k=0}^{t} \bar{r}_k / \bar{r}_t$ and $\omega_t = \alpha_t \bar{r}_t$ for $\bar{r}_t = \max_{k\le t} \max\{\|y_k - x_0\|, \|x_k - x_0\|, r_\epsilon\}$

  $x_{t+1} = \mathrm{Proj}_{\mathcal{K}}(y_t - \alpha_t \eta_{x,t} m_t)$ for $m_t \sim \mathcal{G}(\hat{z}_t)$ and $\hat{z}_t = \frac{\omega_t y_t + \sum_{k=0}^{t-1} \omega_k x_{k+1}}{\sum_{k=0}^{t} \omega_k}$

  $y_{t+1} = \mathrm{Proj}_{\mathcal{K}}(y_t - \alpha_t \eta_{y,t} g_t)$ for $g_t \sim \mathcal{G}(\hat{x}_t)$ and $\hat{x}_t = \frac{\omega_t x_{t+1} + \sum_{k=0}^{t-1} \omega_k x_{k+1}}{\sum_{k=0}^{t} \omega_k}$

end

return $\hat{x}_T$

Presenting U-DoG. Algorithm 1 provides the general template of U-DoG. As in UNIXGRAD [30], each iteration of the algorithm consists of two stochastic gradient steps, with each stochastic gradient queried at a moving average of iterates. Unlike UNIXGRAD, the moving average weights ω_t and the step size multipliers α_t are not fixed in advance, but are instead dynamically set based on the maximum distance moved from the origin, denoted

$$\bar{r}_t := \max_{k\le t} \max\{\|y_k - x_0\|, \|x_k - x_0\|, r_\epsilon\}.$$

The parameter r_ϵ serves as a (loose) lower bound on $\|x_0 - x_\star\|$; typically, $\bar{r}_t$ grows rapidly and then plateaus at a level roughly approximating $\|x_0 - x_\star\|$. When that happens, the sequence $\alpha_t = \sum_{k\le t} \bar{r}_k / \bar{r}_t$ grows linearly in t, similar to $\alpha_t = t + 1$ in UNIXGRAD.

To complete the specification of U-DoG we must set the step size sequence. UNIXGRAD assumes $\mathcal{K}$ the domain has Euclidean diameter D and picks step sizes of the form $\eta_{x,t} = \eta_{y,t} = \frac{\sqrt{2}D}{\sqrt{1+Q_{t-1}}}$ where

$$Q_t := \sum_{k=0}^{t} q_k \quad \text{and} \quad q_t := \alpha_t^2 \|g_t - m_t\|^2. \tag{1}$$

To handle unknown domain size and unbounded domains, U-DoG follows DoG in using $\bar{r}_t$ as the step size numerator in lieu of D. Thus, the U-DoG step size admits the general form

$$\eta_{x,t} = \frac{\bar{r}_t}{\sqrt{G_{x,t}}} \quad \text{and} \quad \eta_{y,t} = \frac{\bar{r}_t}{\sqrt{G_{y,t}}}, \quad \text{where} \quad G_{x,0} \le G_{y,0} \le G_{x,1} \le \cdots. \tag{2}$$

2. Our results hold in the non-Lipschitz or non-smooth cases by setting $L = \infty$ or $\beta = \infty$, respectively. In the non-smooth case, we define $\nabla f(x) := \mathbb{E}\mathcal{G}(x)$ and assume it is a subgradient of f.

In the appendix, we also use the notation

$$\tilde{\eta}_{x,t} = \frac{1}{\sqrt{G_{x,t}}} \text{ and } \tilde{\eta}_{y,t} = \frac{1}{\sqrt{G_{y,t}}}. \tag{3}$$

For bounded domains, setting $G_{x,t} = G_{y,t} = 1 + Q_{t-1}$ recovers the UNIXGRAD guarantees up to logarithmic factors. However, for unbounded domains, ensuring the stability of U-DOG (i.e., that $\bar{r}_t$ never grows much larger than $\|x_0 - x_\star\|$) requires more careful selection of $G_{x,t}, G_{y,t}$. Enforcing iterate stability without compromising the rate of convergence is the main challenge we overcome. To that end, we define a few frequently appearing quantities:

$$r_t \coloneqq \max\{\|y_k - x_0\|, \|x_k - x_0\|\}\ ,\ d_t \coloneqq \|y_t - x_\star\|\ ,\ \bar{d}_t \coloneqq \max_{k\le t} d_k\ ,$$

$$M_t \coloneqq \max_{k\le t}\{\alpha_k^2\|m_k\|^2\} \text{ and } \theta_{t,\delta} \coloneqq \log\frac{60\log(6t)}{\delta}.$$

UNIXGRAD as a special case. For a domain with Euclidean diameter D, setting $r_\epsilon = D\sqrt{2}$ and $G_{x,t} = G_{y,t} = 1 + Q_{t-1}$ recovers UNIXGRAD (with Euclidean distance generating function) exactly, as it implies $\bar{r}_t = D\sqrt{2}$ for all t and hence $\alpha_t = t + 1$.

3. Analysis in the noiseless case

We begin our analysis under the simplifying assumption that gradients are computed exactly.

Assumption 2 *In addition to Assumption 1, we assume that $\mathcal{G}(x) = \nabla f(x)$ with probability 1.*

This noiseless setting allows us to isolate and address the key challenges of exploiting smoothness and stabilizing the iterates.

3.1. General suboptimally bound

Our first result is a bound on the suboptimality of U-DOG for general step sizes; see Section A.1 for complete proof. To interpret Proposition 3 recall that d_0 is the initial distance to the optimum and the definition of Q_t given in (1).

Proposition 3 *In the noiseless setting (Assumption 2), suppose the* U-DOG *step sizes* (2) *satisfy $G_{x,t} \ge Q_{t-1}$ for all $t \ge 0$. Then for every $t \ge 0$ and for any number $s \ge 0$, we have*

$$f(\hat{x}_t) - f(x_\star) \le O\left(\frac{s^{3/2}\beta(\bar{r}_{t+1} + d_0)^2 + (\bar{r}_{t+1} + d_0)\left[\sqrt{\max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t}\right]_+}{\left(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\right)^2}\right). \tag{4}$$

Before sketching the proof of Proposition 3, let us explain how it yields the desired rates of convergence if we momentarily set aside iterate stability and assume $\bar{r}_t \le D$ for all t, e.g., because the domain has diameter D. In this case, we may choose $G_{x,t} = G_{y,t} = Q_{t-1}$ similarly to UNIXGRAD. Substituting $s = 1$ in eq. (4) guarantees suboptimality $O\Big(\frac{\beta D^2}{\left(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\right)^2}\Big)$. As shown in [27, Lemma 3], we have $\max_{t<T}\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1} = \Omega\big(T\log^{-1}(\bar{r}_T/r_\epsilon)\big)$, meaning that for some $t < T$ we obtain the near-optimal rate $O\Big(\frac{\beta D^2}{T^2}\log^2\frac{D}{r_\epsilon}\Big)$. Moreover, since $\alpha_t \le t + 1$ for all t,

when all gradients are bounded by L we have $Q_t = O(L^2 \sum_{k\le t} \alpha_k^2) = O(L^2 t^3)$. Substituting $s = 0$ in eq. (4) and reusing our bound on the denominator gives the near-optimal rate $O\left(\frac{LD}{\sqrt{T}} \log^2 \frac{D}{r_\epsilon}\right)$ in the non-smooth setting. We also see that setting $r_\epsilon = \Omega(D)$ recovers the UNIXGRAD guarantees in the noiseless setting, which is to be expected since $r_\epsilon = D\sqrt{2}$ recovers UNIXGRAD itself as explained in the previous section.

Our proof of Proposition 3 combines ideas from the analyses of UNIXGRAD and DOG. It centers on the weighted "regret" $\mathcal{R}_t := \sum_{k=0}^t \omega_k \langle g_k, x_{k+1} - x_\star \rangle$ where $\omega_k = \alpha_k \bar{r}_k$. This is similar to the weighted regret considered for UNIXGRAD with additional weighting by $\bar{r}_t$ used in the DOG analysis. Algebraic manipulation of $\mathcal{R}_t$ gives (recall that $d_t = \|y_t - x_\star\|$),

$$\mathcal{R}_t \le O\left(\bar{r}_{t+1}^2 \sqrt{Q_t} + \sum_{k=0}^t \left(d_k^2 - d_{k+1}^2\right) \sqrt{G_{y,k}} - \sum_{k=0}^t \|x_{k+1} - y_k\|^2 \sqrt{Q_k}\right).$$

We use a telescoping argument from DOG in order to bound $\sum_{k=0}^t \left(d_k^2 - d_{k+1}^2\right) \sqrt{G_{y,k}}$ by $O\left(\bar{r}_{t+1}(\bar{r}_{t+1} + d_0)\sqrt{G_{y,t}}\right)$. Next, following UNIXGRAD we leverage smoothness to write

$$\|x_{k+1} - y_k\|^2 = \left(\frac{\sum_{i=0}^k \omega_i}{\omega_k}\right)^2 \|\hat{x}_k - \hat{z}_k\|^2 \overset{\text{Lem. 25}}{\ge} \frac{\alpha_k^2}{4} \|\hat{x}_k - \hat{z}_k\|^2 \ge \frac{\alpha_k^2}{4\beta^2} \|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2 = \frac{q_k^2}{4\beta^2},$$

where the last equality is the first time we assumed exact $\mathcal{G}$. We then show that, for all $S \ge 0$,

$$\sum_{k=0}^t \|x_{k+1} - y_k\|^2 \sqrt{Q_k} \ge \sum_{k=0}^t \frac{q_k^2}{\beta^2} \sqrt{Q_k} \ge \Omega\left(S\sqrt{Q_t} - S^{3/2}\beta\right); \tag{5}$$

this is a streamlined version of key arguments in [30, 36] where the authors carefully split the sum above based on the value of the adaptive step size. Taking $S = s \cdot \bar{r}_{t+1}(\bar{r}_{t+1} + d_0)$ and substituting back, we get

$$\mathcal{R}_t \le O\left(s^{3/2}\beta \bar{r}_{t+1}(\bar{r}_{t+1} + d_0)^2 + \bar{r}_{t+1}(\bar{r}_{t+1} + d_0)\left[\sqrt{\max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t}\right]_+\right). \tag{6}$$

To conclude the proof, we use the following UNIXGRAD "anytime online-to-batch conversion" [14] bound:

$$f(\hat{x}_t) - f(x_\star) \le \sum_{k=0}^t \frac{\omega_k}{\sum_{i=0}^t \omega_i} \langle \nabla f(\hat{x}_k), x_{k+1} - x_\star \rangle = \frac{\mathcal{R}_t}{\sum_{k=0}^t \omega_k}, \tag{7}$$

where the last equality is the second and final time the proof uses the noiseless gradient assumption. Dividing eq. (6) by

$$\sum_{k=0}^t \omega_k \overset{\text{Lem. 25}}{\ge} \frac{1}{2}\bar{r}_t \alpha_t^2 = \frac{1}{2}\bar{r}_t \left(\sum_{k=0}^t \bar{r}_k / \bar{r}_t\right)^2 \ge \frac{1}{2}\bar{r}_{t+1} \left(\sum_{k=0}^t \bar{r}_k / \bar{r}_{t+1}\right)^2, \tag{8}$$

and employing (7) yields the suboptimality bound (4).

3.2. Iterate stability

In the discussion following Proposition 3 above, we provisionally imagined that the iterates were bounded ($\bar{r}_t \leq D$ for all t) and argued that in this case simply setting $G_{x,t} = Q_{t-1}$ and $G_{y,t} = Q_t$ suffices for obtaining optimal rates whenever $D = O(d_0)$. However, in unconstrained settings this choice of step size is hopeless, as it makes $\eta_{x,0}$ infinite, implying divergence at the first step![3]

In the following proposition, we identify two conditions that together guarantee the iterates remain appropriately bounded. The complete proof appears in Section A.2.

Proposition 4 *In the noiseless setting (Assumption 2), let $s > 0$ and define $c_t = 12 \log_+^2\left(\frac{s+Q_t}{s}\right)$. If $r_\epsilon \leq d_0$ and the* U-DOG *step sizes (2) satisfy (i) $G_{y,t} \geq c_t^2(s + Q_t)$ (with $G_{x,0} \geq 144s$), and (ii) $\max\{\|x_{t+1} - y_t\|, \|y_{t+1} - x_{t+1}\|\} \leq \frac{2\bar{r}_t}{c_{t-1}}$ for all $t \geq 0$, then we have*

$$\bar{d}_t \leq 2d_0 \;\; \textit{and} \;\; \bar{r}_t \leq 4d_0 \;\; \textit{for all } t \geq 0.$$

Let us briefly explain the two requirements in Proposition 4. Requirement (i) folds two conditions into one. The first is that we increase the UNIXGRAD denominator by a logarithmic factor—this is analogous to the step size attenuation necessary to ensure the stability of DOG (i.e., the T-DOG step size [27, Section 3.3]). The second is more subtle, requiring that $G_{y,t}$ upper bound Q_t (rather than Q_{t-1} as in UNIXGRAD and Proposition 3) and hence depend on $\|g_t - m_t\|$. This is essential for guaranteeing stability but is also the cause of considerable technical difficulty in the noisy setting. Requirement (ii) simply asks that U-DOG iterates at time t move by no more than a fraction of the estimated distance to optimality $\bar{r}_t$; a reasonable requirement if the estimate is good.

The proof of Proposition 4 is a careful application of the T-DOG stability proof [27, Proposition 2] to the U-DOG template. The key to the proof is the following modification of the UNIXGRAD online-to-batch conversion bound (7), which states that for any optimum $x_\star$ we have

$$\mathcal{R}'_t \coloneqq \sum_{k=0}^{t} \eta_{y,k}\alpha_k \left\langle g_k, x_{k+1} - x_\star \right\rangle \overset{(\star)}{=} \sum_{k=0}^{t} \eta_{y,k}\alpha_k \left\langle \nabla f(\hat{x}_k), x_{k+1} - x_\star \right\rangle \geq 0, \tag{9}$$

where $(\star)$ holds only in the noiseless setting. We algebraically manipulate $\mathcal{R}'_t$ similarly to the weighted regret in the proof of Proposition 3. Writing $Q'_t = c_{t-1}^2(s + Q_t)$, we obtain

$$0 \leq \mathcal{R}'_t \leq \sum_{k=0}^{t}\left(d_k^2 - d_{k+1}^2 + \frac{q_k\bar{r}_k^2}{\sqrt{G_{y,k}Q'_k}} + \frac{\sqrt{Q'_k} - \sqrt{G_{x,k}}}{\sqrt{G_{y,k}}}\left(\|x_{k+1} - y_k\|^2 + \|x_{k+1} - y_{k+1}\|^2\right)\right).$$

Our requirements $G_{y,k} \geq c_t^2(s + Q_t)$ (which entails $G_{x,k} \geq G_{y,k-1} \geq c_{t-1}^2(s + Q_{t-1})$) and $\|x_{k+1} - y_k\|^2 + \|x_{k+1} - y_{k+1}\|^2 \leq \frac{8\bar{r}_k^2}{c_{k-1}^2}$, allow us, with some more algebra, to bound the last two summands by $\frac{9q_k\bar{r}_k^2}{c_k(s+Q_k)}$. From here, the proof proceeds identically to the T-DOG analysis [27, Section 3.3]: we get that $\sum_{k=0}^{t}\frac{9q_k^2\bar{r}_k}{c_k(s+Q_k)} \leq \frac{\bar{r}_t^2}{16}$ by the choice of c_t, and substituting back obtain that $d_{t+1}^2 \leq d_0^2 + \frac{\bar{r}_t^2}{16}$, which by straightforward induction implies the desired bounds on $\bar{d}_t$ and $\bar{r}_t$.

3. For constrained domains, however, this choice results in a valid scheme where the first step jumps to the domain boundary. Indeed, UNIXGRAD also behaves this way for sufficiently scaled-up instances since it uses a fixed, arbitrary value for $\eta_{x,0}$. This underscores UNIXGRAD's strong reliance on the bounded domain assumption.

3.3. Rate of convergence in the noiseless case

With the conditional stability guarantee of Proposition 8 in place, we are ready to face a central challenge: finding step sizes $\eta_{x,t}, \eta_{y,t}$ that satisfy the proposition's conditions but still lead to good rates of convergence in the smooth case. Our solution is (recalling the notation $M_t = \max_{k\le t}\{\alpha_k^2\|m_k\|^2\}$):

$$\begin{aligned}\eta_{x,t} &= \frac{\bar{r}_t}{12\log_+^2\left(\frac{\|m_0\|^2+Q_{t-1}}{\|m_0\|^2}\right)\sqrt{\max\{\|m_0\|^2+Q_{t-1},M_t\}}}\\ \eta_{y,t} &= \frac{\bar{r}_t}{12\log_+^2\left(\frac{\|m_0\|^2+Q_{t}}{\|m_0\|^2}\right)\sqrt{\max\{\|m_0\|^2+Q_{t},M_t\}}}.\end{aligned} \tag{10}$$

Clearly, the step sizes (10) satisfy the first condition in Proposition 4 with $s = \|m_0\|^2$. To see why the second condition holds, note that, since $\sqrt{M_t} \ge \alpha_t\|m_t\|$, we have $\eta_{x,t} \le \frac{\bar{r}_t}{c_t\alpha_t\|m_t\|}$. By the contractive property of projections, we therefore have

$$\|x_{t+1} - y_t\| \le \eta_{x,t}\alpha_t\|m_t\| \le \frac{\bar{r}_t}{c_t} \le \frac{2\bar{r}_t}{c_t}.$$

A similar argument also shows that $\|x_{t+1} - y_{t+1}\| \le \frac{2\bar{r}_t}{c_t}$, fulfilling the conditions of Proposition 4 (see Lemma 19).

Now the question becomes: how does the introduction of M_t into the step size affect suboptimality? In the non-smooth case the effect is minimal, as we anyway bound Q_t with $O(L^2t^3)$, and $M_t = O(L^2t^2)$ is of a lower order. In the smooth case, however, M_t is potentially more harmful, since while Proposition 3 allows us to cancel the dependence on Q_t by setting $s = c_t$, it leaves M_t hanging in the numerator, yielding $f(\hat{x}_t) - f(x_\star) \le O\left(\frac{1}{\alpha_t^2}\left(c_t^{3/2}\beta d_0^2 + c_t d_0\sqrt{M_t}\right)\right)$.

Fortunately, smoothness allows us to relate M_t back to the optimality gap $f(\hat{x}_t) - f(x_\star)$. In particular, in the unconstrained setting $\mathcal{K} = \mathbb{R}^n$ we have

$$\|m_t\|^2 \le 2\|g_t - m_t\|^2 + 2\|g_t\|^2 \le 2Q_t/\alpha_t^2 + 4\beta[f(\hat{x}_t) - f(x_\star)],$$

where the last transition used that $g_t = \nabla f(\hat{x}_t)$ in the noiseless setting. Combining this bound with Proposition 3, we obtain

$$f(\hat{x}_t) - f(x_\star) \le O\left(\frac{c_t^{3/2}\beta d_0^2 + \sqrt{c_t^2\beta d_0^2\max_{k\le t}\alpha_k^2[f(\hat{x}_k) - f(x_\star)]}}{\alpha_t^2}\right),$$

from which $f(\hat{x}_t) - f(x_\star) \le O\left(\frac{c_t^2\beta d_0^2}{\alpha_t^2}\right)$ follows by induction. Thus we arrive at our final guarantee in the noiseless case: Theorem 5 (see full proof in Section A.3).

Theorem 5 *In the noiseless setting (Assumption 2) with $\mathcal{K} = \mathbb{R}^n$ and $r_\epsilon \le d_0$, using the step sizes eq. (10), we get that $\bar{d}_T \le 2d_0$, $\bar{r}_T \le 4d_0$ and, for $\tau = \arg\max_{t<T}\sum_{i\le t}\frac{\bar{r}_i}{\bar{r}_{t+1}}$, the suboptimality is*

$$f(\hat{x}_\tau) - f(x_\star) \le O\left(c_{r_\epsilon,T}\min\left\{\frac{\beta d_0^2}{T^2}, \frac{Ld_0}{\sqrt{T}}\right\}\right),$$

where $c_{r_\epsilon,T} = \log_+^4\left(1 + \frac{T\min\{\beta d_0^2, Ld_0\}}{f(x_0) - f(x_\star)}\right)\log_+^2\left(\frac{d_0}{r_\epsilon}\right)$.

4. Analysis in the stochastic case

In this section, we extend the U-DoG guarantees to the noisy case. We start by assuming that the gradient noise is bounded, a setting that captures most of the remaining technical challenges. We then generalize our results to sub-Gaussian noise by means of a black-box reduction [3]. Finally, we specialize the U-DoG guarantee for mini-batches of bounded gradient estimates and conclude with a discussion of the (weak) dependence of our result on problem parameter bounds. Throughout this section, we denote the empirical variance at time t by

$$V_t \coloneqq \frac{1}{t+1}\sum_{k=0}^{t}\big(\|g_t - \nabla f(\hat{x}_t)\|^2 + \|m_t - \nabla f(\hat{z}_t)\|^2\big). \tag{11}$$

We also recall the notation

$$\theta_{t,\delta} \coloneqq \log\frac{60\log(6t)}{\delta}.$$

4.1. Analysis with bounded noise

We formalize the bounded noise assumption as follows.

Assumption 6 *In addition to Assumption 1, we assume that $\|\mathcal{G}(x) - \nabla f(x)\| \le \mathfrak{b}(x)$ with probability 1 for all $x \in \mathcal{K}$, for some (known[4]) function $\mathfrak{b} : \mathcal{K} \to \mathbb{R}_+$.*

For the iterates of U-DoG we define

$$\mathfrak{b}_t \coloneqq \mathfrak{b}(\hat{x}_t) \;\text{ and }\; \bar{\mathfrak{b}}_t \coloneqq \max\Big\{\max_{i\le t}\mathfrak{b}_i, \mathfrak{b}(\hat{z}_0)\Big\}. \tag{12}$$

With the assumption and notation in place, we state the stochastic equivalent of Proposition 3 in the following (see proof in Section B.1).

Proposition 7 *In the bounded noise setting (Assumption 6), suppose the* U-DoG *step sizes (2) satisfy $G_{x,t} \ge Q_{t-1}$ for every $t \ge 0$. Then for any $\mathfrak{B} > 0$, $T \in \mathbb{N}$, and $\delta \in (0,1)$, with probability at least $1 - \delta - \mathbb{P}[\bar{\mathfrak{b}}_{T-1} > \mathfrak{B}]$ we have, for all $t < T$ and $s \ge 0$,*

$$f(\hat{x}_t) - f(x_\star) \le O\left(\mathit{RHS}_{\text{eq. (4)}} + \frac{(1+s)(\bar{r}_{t+1}+d_0)\sqrt{t^3\theta_{t+1,\delta}V_t + (t\theta_{t+1,\delta}\mathfrak{B})^2}}{\big(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\big)^2}\right)$$

where $\mathit{RHS}_{\text{eq. (4)}} = \frac{s^{3/2}\beta(\bar{r}_{t+1}+d_0)^2 + (\bar{r}_{t+1}+d_0)\Big[\sqrt{\max\{G_{y,t},Q_t\}} - s\sqrt{Q_t}\Big]_+}{\big(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\big)^2}$ *as in Proposition 3.*

Proposition 7 is a fairly straightforward extension of its noiseless counterpart. The bound (5) continues to hold if we replace Q_t with $\hat{Q}_t = \sum_{k=0}^t \alpha_k^2 \min\{\|g_k - m_k\|^2, \|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2\}$. Proceeding as in the proof of Proposition 3, we conclude that

$$f(\hat{x}_t) - f(x_\star) \le O\left(\mathrm{RHS}_{\text{eq. (4)}} + \frac{s(\bar{r}_{t+1}+d_0)\big(Q_t^{1/2} - 2\hat{Q}_t^{1/2}\big) + \sum_{k=0}^t \omega_k\langle\nabla f(\hat{x}_k) - g_k, x_{k+1} - x_\star\rangle}{\big(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\big)^2}\right).$$

4. We may view $\mathfrak{b}$ as a coarse upper bound on the true noise magnitude, as it only affects low order terms in our bounds.

We show that $Q_t^{1/2} - 2\hat{Q}_t^{1/2} \leq O\left(\sqrt{t^3 V_t}\right)$ by straightforward manipulation. Furthermore, using a time-uniform empirical-Bernstein-type concentration bound [26, 27] (Lemma 21) to show that (with the appropriate high probability) the martingale difference sum $\sum_{k=0}^{t} \omega_k \langle \nabla f(\hat{x}_k) - g_k, x_{k+1} - x_\star \rangle$ is bounded by $O\left(\bar{r}_t \bar{d}_{t+1} \sqrt{t^3 \theta_{t+1,\delta} V_t + (t\theta_{t+1,\delta}\mathfrak{B})^2}\right)$.

Next, we extend our iterate stability guarantee to the stochastic setting (see proof in section B.2).

Proposition 8 *In the bounded noise setting Assumption 6, let $s > 0$, $T \in \mathbb{N}$ and $\delta \in (0,1)$, and define $c_t = 400\theta_{T,\delta} \log_+^2\left(\frac{s+Q_t}{s}\right)$. Suppose that $r_\epsilon \leq d_0$ and the* U-DoG *step sizes (2) satisfy, with probability 1, for all $t \geq 0$: (i) $G_{y,t} \geq c_t^2(s + Q_t)$ (with $G_{x,0} \geq 400^2\theta_{T,\delta}^2 s$), (ii) $\max\{\|x_{t+1} - y_t\|, \|y_{t+1} - x_{t+1}\|\} \leq \frac{2\bar{r}_t}{c_{t-1}}$, (iii) $\sqrt{G_{y,t}} \geq c_t \alpha_t \max\{\|\nabla f(\hat{x}_t) - g_t\|, \|\nabla f(\hat{x}_t) - m_t\|\}$, and (iv) $\eta_{y,t}$ is independent of g_t given $x_0, \ldots, x_t$. Then, we have with probability of at least $1 - \delta$,*

$$\bar{d}_t \leq 2d_0 \;\; \textit{and} \;\; \bar{r}_t \leq 4d_0 \;\; \textit{for all } t < T.$$

Conditions (i) and (ii) of Proposition 8 are identical to their noiseless counterparts in Proposition 4, while conditions (iii) and (iv) are new and facilitate the application of a concentration bound to the weighed regret $\mathcal{R}'_t$ defined in eq. (9). In particular, the condition (iv) ensures that $\sum_{k=0}^{t} \eta_{y,k}\alpha_k \langle g_k - \nabla f(\hat{x}_k), x_{k+1} - x_\star \rangle$ is a martingale difference sequence, and condition (iii) guarantees boundedness required by our concentration bound (Lemma 22). With this high probability bound in place, the proof continues in the same vein as the noiseless case.

When searching for step sizes meeting the conditions of Proposition 8 we encounter two challenges. First, condition (iii) asks $G_{y,t}$ to be large compared to a quantity depending on the exact gradient $\nabla f(\hat{x}_t)$, which we cannot access directly. We solve it using the bounds given in (12). Simply adding $c_t^2(t+1)^2\bar{\mathfrak{b}}_t^2 \geq c_t^2\alpha_t^2\mathfrak{b}_t^2$ to $G_{y,t}$ guarantees that $\sqrt{G_{y,t}} \geq c_t\alpha_t\|\nabla f(\hat{x}_t) - g_t\|$. Moreover, using $\|u\|^2 + \|v\|^2 \geq \frac{1}{2}\|v+u\|^2$, we have

$$\|g_t - m_t\|^2 + \bar{\mathfrak{b}}_t^2 \geq \|g_t - m_t\|^2 + \|\nabla f(\hat{x}_t) - g_t\|^2 \geq \frac{1}{2}\|\nabla f(\hat{x}_t) - m_t\|^2.$$

Therefore, taking $G_{y,t} = c_t^2(s + 2Q_t + 2(t+1)^2\bar{\mathfrak{b}}_t^2)$ fulfills condition (iii). However, it violates condition (iv) which leads us to the second challenge: how to avoid dependence on g_t? To address this challenge, we employ the somewhat unusual trick of drawing a *fresh stochastic gradient* $\tilde{g}_t \sim \mathcal{G}(\hat{x}_t)$ which is, by construction, independent of g_t given $\hat{x}_t$. We can now replace the forbidden $\|g_t - m_t\|$ with the valid upper bound $2\|\tilde{g}_t - m_t\| + 8\bar{\mathfrak{b}}_t$ and thus satisfy conditions (i) and (iii) without violating condition (iv).

To satisfy condition (ii) we introduce M_t to $G_{y,t}$ as done in the noiseless setting and make another modification to ensure the monotonicity required in (2). Writing,

$$\tilde{q}_t \coloneqq 2\alpha_t^2\|\tilde{g}_t - m_t\|^2 \;\;,\;\; \bar{Q}_t \coloneqq \sum_{k=0}^{t} \max\{q_k, \tilde{q}_k\} \;\text{ and }\; p_t \coloneqq 8(t+1)^2\bar{\mathfrak{b}}_t^2, \tag{13}$$

our final step sizes are:

$$\begin{aligned}
\eta_{x,t} &= \frac{\bar{r}_t}{400\theta_{T,\delta}\log_+^2\left(1 + \frac{p_{t-1} + \bar{Q}_{t-1}}{\|m_0\|^2 + p_0}\right)\sqrt{\max\left\{\|m_0\|^2 + p_0 + p_{t-1} + \bar{Q}_{t-1}, M_t\right\}}} \\
\eta_{y,t} &= \frac{\bar{r}_t}{400\theta_{T,\delta}\log_+^2\left(1 + \frac{p_t + \tilde{q}_t + \bar{Q}_{t-1}}{\|m_0\|^2 + p_0}\right)\sqrt{\max\left\{\|m_0\|^2 + p_0 + p_t + \tilde{q}_t + \bar{Q}_{t-1}, M_t\right\}}}.
\end{aligned} \tag{14}$$

Similar to the T-DoG step sizes [27, Section 3.3], our step sizes depend logarithmically on the desired confidence level δ and double-logarithmically on the maximum iteration budget T.

With all the pieces in place, we now state our main result (see proof in Section B.3).

Theorem 9 *In the bounded noise setting (Assumption 6) with $\mathcal{K} = \mathbb{R}^n$, for any $T \in \mathbb{N}$ and $\delta \in (0, \frac{1}{5})$, consider* U-DoG *with step sizes (14). With probability at least $1-5\delta$, we have $\bar{d}_T \leq 2d_0$, $\bar{r}_T \leq 4d_0$ and for $\tau = \arg\max_{t<T} \sum_{i\leq t} \frac{\bar{r}_i}{\bar{r}_{t+1}}$ and $\mathfrak{b}_\star := \max_{x:\|x-x_\star\|\leq 2d_0}\{\mathfrak{b}(x)\}$ we have*

$$f(\hat{x}_\tau) - f(x_\star) \leq O\left(c_{\delta,r_\epsilon,T}\left(\min\left\{\frac{\beta d_0^2}{T^2}, \frac{Ld_0}{\sqrt{T}}\right\} + \frac{d_0\sqrt{V_T}}{\sqrt{T}} + \frac{d_0\mathfrak{b}_\star}{T}\right)\right), \tag{15}$$

where $c_{\delta,r_\epsilon,T} = \log^2\left(\frac{\log_+(T)}{\delta}\right)\log_+^4\left(1 + T\frac{\mathfrak{b}_\star + \min\{\beta d_0^2, Ld_0\}}{f(x_0)-f(x_\star)}\right)\log_+^2\left(\frac{d_0}{r_\epsilon}\right)$ and V_t, defined in eq. (11), is the empirical noise variance.

We remark that under our assumptions it is straightforward to replace the empirical variance V_t in eq. (15) with its expectation without altering other non-logarithmic terms in the bound, e.g., via Hoeffding's inequality.

4.2. From bounded to sub-Gaussian noise

The bounded noise assumption makes analysis convenient but is not entirely satisfactory since averaging independent bounded-noise estimators does not reduce the probability 1 noise bound, preventing us from making statements about mini-batch scaling. To address this issue, we consider the following standard assumption.

Assumption 10 *In addition to Assumption 1, we assume that $\|\mathcal{G}(x) - \nabla f(x)\|$ is $\sigma^2(x)$-sub-Gaussian for all $x \in \mathcal{K}$, for some (known) $\sigma : \mathcal{K} \to \mathbb{R}_+$. That is,*

$$\mathbb{P}(\|\mathcal{G}(x) - \nabla f(x)\| \geq z) \leq 2\exp\left(-z^2/\sigma^2(x)\right)$$

for all $z \geq 0$ and $x \in \mathcal{K}$.

To move from bounded to sub-Gaussian we utilize a reduction due to Attia and Koren [3] that allows us to essentially replace $\mathfrak{b}(\cdot)$ with $\sigma(\cdot)$ in Theorem 9 at the cost of additional logarithmic factors. To that end, we define $\bar{\sigma}_t := \max\{\max_{i\leq t}\sigma(\hat{x}_k), \sigma(\hat{z}_0)\}$, as well as $\sigma_\star := \max_{x:\|x-x_\star\|\leq 2d_0}\sigma(x)$ and $\varsigma_{t,\delta} := 3\log^{1/2}\left(\frac{15(t+1)^2}{\delta}\right)$. With this notation in hand, we state our guarantee for the sub-Gaussian setting (see proof in Section B.4).

Corollary 11 *Consider the sub-Gaussian noise setting (Assumption 10) with $\mathcal{K} = \mathbb{R}^n$ and $\delta \in (0, \frac{1}{6})$, using the step sizes (14) with $\bar{\mathfrak{b}}_t = \bar{\sigma}_t\varsigma_{t,\delta}$, then with probability at least $1-6\delta$ we get that $\bar{d}_T \leq 2d_0$, $\bar{r}_T \leq 4d_0$, and the suboptimality bound (15) holds for $\mathfrak{b}_\star = \sigma_\star\varsigma_{T-1,\delta}$.*

4.3. Corollary: mini-batch of bounded noise

Finally, we leverage our result for sub-Gaussian noise to demonstrate that U-DoG automatically benefits from increasing mini-batch size (see proof in Section B.5).

Assumption 12 *In addition to Assumption 1, we assume that $\mathcal{G}(x)$ is the average of B unbiased estimates of $\nabla f(x)$, each bounded by L with a known upper bound $\hat{L} \geq L$.*

Corollary 13 *In the mini-batch setting (Assumption 12) with $\mathcal{K} = \mathbb{R}^n$, for any $T \in \mathbb{N}$ and $\delta \in (0, \frac{1}{6})$, consider* U-DoG *with step sizes (14) where $\bar{\mathfrak{b}}_t = \sqrt{2}\frac{\hat{L}}{\sqrt{B}}\varsigma_{t,\delta}$. With probability at least $1-6\delta$ we have $\bar{d}_T \leq 2d_0$, $\bar{r}_T \leq 4d_0$ and, for $\tau = \arg\max_{t<T} \sum_{i\leq t} \frac{\bar{r}_i}{\bar{r}_{t+1}}$,*

$$f(\hat{x}_\tau) - f(x_\star) \leq O\left(c_{\delta,r_\epsilon,T}\left(\frac{\beta d_0^2}{T^2} + \frac{\left(L + \hat{L}/\sqrt{T}\right)d_0}{\sqrt{TB}}\right)\right),$$

where $c_{\delta,r_\epsilon,T} = \sqrt{\log_+\left(\frac{T}{\delta}\right)}\log^2\left(\frac{\log_+(T)}{\delta}\right)\log_+^4\left(1 + T\frac{\hat{L}+\min\left\{\beta d_0^2, Ld_0\right\}}{f(x_0)-f(x_\star)}\right)\log_+^2\left(\frac{d_0}{r_\epsilon}\right)$.

4.4. Discussion: how parameter-free is our algorithm?

With our results established, we now discuss in more detail the extent to which our algorithms and complexity bounds are free of a-priori knowledge of problem parameters. U-DoG requires a lower bound r_ϵ on the initial distance to the optimum d_0, and pointwise upper bounds $\mathfrak{b}_t$ on the noise magnitude at each iteration. Theorem 9 provides suboptimality bounds that depend poly-logarithmically on $\frac{d_0}{r_\epsilon}$ which quantifies how r_ϵ underestimates d_0. Many works [e.g., 7, 9, 15, 16, 27, 28, 39, 41, 49] treat such logarithmic dependence as the *definition* of a parameter-free algorithm, and in that strict sense our method is certainly parameter-free. The noise bounds impact our suboptimality guarantees polynomially via the additive term $\mathfrak{b}_\star/T$ where $\mathfrak{b}_\star := \max_{x:\|x-x_\star\|\leq 2d_0}\{\mathfrak{b}(x)\}$, potentially implying greater sensitivity to problem parameters. Neverthless, we argue that our method fully deserve the title "parameter-free" for the following reasons.

1. **The noise bound only contributes a low-order error term.** To see why $\mathfrak{b}_\star/T$ is low-order, let $\hat{\mathfrak{b}}_T$ be the largest stochastic gradient error in the first T iterations of U-DoG. Then the empirical variance satisfies $V_T = O\left(\hat{\mathfrak{b}}_T^2\right)$ and the noise-dependent part of Theorem 9 is $\widetilde{O}\left(\frac{d_0\hat{\mathfrak{b}}_T}{\sqrt{T}} + \frac{d_0\mathfrak{b}_\star}{T}\right)$. Therefore, as long as $\mathfrak{b}_\star/\hat{\mathfrak{b}}_T = O(\sqrt{T})$ (i.e. the noise bound is loose by $\sqrt{T}$ or less) we get the near-optimal dependence on the unknown, true noise magnitude $\hat{\mathfrak{b}}_T$.

2. **The low-order term and noise bound assumption are unavoidable.** Recent work [4, Theorem 6] proves that *any* algorithm with logarithmic dependence on uncertainty in distance to optimality *must* suffer the low-order error term $\mathfrak{b}_\star/T$, and hence also require an a-priori noise bound (concurrent work [10, 31] also shows similar results). Moreover, prior parameter-free algorithms assume known bounds on stochastic gradient magnitude, which is stronger than assuming noise bounds. In this sense, our method is as parameter-free as it gets.

3. **The noise bound is often easy to obtain and vanishes as batch size grows.** Corollary 13 and Assumption 12 give a general setting where a noise bound is readily available. For a concrete instantiation, consider logistic regression with normalized covariates. In this case $\hat{L} = 1$ and the noise bound at batch size B is $\widetilde{O}(1/B)$, which decreases as the batch size grows.

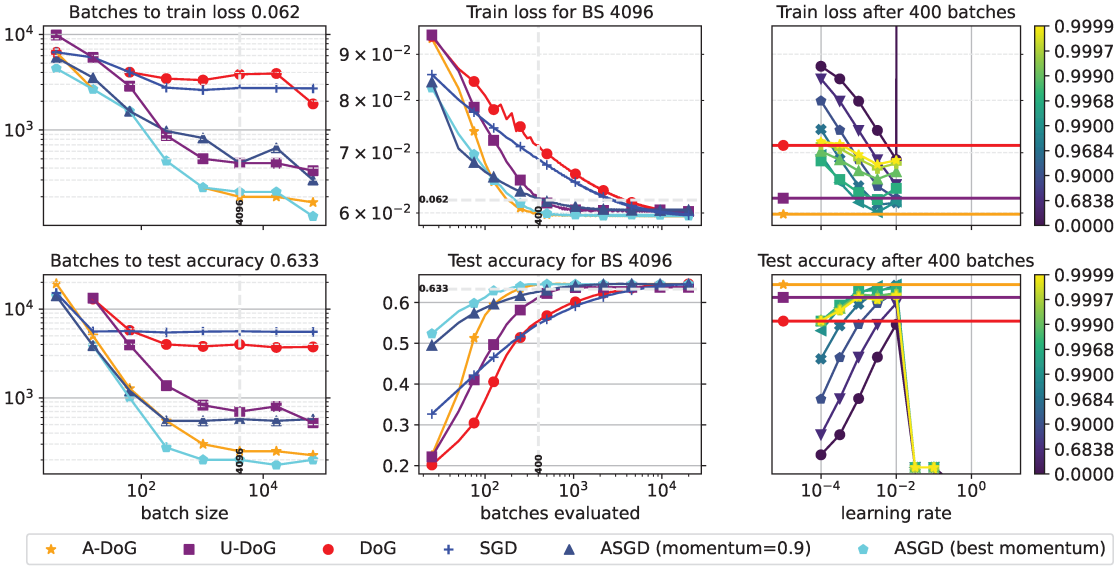

Figure 1: Training a linear model with ViT-32 features and least-squares loss on SVHN. Top: Train loss. Bottom: Test accuracy after iterate averaging. First column: Batch size scaling of complexity to reach target performance. Second column: Learning curves. Third column: ASGD performance at all learning rates and momenta, contrasted with DoG variants.

5. Experiments

We test U-DoG on a suite of experiments on convex and non-convex learning problems. We also heuristically derive and experiment with an algorithm we call A-DoG, which integrates ideas from AcceleGrad [36] and DoG. Namely, it uses the AcceleGrad step with DoG numerator and α_t as in U-DoG. The pseudocode for A-DoG is given in Algorithm 2 in Section G.2.

We compare our algorithms to DoG as well as carefully tuned SGD with constant Nesterov momentum (ASGD for short) across a wide range of batch sizes. Detailed experimental results and analyses, as well as implementation details, are presented in Appendix G.

Our testbed consists of multiple classification problems based on the VTAB benchmark [67] and libsvm datasets [12], which we solve with both multiclass log loss and least squares loss, as well as a synthetic noiseless linear regression problem (see Section G.3). In addition, we perform preliminary experiments in the non-convex setting, including training neural networks from scratch on CIFAR-10 and VTAB datasets, and fine-tuning a CLIP model on ImageNet (see Section G.4).

On convex optimization problems, both U-DoG and A-DoG often substantially improve over DoG, with A-DoG achieving results comparable to well-tuned ASGD and outperforming U-DoG, likely by avoiding extra-gradient computations. Figure 1 illustrates these results on a particular dataset and least-squares loss function configuration and Section G.3 repeats this figure for additional configurations. The left panels in the figure show the rate of convergence of A-DoG, U-DoG and ASGD plateaus at a larger batch size compared to DoG and SGD without momentum. This is the typical effect of acceleration in stochastic optimization [56], and is also supported by Corollary 13 which shows that, for sufficiently large batch size, U-DoG converges at rate scaling as $1/T^2$. In contrast, non-accelerated methods like DoG and SGD converge with rate scaling as $1/T$. The right panels of the figure show that, at a tight computational budget, the performance

of ASGD is very sensitive to the tuning of both step size and momentum, with only the very best values matching the performance of A-DoG. When using logarithmic instead of least-squares loss, the test accuracy becomes more robust to large step size choices (see Figure 2 in the appendix). This is partly because the log loss is Lispchitz which prevents complete divergence at any fixed step size.

In our preliminary non-convex experiments on neural network models (reported in detail in Sections G.3 and G.4), we find that U-DoG often fails to converge to competitive results, while A-DoG is competitive with DoG on most VTAB tasks, but under-performs it for CIFAR-10 and ImageNet fine-tuning, indicating that it is not a yet a viable general-purpose neural network optimizer.

Acknowledgments

We thank Konstantin Mishchenko for helpful discussion. This work was supported by the NSF-BSF program, under NSF grant #2239527 and BSF grant #2022663. MI acknowledges support from the Israeli Council of Higher Education. OH acknowledges support from Pitt Momentum Funds, and AFOSR grant #FA955023-1-0242. YC acknowledges support from the Israeli Science Foundation (ISF) grant no. 2486/21 and the Alon Fellowship.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015.

[2] E. Alpaydin and Fevzi. Alimoglu. Pen-Based Recognition of Handwritten Digits. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C5MG6K.

[3] Amit Attia and Tomer Koren. SGD with AdaGrad stepsizes: Full adaptivity with high probability to unknown parameters, unbounded gradients and affine variance. In *International Conference on Machine Learning (ICML)*, 2023.

[4] Amit Attia and Tomer Koren. How free is parameter-free stochastic optimization? In *International Conference on Machine Learning (ICML)*, 2024.

[5] Charles Beattie, Joel Z Leibo, Denis Teplyashin, Tom Ward, Marcus Wainwright, Heinrich Küttler, Andrew Lefrancq, Simon Green, Víctor Valdés, Amir Sadik, et al. Deepmind lab. *arXiv:1612.03801*, 2016.

[6] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

[7] Aditya Bhaskara, Ashok Cutkosky, Ravi Kumar, and Manish Purohit. Online learning with imperfect hints. In *International Conference on Machine Learning (ICML)*, 2020.

[8] Jock Blackard. Covertype. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50K5N.

[9] Yair Carmon and Oliver Hinder. Making SGD parameter-free. In *Conference on Learning Theory (COLT)*, 2022.

[10] Yair Carmon and Oliver Hinder. The price of adaptivity in stochastic convex optimization. In *Conference on Learning Theory (COLT)*, 2024.

[11] Yair Carmon, Danielle Hausler, Arun Jambulapati, Yujia Jin, and Aaron Sidford. Optimal and adaptive monteiro-svaiter acceleration. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

[12] Chih-Chung Chang and Chih-Jen Lin. LIBSVM: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2011.

[13] Gong Cheng, Junwei Han, and Xiaoqiang Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[14] Ashok Cutkosky. Anytime online-to-batch, optimism and acceleration. In *International Conference on Machine Learning (ICML)*, pages 1446–1454, 2019.

[15] Ashok Cutkosky. Artificial constraints and hints for unbounded online learning. In *Conference on Learning Theory (COLT)*, 2019.

[16] Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Conference on Learning Theory (COLT)*, 2018.

[17] Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by D-adaptation. In *International Conference on Machine Learning (ICML)*, 2023.

[18] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[19] Jelena Diakonikolas and Lorenzo Orecchia. Accelerated extra-gradient descent: A novel accelerated first-order method. In *Innovations in Theoretical Computer Science (ITCS)*, 2018.

[20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021.

[21] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12(7), 2011.

[22] Vineet Gupta, Tomer Koren, and Yoram Singer. A unified approach to adaptive regularization in online and stochastic optimization. *arXiv:1706.06569*, 2017.

[23] Charles R. Harris, K. Jarrod Millman, Stéfan J. van der Walt, Ralf Gommers, Pauli Virtanen, David Cournapeau, Eric Wieser, Julian Taylor, Sebastian Berg, Nathaniel J. Smith, Robert Kern, Matti Picus, Stephan Hoyer, Marten H. van Kerkwijk, Matthew Brett, Allan Haldane, Jaime Fernández del Río, Mark Wiebe, Pearu Peterson, Pierre Gérard-Marchant, Kevin Sheppard, Tyler Reddy, Warren Weckesser, Hameer Abbasi, Christoph Gohlke, and Travis E. Oliphant. Array programming with NumPy. *Nature*, 585(7825):357–362, 2020.

[24] Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[25] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020.

[26] Steven R Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

[27] Maor Ivgi, Oliver Hinder, and Yair Carmon. DoG is SGD's best friend: A parameter-free dynamic step size schedule. In *International Conference on Machine Learning (ICML)*, 2023. We refer to the latest arXiv version: https://arxiv.org/abs/2302.12022.

[28] Andrew Jacobsen and Ashok Cutkosky. Parameter-free mirror descent. In *Conference on Learning Theory (COLT)*, 2022.

[29] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. CLEVR: A diagnostic dataset for compositional language and elementary visual reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[30] Ali Kavis, Kfir Y Levy, Francis Bach, and Volkan Cevher. UniXGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[31] Ahmed Khaled and Chi Jin. Tuning-free stochastic optimization. In *International Conference on Machine Learning (ICML)*, 2024.

[32] Ahmed Khaled, Konstantin Mishchenko, and Chi Jin. DoWG unleashed: An efficient universal parameter-free gradient descent method. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

[33] Diederik P Kingma and Jimmy Ba. ADAM: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[34] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[35] Guanghui Lan. An optimal method for stochastic composite optimization. *Mathematical Programming*, 133(1):365–397, 2012.

[36] Kfir Y Levy, Alp Yurtsever, and Volkan Cevher. Online adaptive methods, universality and acceleration. *Advances in Neural Information Processing Systems (NeurIPS)*, 2018.

[37] Ilya Loshchilov and Frank Hutter. SGDR: Stochastic gradient descent with warm restarts. *International Conference on Learning Representations*, 2017.

[38] H Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *The Journal of Machine Learning Research*, 18(1):3117–3166, 2017.

[39] H Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in Hilbert spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory (COLT)*, 2014.

[40] H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *arXiv:1002.4908*, 2010.

[41] Zakaria Mhammedi and Wouter M Koolen. Lipschitz and comparator-norm adaptivity in online learning. In *Conference on Learning Theory (COLT)*, 2020.

[42] Konstantin Mishchenko and Aaron Defazio. Prodigy: An expeditiously adaptive parameter-free learner. *arXiv:2306.06101*, 2023.

[43] Arkadi Nemirovski. Prox-method with rate of convergence $o(1/t)$ for variational inequalities with Lipschitz continuous monotone operators and smooth convex-concave saddle point problems. *SIAM Journal on Optimization*, 15(1):229–251, 2004.

[44] Yurii Nesterov. A method of solving a convex programming problem with convergence rate $O(1/k^2)$. *Soviet Mathematics Doklady*, 27(2):372–376, 1983.

[45] Yurii Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

[46] Yuval Netzer, Tao Wang, Adam Coates, A. Bissacco, Bo Wu, and A. Ng. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning 2011*, 2011.

[47] Francesco Orabona. Dimension-free exponentiated gradient. *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[48] Francesco Orabona. A modern introduction to online learning. *arXiv:1912.13213*, 2021.

[49] Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[50] Courtney Paquette and Katya Scheinberg. A stochastic line search method with expected complexity analysis. *SIAM Journal on Optimization*, 30(1):349–376, 2020.

[51] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[52] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[53] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

[54] Sasha Rakhlin and Karthik Sridharan. Optimization, learning, and games with predictable sequences. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2013.

[55] Sashank J Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of Adam and beyond. In *International Conference on Learning Representations (ICLR)*, 2018.

[56] Christopher J Shallue, Jaehoon Lee, Joseph Antognini, Jascha Sohl-Dickstein, Roy Frostig, and George E Dahl. Measuring the effects of data parallelism neural network training. *Journal of Machine Learning Research*, 20:1–49, 2019.

[57] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *International Conference on Machine Learning (ICML)*, 2013.

[58] Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning (ICML)*, 2018.

[59] Matthew Streeter and H Brendan McMahan. No-regret algorithms for unconstrained online convex optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2012.

[60] Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[61] Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, C J Carey, İlhan Polat, Yu Feng, Eric W. Moore, Jake VanderPlas, Denis Laxalde, Josef Perktold, Robert Cimrman, Ian Henriksen, E. A. Quintero, Charles R. Harris, Anne M. Archibald, Antônio H. Ribeiro, Fabian Pedregosa, Paul van Mulbregt, and SciPy 1.0 Contributors. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272, 2020.

[62] Wes McKinney. Data Structures for Statistical Computing in Python. In *Proceedings of the 9th Python in Science Conference*, 2010.

[63] Ross Wightman. PyTorch image models. https://github.com/rwightman/pytorch-image-models, 2019.

[64] Jianxiong Xiao, James Hays, Krista A Ehinger, Aude Oliva, and Antonio Torralba. Sun database: Large-scale scene recognition from abbey to zoo. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[65] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119(1):3–22, 2016.

[66] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016.

[67] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv:1910.04867*, 2019.

Contents

Appendix A. Proof for Section 3 (the noiseless setting)

A.1. Proof of Proposition 3

Proof Define

$$\rho_t := \frac{1}{\sqrt{Q_t}} \text{ and}$$
$$\hat{Q}_t := \sum_{k=0}^{t} \alpha_k^2 \min\left\{\|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2, \|g_k - m_k\|^2\right\}.$$

Note that in the noiseless setting $\hat{Q}_T = Q_T$. However, most of the proof carries over to the noisy setting as well. Therefore, until a later stage of the proof, we do not use that $m_t = \nabla f(\hat{z}_t)$, $g_t = \nabla f(\hat{x}_t)$ and $\hat{Q}_t = Q_t$ in the noiseless setting.

Recall the notation $\tilde{\eta}_{x,t} = \frac{1}{\sqrt{G_{x,t}}}$ and $\tilde{\eta}_{y,t} = \frac{1}{\sqrt{G_{y,t}}}$. Algebraic manipulation gives us that for all $k \geq 0$

$$\begin{aligned}\bar{r}_k \alpha_k \langle g_k, x_{k+1} - x_\star \rangle \leq \frac{\bar{r}_k^2 \alpha_k^2 \rho_k}{2} \|g_k - m_k\|^2 - \sum_{k=0}^{t} \frac{1}{2\rho_k} \|x_{k+1} - y_k\|^2 \\ + \left(\frac{1}{2\rho_k} - \frac{1}{2\tilde{\eta}_{x,k}}\right)\left(\|x_{k+1} - y_k\|^2 + \|x_{k+1} - y_{k+1}\|^2\right) \\ + \frac{1}{2\tilde{\eta}_{y,k}}\left(\|x_\star - y_k\|^2 - \|x_\star - y_{k+1}\|^2\right);\end{aligned}$$

see Lemma 16 for a proof. Therefore, by summing over both sides of the inequality we get that for all $t \geq 0$

$$\begin{aligned}\sum_{k=0}^{t} \bar{r}_k \alpha_k \langle g_k, x_{k+1} - x_\star \rangle \leq \underbrace{\frac{\bar{r}_t^2}{2} \sum_{k=0}^{t} \frac{\alpha_k^2 \|g_k - m_k\|^2}{\sqrt{\sum_{j=0}^{k} \alpha_j^2 \|g_j - m_j\|^2}}}_{(A)} \underbrace{- \sum_{k=0}^{t} \frac{1}{2\rho_k} \|x_{k+1} - y_k\|^2}_{(B)} \\ \underbrace{+ 4\bar{r}_{t+1}^2 \sum_{k=0}^{t} \left[\frac{1}{\rho_k} - \frac{1}{\tilde{\eta}_{x,k}}\right]_+}_{(C)} \underbrace{+ \frac{1}{2} \sum_{k=0}^{t} \frac{1}{\tilde{\eta}_{y,k}} \left(d_k^2 - d_{k+1}^2\right)}_{(D)}.\end{aligned}$$

Bounding (A)**:** We have $\sum_{k=0}^{t} \frac{\alpha_k^2 \|g_k - m_k\|^2}{\sqrt{\sum_{j=0}^{k} \alpha_j^2 \|g_j - m_j\|^2}} \leq 2\sqrt{\sum_{k=0}^{t} \alpha_k^2 \|g_k - m_k\|^2}$; see Lemma 28 with $s_k = \alpha_k^2 \|g_k - m_k\|^2$, and therefore

$$\frac{\bar{r}_t^2}{2} \sum_{k=0}^{t} \frac{\alpha_k^2 \|g_k - m_k\|^2}{\sqrt{\sum_{j=0}^{k} \alpha_j^2 \|g_j - m_j\|^2}} = \frac{\bar{r}_t^2}{\rho_t}.$$

Bounding (B)**:** We have that for all $k \geq 0$

$$\begin{aligned}\|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2 &\overset{(1)}{\leq} \beta^2 \|\hat{x}_k - \hat{z}_k\|^2 \\ &= \frac{\beta^2 \bar{r}_k^2 \alpha_k^2}{\left(\sum_{0=1}^k \bar{r}_i \alpha_i\right)^2} \|x_{k+1} - y_k\|^2 \\ &\overset{(2)}{\leq} \frac{4\beta^2}{\alpha_k^2} \|x_{k+1} - y_k\|^2,\end{aligned}$$

where (1) is from the β-smoothness of f, and (2) is because $\bar{r}_k \alpha_k^2 \leq 2 \sum_{0=1}^k \bar{r}_i \alpha_i$ by Lemma 25 . Therefore,

$$-\|x_{k+1} - y_k\|^2 \leq -\frac{\alpha_k^2 \|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2}{4\beta^2}.$$

Thus,

$$-\sum_{k=0}^t \frac{1}{2\rho_k} \|x_{k+1} - y_k\|^2 \leq -\sum_{k=0}^t \frac{\alpha_k^2 \|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2}{8\beta^2 \rho_k}$$

Bounding (C)**:** As $\frac{1}{\rho_k} - \frac{1}{\tilde{\eta}_{x,k}}$ is not necessarily non-negative for all $k \in \{0, \ldots, t\}$ we define the set of indices for which it is non-negative as

$$I \triangleq \left\{k \in \{0, 1, \ldots, t\} \ : \ \frac{1}{\rho_t} - \frac{1}{\tilde{\eta}_{x,t}} \geq 0\right\}.$$

Define i_k as the k-th smallest index in I, and define $i_{|I|+1} \coloneqq t + 1$. We note that for all $k \in I$ then $i_k \leq i_{k+1} - 1 \leq t$. Therefore,

$$\begin{aligned}(C) &= 4\bar{r}_{t+1}^2 \sum_{k=1}^{|I|} \left(\frac{1}{\rho_{i_k}} - \frac{1}{\tilde{\eta}_{x,i_k}}\right) \leq 4\bar{r}_{t+1}^2 \sum_{k=1}^{|I|} \left(\frac{1}{\rho_{[i_{k+1}-1]}} - \frac{1}{\tilde{\eta}_{x,i_k}}\right) \\ &\leq \frac{4\bar{r}_{t+1}^2}{\rho_t} + 4\bar{r}_{t+1}^2 \sum_{k=2}^{|I|} \left(\frac{1}{\rho_{[i_k-1]}} - \frac{1}{\tilde{\eta}_{x,i_k}}\right) \leq \frac{4\bar{r}_{t+1}^2}{\rho_t} + 4\bar{r}_{t+1}^2 \sum_{k=0}^{t-1} \left[\frac{1}{\rho_k} - \frac{1}{\tilde{\eta}_{x,k+1}}\right]_+.\end{aligned}$$

Bounding (D)**:**

$$\begin{aligned}\frac{1}{2} \sum_{k=0}^t \frac{1}{\tilde{\eta}_{y,t}} \left(d_k^2 - d_{k+1}^2\right) &= \frac{d_0^2}{2\tilde{\eta}_{y,0}} - \frac{d_{t+1}^2}{2\tilde{\eta}_{y,t}} + \frac{1}{2} \sum_{k=0}^t \left(\frac{1}{\tilde{\eta}_{y,k}} - \frac{1}{\tilde{\eta}_{y,k-1}}\right) d_k^2 \\ &\leq \frac{d_0^2}{2\tilde{\eta}_{y,0}} - \frac{d_{t+1}^2}{2\tilde{\eta}_{y,t}} + \frac{\bar{d}_t^2}{2} \sum_{k=0}^t \left(\frac{1}{\tilde{\eta}_{y,k}} - \frac{1}{\tilde{\eta}_{y,k-1}}\right).\end{aligned}$$

By performing telescopic summation we obtain

$$(D) \leq \frac{\bar{d}_{t+1}^2 - d_{t+1}^2}{2\tilde{\eta}_{y,t}}.$$

Let $s \in \arg\max_{k \le t+1} d_k$, we have that $\bar{d}_{t+1}^2 - d_{t+1}^2 = \bar{d}_s^2 - d_{t+1}^2 = \big(\bar{d}_s - d_{t+1}\big)\big(\bar{d}_s + d_{t+1}\big) \le \|y_s - y_{t+1}\|\big(\bar{d}_s + \bar{d}_{t+1}\big) \le (\bar{r}_s + r_{t+1})\big(\bar{d}_s + d_{t+1}\big) \le 4\bar{r}_{t+1}\bar{d}_{t+1}$. Thus,

$$(D) \le \frac{2\bar{r}_{t+1}\bar{d}_{t+1}}{\tilde{\eta}_{y,t}}.$$

Bounding $(A)+(B)+(C)+(D)$**:** Combining all of the above, we obtain that

$$\sum_{k=0}^{t} \bar{r}_k \alpha_k \langle g_k, x_{k+1} - x_\star \rangle \le 5\bar{r}_{t+1}\big(\bar{r}_{t+1} + \bar{d}_{t+1}\big) \max\left\{\frac{1}{\rho_t}, \frac{1}{\tilde{\eta}_{y,t}}\right\} + 4\bar{r}_{t+1}^2 \sum_{k=0}^{t-1}\left[\frac{1}{\rho_k} - \frac{1}{\tilde{\eta}_{x,k+1}}\right]_+ - \sum_{k=0}^{t} \frac{\alpha_k^2 \|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2}{8\beta^2\rho_k}.$$

Therefore, as for any we have that $G_{x,k} \ge Q_{k-1}$,

$$\sum_{k=0}^{t} \bar{r}_k \alpha_k \langle g_k, x_{k+1} - x_\star \rangle \le 5\bar{r}_{t+1}\big(\bar{r}_{t+1} + \bar{d}_{t+1}\big)\sqrt{\max\{G_{y,t}, Q_t\}} - \sum_{k=0}^{t} \frac{\alpha_k^2 \|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2}{8\beta^2\rho_k}.$$

Let $s \ge 0$ and recall that $\frac{1}{\rho_k} = \sqrt{Q_k}$. We get that

$$\begin{aligned}\sum_{k=0}^{t} \bar{r}_k \alpha_k \langle g_k, x_{k+1} - x_\star \rangle &\le 10s\bar{r}_{t+1}\big(\bar{r}_{t+1} + \bar{d}_{t+1}\big)\sqrt{\hat{Q}_t} - \sum_{k=0}^{t} \frac{\alpha_k^2 \|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2}{8\beta^2}\sqrt{Q_k} \\ &\quad + 5\bar{r}_{t+1}\big(\bar{r}_{t+1} + \bar{d}_{t+1}\big)\left(s\sqrt{Q_t} - 2s\sqrt{\hat{Q}_t}\right) \\ &\quad + 5\bar{r}_{t+1}\big(\bar{r}_{t+1} + \bar{d}_{t+1}\big)\left(\sqrt{\max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t}\right). \end{aligned} \tag{16}$$

We have that

$$\begin{aligned}&10s\bar{r}_{t+1}\big(\bar{r}_{t+1} + \bar{d}_{t+1}\big)\sqrt{\hat{Q}_t} - \sum_{k=0}^{t} \frac{\alpha_t^2 \|\nabla f(\hat{x}_t) - \nabla f(\hat{z}_t)\|^2}{8\beta^2}\sqrt{Q_k} \\ &\quad \le 10s\bar{r}_{t+1}\big(\bar{r}_{t+1} + \bar{d}_{t+1}\big)\sqrt{\hat{Q}_t} - \sum_{k=0}^{t} \frac{\alpha_k^2 \min\big\{\|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2, \|g_k - m_k\|^2\big\}}{8\beta^2}\sqrt{\hat{Q}_k}.\end{aligned}$$

Define $B_k^2 = \alpha_k^2 \min\big\{\|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2, \|g_k - m_k\|^2\big\}$, $c_1 = 10s\bar{r}_{t+1}(\bar{r}_{t+1} + \bar{d}_{t+1})$, and $c_2 = 8\beta^2$. Lemma 27 gives us that for all $t \ge 0$

$$c_1\sqrt{\sum_{k=0}^{t} B_k^2} - \sum_{k=0}^{t} \frac{B_k^2}{c_2}\sqrt{\sum_{j=0}^{k} B_j^2} \le 2c_1^{3/2}c_2^{1/2}.$$

Therefore,

$$\begin{aligned}&10s\bar{r}_{t+1}\big(\bar{r}_{t+1} + \bar{d}_{t+1}\big)\sqrt{\hat{Q}_t} - \sum_{k=0}^{t} \frac{\alpha_t^2 \|\nabla f(\hat{x}_t) - \nabla f(\hat{z}_t)\|^2}{8\beta^2}\sqrt{Q_k} \\ &\quad \le 2\big(10s\bar{r}_{t+1}\big(\bar{r}_{t+1} + \bar{d}_{t+1}\big)\big)^{3/2}(8\beta)^{1/2} \le 180s^{3/2}\bar{r}_{t+1}\big(\bar{r}_{t+1} + \bar{d}_{t+1}\big)^2\beta.\end{aligned}$$

Combining this result with eq. (16) yields that for all $t \geq 0$ and $s \geq 0$

$$\begin{aligned}\sum_{k=0}^{t} \bar{r}_k \alpha_k \langle g_k, x_{k+1} - x_\star \rangle \leq\ & 180 s^{3/2} \bar{r}_{t+1} \big(\bar{r}_{t+1} + \bar{d}_{t+1}\big)^2 \beta \\ & + 5 \bar{r}_{t+1} \big(\bar{r}_{t+1} + \bar{d}_{t+1}\big) \left(s\sqrt{Q_t} - 2s\sqrt{\hat{Q}_t} \right) \\ & + 5 \bar{r}_{t+1} \big(\bar{r}_{t+1} + \bar{d}_{t+1}\big) \left(\sqrt{\max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t} \right). \end{aligned} \tag{17}$$

Lemma 14 gives us that

$$f(\hat{x}_t) - f(x_\star) \leq \frac{1}{\sum_{k=0}^{t} \bar{r}_k \alpha_k} \sum_{k=0}^{t} \bar{r}_k \alpha_k \langle \nabla f(\hat{x}_k), x_{k+1} - x_\star \rangle .$$

Now, by additionally using the fact that in the noiseless setting

$$\hat{Q}_t = Q_t \ \text{ and}$$
$$\sum_{k=0}^{t} \bar{r}_k \alpha_k \langle \nabla f(\hat{x}_k), x_{k+1} - x_\star \rangle = \sum_{k=0}^{t} \bar{r}_k \alpha_k \langle g_k, x_{k+1} - x_\star \rangle$$

we get that

$$\begin{aligned} f(\hat{x}_t) - f(x_\star) \leq\ & 180 s^{3/2} \frac{\bar{r}_{t+1}}{\sum_{k=0}^{t} \bar{r}_k \alpha_k} \beta \big(\bar{r}_{t+1} + \bar{d}_{t+1}\big)^2 \\ & + 5 \frac{\bar{r}_{t+1}}{\sum_{k=0}^{t} \bar{r}_k \alpha_k} \big(\bar{r}_{t+1} + \bar{d}_{t+1}\big) \left(\sqrt{\max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t} \right). \end{aligned}$$

Finally, by using the fact that $\bar{d}_{t+1} \leq d_0 + \bar{r}_{t+1}$ and because $\bar{r}_k \alpha_k^2 \leq 2 \sum_{0=1}^{k} \bar{r}_i \alpha_i$ for all $k \geq 0$ (Lemma 25), we obtain that

$$f(\hat{x}_t) - f(x_\star) \leq O\left(\frac{s^{3/2} \beta (\bar{r}_{t+1} + d_0)^2 + (\bar{r}_{t+1} + d_0) \big[\sqrt{\max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t} \big]_+}{\big(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\big)^2} \right).$$

■

A.2. Proof of Proposition 4

Proof For any $h > 0$ (in this case $h = 12$), define

$$c_t \coloneqq h \log_+^2 \left(\frac{s + Q_t}{s} \right) \ \text{ and}$$
$$\rho_t \coloneqq \frac{1}{c_{t-1} \sqrt{s + Q_t}}.$$

Lemma 16 gives us that, for all $t \geq 0$,

$$\bar{r}_t \alpha_t \langle g_t, x_{t+1} - x_\star \rangle \leq \frac{\bar{r}_t^2 \alpha_t^2 \rho_t}{2} \|g_t - m_t\|^2 + \left(\frac{1}{2\rho_t} - \frac{1}{2\tilde{\eta}_{x,t}} \right) \left(\|x_{t+1} - y_t\|^2 + \|x_{t+1} - y_{t+1}\|^2 \right) + \frac{1}{2\tilde{\eta}_{y,t}} \left(\|x_\star - y_t\|^2 - \|x_\star - y_{t+1}\|^2 \right).$$

From the definitions of ρ_t and $\tilde{\eta}_{x,t} = 1/\sqrt{G_{x,t}} \leq 1/\sqrt{G_{y,t-1}} \leq 1/\rho_{t-1}$ we obtain that

$$\frac{1}{2\rho_t} - \frac{1}{2\tilde{\eta}_{x,t}} \leq \frac{c_{t-1}^2}{2} \rho_t (Q_t - Q_{t-1});$$

See proof in Lemma 26. Now, because we also have that $\max\{\|x_{t+1} - y_t\|, \|y_{t+1} - x_{t+1}\|\} \leq \frac{2\bar{r}_t}{c_{t-1}}$, we get

$$\bar{r}_t \alpha_t \langle g_t, x_{t+1} - x_\star \rangle \leq \frac{9}{2} \bar{r}_t^2 \rho_t \alpha_t^2 \|g_t - m_t\|^2 + \frac{1}{2\tilde{\eta}_{y,t}} \left(d_t^2 - d_{t+1}^2 \right).$$

Thus,

$$2\tilde{\eta}_{y,t} \bar{r}_t \alpha_t \langle g_t, x_{t+1} - x_\star \rangle \leq 9 \bar{r}_t^2 \tilde{\eta}_{y,t} \rho_t \alpha_t^2 \|g_t - m_t\|^2 + \left(d_t^2 - d_{t+1}^2 \right).$$

Consequentially, by summing the two sides of the inequality, we get that for all $t \geq 0$

$$\begin{aligned} 2 \sum_{k=0}^{t} \tilde{\eta}_{y,k} \bar{r}_k \alpha_k \langle g_k, x_{k+1} - x_\star \rangle &\leq 9 \sum_{k=0}^{t} \bar{r}_k^2 \tilde{\eta}_{y,k} \rho_k \alpha_k^2 \|g_k - m_k\|^2 + \sum_{k=0}^{t} \left(d_k^2 - d_{k+1}^2 \right) \\ &\leq \frac{9 \bar{r}_t^2}{h^2} \sum_{k=0}^{t} \frac{Q_k - Q_{k-1}}{(s + Q_k) \log_+^2 \left(\frac{s+Q_k}{s} \right)} + \sum_{k=0}^{t} \left(d_k^2 - d_{k+1}^2 \right). \end{aligned}$$

Lemma 30 gives us that

$$\sum_{k=0}^{t} \frac{Q_k - Q_{k-1}}{(s + Q_k) \log_+^2 \left(\frac{s+Q_k}{s} \right)} \leq 1.$$

Therefore, we obtain that

$$2 \sum_{k=0}^{t} \tilde{\eta}_{y,k} \bar{r}_k \alpha_k \langle g_k, x_{k+1} - x_\star \rangle \leq \frac{9 \bar{r}_t^2}{h^2} + \sum_{k=0}^{t} \left(d_k^2 - d_{k+1}^2 \right).$$

Thus,

$$\begin{aligned} 2 \sum_{k=0}^{t} \tilde{\eta}_{y,k} \bar{r}_k \alpha_k \langle \nabla f(\hat{x}_k), x_{k+1} - x_\star \rangle \leq & \frac{9 \bar{r}_t^2}{h^2} + 2 \sum_{k=0}^{t} \tilde{\eta}_{y,k} \bar{r}_k \alpha_k \langle \nabla f(\hat{x}_k) - g_k, x_{k+1} - x_\star \rangle \\ & + \sum_{k=0}^{t} \left(d_k^2 - d_{k+1}^2 \right). \end{aligned}$$

Consequentially, as Lemma 17 gives us that

$$\sum_{k=0}^{t} \tilde{\eta}_{y,k} \bar{r}_k \alpha_k \langle \nabla f(\hat{x}_k), x_{k+1} - x_\star \rangle \geq 0,$$

we get that

$$0 \leq \frac{9\bar{r}_t^2}{h^2} + 2\sum_{k=0}^{t} \eta_{y,k} \alpha_k \langle \nabla f(\hat{x}_k) - g_k, x_{k+1} - x_\star \rangle + \sum_{k=0}^{t} \left(d_k^2 - d_{k+1}^2\right).$$

Therefore, we get that for all $t \geq 0$

$$d_{t+1}^2 \leq \frac{9\bar{r}_t^2}{h^2} + 2\sum_{k=0}^{t} \eta_{y,k} \alpha_k \langle \nabla f(\hat{x}_k) - g_k, x_{k+1} - x_\star \rangle + d_0^2. \tag{18}$$

As we are in the noiseless case, and $h = 12$, we get that for all $t \geq 0$

$$\begin{aligned} d_{t+1}^2 &\leq \frac{\bar{r}_t^2}{16} + d_0^2 \\ &\leq \left(d_0 + \frac{1}{4}\bar{r}_t\right)^2. \end{aligned}$$

Finally, Lemma 18 now gives us that for all $t \geq 0$

$$d_t \leq 2d_0 \;\text{ and }\; r_t \leq 4d_0.$$

■

A.3. Proof of Theorem 5

Proof Define

$$c_t = 12 \log_+^2 \left(\frac{\|m_0\|^2 + Q_t}{\|m_0\|^2} \right).$$

From Lemma 19, we get that for all $t \geq 0$ the distance between iterates is not large:

$$\max\{\|x_{t+1} - y_t\|, \|x_{t+1} - y_{t+1}\|\} \leq \frac{2\bar{r}_t}{c_{t-1}}.$$

Now, we fulfill all the conditions for Proposition 4 and therefore, for all $t \geq 0$

$$\bar{d}_t \leq 2d_0 \;\text{ and }\; \bar{r}_t \leq 4d_0.$$

Proposition 3 gives that for all $t \geq 0$ and for all $s \geq 0$

$$f(\hat{x}_t) - f(x_\star) \leq O\left(\frac{s^{3/2}\beta(\bar{r}_{t+1} + d_0)^2 + (\bar{r}_{t+1} + d_0)\left[\sqrt{G_{y,t}} - s\sqrt{Q_t}\right]_+}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2} \right).$$

By using the fact that $\bar{r}_t \leq 4d_0$, we get that for all $t \geq 0$

$$f(\hat{x}_t) - f(x_\star) \leq O\left(\frac{s^{3/2}\beta d_0^2 + d_0\left[\sqrt{G_{y,t}} - s\sqrt{Q_t}\right]_+}{\left(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\right)^2}\right). \tag{19}$$

Recall that

$$\tau = \arg\max_{t<T} \sum_{i\leq t} \frac{\bar{r}_i}{\bar{r}_{t+1}}.$$

To show the non-smooth rate, we set $s = 0$ and obtain

$$\sqrt{G_{y,\tau}} \leq c_T \sqrt{\max_{k\leq T-1}\left\{\alpha_k^2\|m_k\|^2\right\} + \sum_{k=0}^{T-1}\alpha_k^2\|g_k - m_k\|^2} \leq c_T\sqrt{T^2L^2 + T^3L^2} \leq 2LT^{3/2}c_T.$$

This result, with eq. (19), gives us that

$$f(\hat{x}_\tau) - f(x_\star) \leq O\left(\frac{Ld_0T^{3/2}}{\left(\sum_{k=0}^\tau \bar{r}_k/\bar{r}_{t+1}\right)^2}c_T\right). \tag{20}$$

To show the smooth rate, setting $s = 2c_{t+1}$ yields

$$\sqrt{G_{y,t}} - s\sqrt{Q_t} \leq c_{t+1}\left(\sqrt{Q_t + M_t} - 2\sqrt{Q_t}\right) \leq c_{t+1}\left(\sqrt{M_t} - \sqrt{Q_t}\right).$$

For some $\kappa_t \leq t$ we have that $\sqrt{M_t} = \alpha_{\kappa_t}\|m_{\kappa_t}\|$. In addition, the smoothness of f implies that $\|\nabla f(z)\|^2 \leq 2\beta[f(z) - f(x_\star)]$ for all $z \in \mathcal{X}$. Combining this fact with the triangle inequality gives us that, in the noiseless setting,

$$\alpha_{\kappa_t}\|m_{\kappa_t}\| = \alpha_{\kappa_t}\|\nabla f(\hat{z}_{\kappa_t})\| \leq \alpha_{\kappa_t}\|\nabla f(\hat{x}_{\kappa_t}) - \nabla f(\hat{z}_{\kappa_t})\| + \alpha_{\kappa_t}\sqrt{2\beta}\sqrt{f(\hat{x}_{\kappa_t}) - f(x_\star)}.$$

Thus,

$$\sqrt{M_t} \leq \sqrt{Q_t} + \alpha_{\kappa_t}\sqrt{2\beta}\sqrt{f(\hat{x}_{\kappa_t}) - f(x_\star)}.$$

Therefore,

$$\sqrt{G_{y,t}} - s\sqrt{Q_t} \leq \alpha_{\kappa_t}\sqrt{2c_{t+1}^2\beta}\sqrt{f(\hat{x}_{\kappa_t}) - f(x_\star)}.$$

This result, together with eq. (19), give us that for all $t \geq 0$, there exist $\kappa_t \leq t$ such as

$$f(\hat{x}_t) - f(x_\star) \leq O\left(\frac{c_{t+1}^{3/2}\beta d_0^2 + \alpha_{\kappa_t}\sqrt{c_{t+1}^2\beta d_0^2}\sqrt{f(\hat{x}_{\kappa_t}) - f(x_\star)}}{\left(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\right)^2}\right).$$

Using the previous inequality and Lemma 15 we obtain that for all $t \geq 0$ that

$$f(\hat{x}_t) - f(x_\star) \leq O\left(\frac{\beta d_0^2}{\left(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\right)^2}c_{t+1}^2\right). \tag{21}$$

Combining the result from eq. (20) and eq. (21) gives

$$f(\hat{x}_\tau) - f(x_\star) \leq O\left(\frac{\min\{\beta d_0^2, L d_0 T^{3/2}\}}{\left(\sum_{k=0}^{\tau} \bar{r}_k/\bar{r}_{t+1}\right)^2} c_T^2\right). \tag{22}$$

Lemma 29 gives us that

$$\sum_{k=0}^{\tau} \bar{r}_k/\bar{r}_{t+1} \geq \frac{1}{e}\left(\frac{T}{\log_+(\bar{r}_T/r_\epsilon)} - 1\right).$$

Thus, if $T \geq 2\log_+(\bar{r}_T/r_\epsilon)$ then

$$\frac{1}{\sum_{k=0}^{\tau} \bar{r}_k/\bar{r}_{t+1}} \leq O\left(\frac{1}{T}\log_+\left(\frac{\bar{r}_T}{r_\epsilon}\right)\right).$$

Therefore, from eq. (22), we obtain

$$f(\hat{x}_\tau) - f(x_\star) \leq O\left(\frac{\min\{\beta d_0^2, L d_0 T^{3/2}\}}{T^2} c_T^2 \log_+^2\left(\frac{\bar{r}_T}{r_\epsilon}\right)\right). \tag{23}$$

We have that

$$\begin{aligned} c_T &\leq O\left(\log_+^2\left(\frac{\|m_0\|^2 + Q_{T-1}}{\|m_0\|^2}\right)\right) \\ &\overset{(i)}{\leq} O\left(\log_+^2\left(1 + \frac{T^3 \min\{\beta d_0, L\}}{\|\nabla f(\hat{z}_0)\|^2}\right)\right) \leq O\left(\log_+^2\left(1 + T\frac{\min\{\beta d_0^2, L d_0\}}{f(x_0) - f(x_\star)}\right)\right), \end{aligned}$$

due to (i) the noiseless setting and f being β-smooth and L-Lipschitz, and (ii) convexity, which implies $f(x_0) - f(x_\star) \leq d_0\|\nabla f(\hat{z}_0)\|$ Finally, from eq. (23), we obtain

$$f(\hat{x}_\tau) - f(x_\star) \leq O\left(\frac{\min\{\beta d_0^2, L d_0 T^{3/2}\}}{T^2} \log_+^4\left(1 + T\frac{\min\{\beta d_0^2, L d_0\}}{f(x_0) - f(x_\star)}\right) \log_+^2\left(\frac{d_0}{r_\epsilon}\right)\right). \tag{24}$$

Finally, for $T \leq 2\log_+(\bar{r}_T/r_\epsilon)$ the theorem holds trivially since $f(\hat{x}_\tau) - f(x_\star) \leq \min\{\beta \bar{d}_\tau^2, L\bar{d}_\tau\}$ and $\bar{d}_\tau \leq 2d_0$ by Proposition 8. Therefore,

$$f(\hat{x}_\tau) - f(x_\star) \leq O\left(\min\{\beta d_0^2, L d_0\}\right) \leq O\left(\frac{\min\{\beta \bar{d}_0^2, L\bar{d}_0\}}{T^2} \log_+^2(\bar{r}_T/r_\epsilon)\right),$$

and so the bound Equation (24) holds in all cases, concluding the proof. ■

Appendix B. Proofs for Section 4 (the stochastic setting)

B.1. Proof of Proposition 7

Proof Define

$$\hat{Q}_t := \sum_{k=0}^{t} \alpha_k^2 \min\{\|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2, \|g_k - m_k\|^2\}.$$

Our proof continues from eq. (17) in the proof Proposition 3, which also holds for stochastic gradients.

$$\begin{aligned}\sum_{k=0}^{t} \bar{r}_k \alpha_k \langle g_k, x_{k+1} - x_\star \rangle \leq\ & 180 s^{3/2} \bar{r}_{t+1} \left(\bar{r}_{t+1} + \bar{d}_{t+1}\right)^2 \beta \\ & + 5\bar{r}_{t+1}\left(\bar{r}_{t+1} + \bar{d}_{t+1}\right)\left(s\sqrt{Q_t} - 2s\sqrt{\hat{Q}_t}\right) \\ & + 5\bar{r}_{t+1}\left(\bar{r}_{t+1} + \bar{d}_{t+1}\right)\left(\sqrt{\max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t}\right).\end{aligned}$$

For all $k \geq 0$

$$\begin{aligned}\|g_k - m_k\|^2 &\leq 2\|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2 + 2\|(g_k - \nabla f(\hat{x}_k)) - (m_k - \nabla f(\hat{z}_k))\|^2 \\ &\leq 2\min\{\|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2, \|g_k - m_k\|^2\} + 4\|m_t - \nabla f(\hat{z}_t)\|^2 + 4\|g_t - \nabla f(\hat{x}_t)\|^2.\end{aligned}$$

Thus, for all $k \geq 0$

$$\|g_k - m_k\|^2 \leq 2\min\{\|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2, \|g_k - m_k\|^2\} + 4\|m_t - \nabla f(\hat{z}_t)\|^2 + 4\|g_t - \nabla f(\hat{x}_t)\|^2.$$

Multiplying by α_k^2, summing and recalling that $\alpha_k \leq k+1$ implies $Q_t \leq 2\hat{Q}_t + 4(t+1)^3 V_t$, where $V_t = \frac{1}{t+1}\sum_{k=0}^{t}\left(\|g_t - \nabla f(\hat{x}_t)\|^2 + \|m_t - \nabla f(\hat{z}_t)\|^2\right)$ is the empirical variance. Substituting into eq. (17), we get that

$$\begin{aligned}\sum_{k=0}^{t} \bar{r}_k \alpha_k \langle g_k, x_{k+1} - x_\star \rangle \leq\ & 180 s^{3/2} \bar{r}_{t+1} \left(\bar{r}_{t+1} + \bar{d}_{t+1}\right)^2 \beta \\ & + 5\bar{r}_{t+1}\left(\bar{r}_{t+1} + \bar{d}_{t+1}\right)\left(\sqrt{\max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t}\right) \\ & + 10 s\bar{r}_{t+1}\left(\bar{r}_{t+1} + \bar{d}_{t+1}\right)\sqrt{(t+1)^3 V_t}. \qquad (25)\end{aligned}$$

Lemma 21 gives us that with probability of at least $1-\delta-\mathbb{P}\left[\bar{\mathfrak{b}}_{T-1} > \mathfrak{B}\right]$, for all $t \in \{0, 1, \ldots, T-1\}$,

$$\left|\sum_{k=0}^{t} \bar{r}_t \alpha_k \langle \nabla f(\hat{x}_k) - g_k, x_{k+1} - x_\star \rangle\right| \leq 8\alpha_t \bar{r}_t (\bar{r}_{t+1} + d_0)\sqrt{\theta_{t+1,\delta}\sum_{k=0}^{t}\|\nabla f(\hat{x}_k) - g_k\|^2 + (\theta_{t+1,\delta}\mathfrak{B})^2}.$$

Using the previous equality and the definition of V_t we obtain that

$$\begin{aligned}
&\sum_{k=0}^{t} \bar{r}_k \alpha_k \left\langle \nabla f(\hat{x}_k), x_{k+1} - x_\star \right\rangle \\
&\qquad = \sum_{k=0}^{t} \bar{r}_k \alpha_k \left\langle g_k, x_{k+1} - x_\star \right\rangle + \sum_{k=0}^{t} \bar{r}_k \alpha_k \left\langle \nabla f(\hat{x}_k) - g_k, x_{k+1} - x_\star \right\rangle \\
&\qquad \leq \sum_{k=0}^{t} \bar{r}_k \alpha_k \left\langle g_k, x_{k+1} - x_\star \right\rangle + 8\alpha_t \bar{r}_t (\bar{r}_{t+1} + d_0) \sqrt{(t+1)\theta_{t+1,\delta} V_t + (\theta_{t+1,\delta}\mathfrak{B})^2}. \quad (26)
\end{aligned}$$

Lemma 14 gives us that

$$f(\hat{x}_t) - f(x_\star) \leq \frac{1}{\sum_{k=0}^{t} \bar{r}_k \alpha_k} \sum_{k=0}^{t} \bar{r}_k \alpha_k \left\langle \nabla f(\hat{x}_k), x_{k+1} - x_\star \right\rangle .$$

By combining the above inequality with eq. (25) and eq. (26), we obtain

$$\begin{aligned}
f(\hat{x}_t) - f(x_\star) \leq\ & 180 s^{3/2} \frac{\bar{r}_{t+1}}{\sum_{k=0}^{t} \bar{r}_k \alpha_k} \beta \left(\bar{r}_{t+1} + \bar{d}_{t+1}\right)^2 \\
& + 5 \frac{\bar{r}_{t+1}}{\sum_{k=0}^{t} \bar{r}_k \alpha_k} \left(\bar{r}_{t+1} + \bar{d}_{t+1}\right) \left(\sqrt{\max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t} \right) \\
& + 10(1+s) \frac{\bar{r}_{t+1}}{\sum_{k=0}^{t} \bar{r}_k \alpha_k} \left(\bar{r}_{t+1} + \bar{d}_{t+1}\right) \sqrt{(t+1)^3 V_t + (\theta_{t+1,\delta}\mathfrak{B})^2}.
\end{aligned}$$

Now, as Lemma 25 gives us that $\bar{r}_t \alpha_t^2 \leq 2 \sum_{k=0}^{t} \bar{r}_k \alpha_k$, we obtain that

$$\begin{aligned}
f(\hat{x}_t) - f(x_\star) \leq\ & 360 s^{3/2} \frac{1}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2} \beta \left(\bar{r}_{t+1} + \bar{d}_{t+1}\right)^2 \\
& + 10 \frac{1}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2} \left(\bar{r}_{t+1} + \bar{d}_{t+1}\right) \left(\sqrt{\max\{G_{y,t}, Q_t\}} - s\sqrt{Q_t} \right) \\
& + 20 \frac{1}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2} \left(\bar{r}_{t+1} + \bar{d}_{t+1}\right) \sqrt{(t+1)^3 V_t + (\theta_{t+1,\delta}\mathfrak{B})^2}.
\end{aligned}$$

Finally, because that $\bar{d}_{t+1} \leq d_0 + \bar{r}_{t+1}$, we get that for any $\mathfrak{B} > 0$ with probability of at least $1 - \delta - \mathbb{P}[\bar{\mathfrak{b}}_{T-1} > \mathfrak{B}]$ we have that for all $t < T$ and for any number $s \geq 0$

$$f(\hat{x}_t) - f(x_\star) \leq O\left(\text{RHS}_{eq.\ (4)} + \frac{(1+s)(\bar{r}_{t+1} + d_0) \sqrt{t^3 \theta_{t,\delta} V_t + (t\theta_{t,\delta}\mathfrak{B})^2}}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2} \right)$$

where $\text{RHS}_{eq.\ (4)} = \frac{s^{3/2}\beta(\bar{r}_{t+1}+d_0)^2 + (\bar{r}_{t+1}+d_0)\left[\sqrt{\max\{G_{y,t},Q_t\}} - s\sqrt{Q_t}\right]_+}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2}$ is the error term appearing in Proposition 3. ■

B.2. Proof of Proposition 8

Proof The proof continues from eq. (18) in the proof of Proposition 4, which also holds for stochastic gradients. Substituting $h = 400$ in eq. (18) gives, for all $t \geq 0$,

$$d_{t+1}^2 \leq \frac{9\bar{r}_t^2}{400^2} + 2\sum_{k=0}^{t} \eta_{y,k}\alpha_k \left\langle \nabla f(\hat{x}_k) - g_k, x_{k+1} - x_\star \right\rangle + d_0^2.$$

Now, Lemma 22 gives us that with probability at least $1-\delta$, for all $t < T$

$$\begin{aligned}\left|\sum_{k=0}^{t} \alpha_k \eta_{y,k} \left\langle g_k - \nabla f(\hat{x}_k), x_{k+1} - x_\star \right\rangle\right| &\leq \frac{12\theta_{t+1,\delta}}{400\theta_{T,\delta}} \bar{r}_t(\bar{r}_{t+1} + d_0) \\ &\leq \frac{12\theta_{t+1,\delta}}{400\theta_{T,\delta}} (\bar{r}_t\bar{r}_{t+1} + \bar{r}_t d_0) \leq \frac{12}{400}\left(1 + \frac{3}{400}\right)\bar{r}_t^2 + \frac{12}{400}\bar{r}_t d_0.\end{aligned}$$

Therefore,

$$d_{t+1}^2 \leq \frac{81\bar{r}_t^2}{400^2} + \frac{24}{400}\bar{r}_t^2 + \frac{24}{400}\bar{r}_t d_0 + d_0^2 \leq \frac{\bar{r}_t^2}{16} + \frac{\bar{r}_t d_0}{2} + d_0.$$

Thus, with probability of at least $1-\delta$, for all $t < T$

$$d_{t+1}^2 \leq \left(d_0 + \frac{1}{4}\bar{r}_t\right)^2.$$

Finally, Lemma 18 gives us that with probability of at least $1-\delta$ for all $t < T$

$$d_t \leq 2d_0 \;\text{ and }\; r_t \leq 4d_0.$$

■

B.3. Proof of Theorem 9

Proof Recall the notation

$$\tilde{q}_t \coloneqq 2\alpha_t^2\|\tilde{g}_t - m_t\|^2 \;,\; \bar{Q}_t \coloneqq \sum_{k=0}^{t} \max\{q_k, \tilde{q}_k\} \;\text{ and }\; p_t \coloneqq 8(t+1)^2\bar{\mathfrak{b}}_t^2,$$

and that our step sizes are of the form (2) with

$$G_{y,t} = \hat{c}_t^2 \max\left\{\|m_0\|^2 + p_0 + p_t + \tilde{q}_t + \bar{Q}_{t-1}, M_t\right\},$$

where

$$\hat{c}_t = 400\theta_{T,\delta}\log_+^2\left(1 + \frac{p_t + \tilde{q}_t + \bar{Q}_{t-1}}{\|m_0\|^2 + p_0}\right).$$

We begin by verifying the conditions of Proposition 8 with $s = \|m_0\|^2 + p_0$, where condition (iv) holds by construction. By Assumption 6 we have

$$\|g_t - \tilde{g}_t\|^2 \leq 2\|g_t - \nabla f(\hat{x}_t)\|^2 + 2\|\tilde{g}_t - \nabla f(\hat{x}_t)\|^2 \leq 4\bar{\mathfrak{b}}_t^2.$$

Therefore, since $t+1 \geq \alpha_t$, we have

$$\tilde{q}_t + p_t \geq \alpha_t^2\big(2\|\tilde{g}_t - m_t\|^2 + 2\|g_t - \tilde{g}_t\|^2\big) \geq \alpha_t^2\|g_t - m_t\|^2 = q_t,$$

and consequently

$$\tilde{q}_t + p_t + \bar{Q}_{t-1} \geq Q_t.$$

Defining

$$c_t = 400\theta_{T,\delta}\log_+^2\left(1 + \frac{Q_t}{\|m_0\|^2 + p_0}\right),$$

we conclude that

$$G_{y,t} \geq c_t^2 \max\big\{\|m_0\|^2 + p_0 + Q_t, M_t\big\} \geq c_t^2(\|m_0\|^2 + p_0 + Q_t)$$

so that condition (i) of Proposition 8 holds. Next, since

$$G_{y,t} \geq c_t^2 \max\{Q_t, M_t\} \geq c_t^2\alpha_t^2 \max\{\|g_t - m_t\|^2, \|m_t\|^2\},$$

Lemma 19 guarantees condition (ii) of Proposition 8. Finally, we note that

$$p_t \geq 8\alpha_t^2 \max\{\|g_t - \nabla f(\hat{x}_t)\|^2, \|\tilde{g}_t - \nabla f(\hat{x}_t)\|^2\}$$

and

$$p_t + \tilde{q}_t \geq \alpha_t^2\big(2\|m_t - \tilde{g}_t\|^2 + 2\|\tilde{g}_t - \nabla f(\hat{x}_t)\|^2\big) \geq \alpha_t^2\|m_t - \nabla f(\hat{x}_t)\|^2.$$

Therefore, as $\sqrt{G_{y,t}} \geq c_t\sqrt{p_t + \tilde{q}_t}$, condition (iii) of Proposition 8 holds.

As all the conditions for Proposition 8 hold, with probability of at least $1-\delta$, for all $t \geq 0$

$$\bar{d}_t \leq 2d_0 \;\text{ and }\; \bar{r}_t \leq 4d_0.$$

Recalling that $\mathfrak{b}_\star := \max_{x:\|x-x_\star\|\leq 2d_0}\{\mathfrak{b}(x)\}$, this also implies that $\mathbb{P}[\bar{\mathfrak{b}}_{T-1} > \mathfrak{b}_\star] \leq \delta$.

We now combine the conclusions of Proposition 8 with Proposition 3 to obtain a suboptimality bound for U-DoG. Substituting $\mathbb{P}(\bar{r}_T \leq 4d_0) \leq \delta$ and $\mathbb{P}[\bar{\mathfrak{b}}_{T-1} > \mathfrak{b}_\star] \leq \delta$ into Proposition 7 we get that, with probability at least $1-3\delta$, for all $t < T$ and $s \geq 0$,

$$f(\hat{x}_t) - f(x_\star) \leq O\left(\frac{s^{3/2}\beta d_0^2 + d_0\big[\sqrt{G_{y,t}} - s\sqrt{Q_t}\big]_+ + (1+s)d_0\sqrt{t^3\theta_{t+1,\delta}V_t + (t\theta_{t+1,\delta}\mathfrak{b}_\star)^2}}{\big(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\big)^2}\right). \tag{27}$$

To simplify $G_{y,t}$ in the bound above, we invoke Lemma 23 which gives that, with probability at least $1-\delta-\mathbb{P}[\bar{\mathfrak{b}}_{T-1} > \mathfrak{b}_\star] \geq 1-2\delta$, for all $t < T$,

$$\bar{Q}_t \leq 5Q_t + 80(t+1)^3\sqrt{\theta_{t+1,\delta}}V_t + 2(t+1)^2\theta_{t+1,\delta}\mathfrak{b}_\star^2,$$

and hence

$$\sqrt{G_{y,t}} \leq \hat{c}_t\sqrt{\bar{Q}_t + 2M_t + 2p_t} = O\Big(\hat{c}_t\sqrt{Q_t} + \hat{c}_t\sqrt{M_t} + \hat{c}_t\theta_{T,\delta}\sqrt{t^3V_t + t^2\mathfrak{b}_\star^2}\Big).$$

Combining this with the bound (27) and replacing s with $s\hat{c}_t\sqrt{3}$, we get that with probability at least $1-5\delta$, for all $t<T$ and $s\geq 0$,

$$f(\hat{x}_t)-f(x_\star)\leq O\left(\frac{s^{3/2}\hat{c}_t^{3/2}\beta d_0^2+\hat{c}_t d_0\Big(\big[(1-s)\sqrt{Q_t}+\sqrt{M_t}\big]_+ +(1+s)\theta_{T,\delta}\sqrt{t^3V_t+t^2\mathfrak{b}_\star^2}\Big)}{\left(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\right)^2}\right). \tag{28}$$

The remainder of the proof parallels the proof of Theorem 5, where we specialize our bound to the Lipschitz and smooth cases by choosing different values of s. For the Lipschitz case, we use the facts that

$$Q_t\leq 4\sum_{k\leq T}\alpha_t^2\big(\|\nabla f(\hat{x}_k)\|^2+\|\nabla f(\hat{z}_k)\|^2+\|g_k-\nabla f(\hat{x}_k)\|^2+\|m_k-\nabla f(\hat{z}_k)\|^2\big)=O(L^2T^3+V_TT^3)$$

$Q_t=O(L^2T^3)$ and $M_t\leq O(L^2T^2)$ and (under the event $\bar{d}_T\leq 2d_0$)

$$M_t\leq\max_{k\leq T}\big\{2\alpha_t^2\big(\|\nabla f(\hat{z}_k)\|^2+\|m_k-\nabla f(\hat{z}_k)\|^2\big)\big\}=O(L^2T^2+\mathfrak{b}_\star^2T^2),$$

giving the suboptimality bound. Substituting these expression and $s=0$ into (28) we get, for all $t<T$,

$$f(\hat{x}_t)-f(x_\star)\leq O\left(\hat{c}_t\frac{Ld_0T^{3/2}+d_0\theta_{T,\delta}\sqrt{T^3V_T+T^2\mathfrak{b}_\star^2}}{\left(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\right)^2}\right). \tag{29}$$

For the smooth case and any $t<T$, let $\kappa_t\leq t$ be such that For some $\kappa_t\leq t$ we have that

$$\sqrt{M_t}=\alpha_{\kappa_t}\|m_{\kappa_t}\|.$$

The smoothness of f implies that $\|\nabla f(z)\|^2\leq 2\beta[f(z)-f(x_\star)]$ for all $z\in\mathcal{X}$. Combining this fact with the triangle inequality gives us that

$$\alpha_{\kappa_t}\|m_{\kappa_t}\|\leq\alpha_{\kappa_t}\|\nabla f(\hat{x}_{\kappa_t})-\nabla f(\hat{z}_{\kappa_t})\|+\alpha_{\kappa_t}\|m_{\kappa_t}-\nabla f(\hat{z}_{\kappa_t})\|+\alpha_{\kappa_t}\sqrt{2\beta}\sqrt{f(\hat{x}_{\kappa_t})-f(x_\star)}$$

and therefore,

$$\sqrt{M_t}\leq\sqrt{Q_t}+\sqrt{(t+1)^3V_t}+\alpha_{\kappa_t}\sqrt{2\beta}\sqrt{f(\hat{x}_{\kappa_t})-f(x_\star)}.$$

Substituting into eq. (28) and taking $s=2$, we get, for all $t<T$,

$$f(\hat{x}_t)-f(x_\star)\leq O\left(\frac{\hat{c}_t^{3/2}\beta d_0^2+\hat{c}_t d_0\theta_{T,\delta}\sqrt{T^3V_T+T^2\mathfrak{b}_\star^2}+\alpha_{\kappa_t}\sqrt{\hat{c}_{t+1}^2\beta d_0^2}\sqrt{f(\hat{x}_{\kappa_t})-f(x_\star)}}{\left(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\right)^2}\right).$$

Applying Lemma 15 and noting that $\theta_{T,\delta}\leq\hat{c}_t$ simplifies the bound to

$$f(\hat{x}_t)-f(x_\star)\leq O\left(\frac{\hat{c}_T^2\beta d_0^2+\hat{c}_T\theta_{T,\delta}d_0\sqrt{T^3V_T+T^2\mathfrak{b}_\star^2}}{\left(\sum_{k=0}^t \bar{r}_k/\bar{r}_{t+1}\right)^2}\right). \tag{30}$$

Combining the bounds eq. (29) and eq. (30) and noting that $\theta_{T,\delta} \leq \hat{c}_T$, we conclude that, with probability at least $1-5\delta$, for all $t < T$,

$$f(\hat{x}_t) - f(x_\star) \leq O\left(\hat{c}_T^2 \cdot \frac{\min\left\{\beta d_0^2, L d_0 T^{3/2}\right\} + d_0\sqrt{T^3 V_{T-1} + T^2 \mathfrak{b}_\star^2}}{\left(\sum_{k=0}^{\tau} \bar{r}_k/\bar{r}_{t+1}\right)^2}\right).$$

For $\tau = \arg\max_{t<T} \sum_{i\leq t} \frac{\bar{r}_i}{\bar{r}_{t+1}}$, Lemma 29 gives us that

$$\sum_{k=0}^{\tau} \bar{r}_k/\bar{r}_{t+1} \geq \frac{1}{e}\left(\frac{T}{\log_+(\bar{r}_T/r_\epsilon)} - 1\right).$$

Thus, for $T \geq 2\log_+(\bar{r}_T/r_\epsilon)$ we get (under the event $\bar{r}_T \leq 4d_0$)

$$f(\hat{x}_\tau) - f(x_\star) \leq O\left(\hat{c}_T^2 \log_+^2\left(\frac{d_0}{r_\epsilon}\right) \cdot \frac{\min\left\{\beta d_0^2, L d_0 T^{3/2}\right\} + d_0\sqrt{T^3 V_{T-1} + T^2 \mathfrak{b}_\star^2}}{T^2}\right),$$

which establishes the theorem, since

$$\begin{aligned}
\hat{c}_T &\overset{(i)}{\leq} O\left(\log_+^2\left(1 + \frac{T^2\mathfrak{b}_\star^2 + \bar{Q}_{T-1}}{\|\nabla f(\hat{z}_0)\|^2}\right)\right) \leq O\left(\log_+^2\left(1 + \frac{T^3\mathfrak{b}_\star^2 + T^3\sum_{k=0}^{T-1}\|\nabla f(\hat{x}_k) - \nabla f(\hat{z}_k)\|^2}{\|\nabla f(\hat{z}_0)\|^2}\right)\right) \\
&\leq O\left(\log_+^2\left(1 + \frac{T^3\mathfrak{b}_\star^2 + T^3\min\{\beta d_0, L\}}{\|\nabla f(\hat{z}_0)\|^2}\right)\right) \overset{(ii)}{\leq} O\left(\log_+^2\left(1 + \frac{T^3\mathfrak{b}_\star^2 d_0^2 + T^3\min\left\{\beta d_0^3, L d_0^2\right\}}{f(x_0) - f(x_\star)}\right)\right) \\
&= O\left(\log_+^2\left(1 + T\frac{\mathfrak{b}_\star d_0 + \min\left\{\beta d_0^2, L d_0\right\}}{f(x_0) - f(x_\star)}\right)\right),
\end{aligned}$$

where (i) is because $\|\nabla f(\hat{z}_0)\|^2 \leq \|\nabla f(\hat{z}_0) - m_0 + m_0\|^2 \leq 2\|m_0\|^2 + p_0$, and (ii) is from convexity: $f(x_0) - f(x_\star) \leq d_0\|\nabla f(\hat{z}_0)\|$.

Finally, when $T \leq 2\log_+(\bar{r}_T/r_\epsilon)$ the required bound is immediate from problem geometry, as explained at the end of the proof of Theorem 5. ■

B.4. Proof of Corollary 11

Proof Define

$$\delta_t' = \frac{\delta}{5(t+1)^2}.$$

A black-box reduction from sub-Gaussian to bounded stochastic gradient (Lemma 31) shows that at each iteration t, with probability at least $1-\delta_t'$, a call to a σ^2-sub-Gaussian subgradient oracle produces an identical result to a call to an alternative stochastic gradient that is bounded by $3\sigma\sqrt{\log(3/\delta_t')}$.

We apply Theorem 9 to U-DoG with the alternative, bounded stochastic gradient oracle. Thus, for this setting, with probability at least $1-5\delta$, we have $\bar{d}_T \leq 2d_0$, $\bar{r}_T \leq 4d_0$, and the suboptimality bound (15) holds for $\mathfrak{b}_\star = \sigma_\star\varsigma_{T-1,\delta}$. To conclude the proof we use Lemma 31 to show that the

algorithm described above produces output different than U-DoG with the original sub-Gaussian oracle as at most

$$3\sum_{t=0}^{\infty}\delta_t' \leq \frac{3\delta}{5}\sum_{t=1}^{\infty}\frac{1}{t^2} \leq \frac{3\cdot\pi^2}{5\cdot 6}\delta \leq \delta,$$

where the factor of 3 comes from the fact that every U-DoG iteration involves 3 stochastic gradient queries. ■

B.5. Proof of Corollary 13

Proof A mini-batch of B gradient oracle results, each with noise bounded by L, is a $\frac{2L^2}{B}$-sub-Gaussian (see Lemma 24), and we can therefore apply Corollary 11 with $\sigma_t^2 = \frac{2L^2}{B}$. Moreover, reusing the sub-Gaussian-to-bounded reduction in the proof of Corollary 11 (Section B.4) we get that, with probability at least $1-6\delta$,

$$\sqrt{V_T} \leq \frac{\sqrt{2}L}{\sqrt{B}}\varsigma_{T,\delta}$$

holds in addition to the suboptimality bound given by Corollary 11. Substituting the above bound on $\sqrt{V_T}$ along with $\mathfrak{b}_\star \leq \sqrt{2}\frac{L}{\sqrt{B}}\zeta_{T,\delta}$ concludes the proof. ■

Appendix C. Suboptimality lemmas

C.1. Weighted regret to suboptimality conversion (Lemma 14)

The following lemma is a straightforward reproduction of Lemma 1 from Kavis et al. [30] with minor changes. In addition, we use the proof of the following lemma as a starting point for the proof of Lemma 17.

Lemma 14 (Kavis et al. [30]) *For any sequence of positive numbers $\omega_0, \omega_1, \omega_2, \ldots$, define*

$$\hat{x}_t := \frac{\sum_{k=0}^{t}\omega_k x_{k+1}}{\sum_{k=0}^{t}\omega_k}.$$

We have that for any $T > 0$

$$f(\hat{x}_{T-1}) - f(x_\star) \leq \frac{1}{\sum_{t=0}^{T-1}\omega_t}\sum_{t=0}^{T-1}\omega_t\left\langle \nabla f(\hat{x}_t), x_{t+1} - x_\star\right\rangle.$$

Proof For any $t \geq 0$ we have that

$$\begin{aligned}
\omega_t \left\langle \nabla f(\hat{x}_t), x_{t+1} - x_\star \right\rangle &= \omega_t \left\langle \nabla f(\hat{x}_t), \frac{\sum_{k=0}^{t} \omega_k}{\omega_t} \hat{x}_t - \frac{\sum_{k=0}^{t-1} \omega_k}{\omega_t} \hat{x}_{t-1} - x_\star \right\rangle \\
&= \omega_t \left\langle \nabla f(\hat{x}_t), \frac{\sum_{k=0}^{t} \omega_k}{\omega_t} (\hat{x}_t - x_\star) - \frac{\sum_{k=0}^{t-1} \omega_k}{\omega_t} (\hat{x}_{t-1} - x_\star) \right\rangle \\
&= \sum_{k=0}^{t} \omega_k \left\langle \nabla f(\hat{x}_t), \hat{x}_t - x_\star \right\rangle - \sum_{k=0}^{t-1} \omega_k \left\langle \nabla f(\hat{x}_t), \hat{x}_{t-1} - x_\star \right\rangle \\
&= \omega_t \left\langle \nabla f(\hat{x}_t), \hat{x}_t - x_\star \right\rangle + \sum_{k=0}^{t-1} \omega_k \left\langle \nabla f(\hat{x}_t), \hat{x}_t - \hat{x}_{t-1} \right\rangle .
\end{aligned}$$

By using the convexity of f, we get

$$\omega_t \left\langle \nabla f(\hat{x}_t), x_{t+1} - x_\star \right\rangle \geq \omega_t (f(\hat{x}_t) - f(x_\star)) + \sum_{k=0}^{t-1} \omega_k (f(\hat{x}_t) - f(\hat{x}_{t-1})). \tag{31}$$

Therefore, for any $T > 0$

$$\begin{aligned}
\sum_{t=0}^{T-1} \omega_t \left\langle \nabla f(\hat{x}_t), x_{t+1} - x_\star \right\rangle &\geq \sum_{t=0}^{T-1} \omega_t (f(\hat{x}_t) - f(x_\star)) + \sum_{t=0}^{T-1} \sum_{k=0}^{t-1} \omega_k (f(\hat{x}_t) - f(\hat{x}_{t-1})) \\
&= \sum_{t=0}^{T-1} \omega_t (f(\hat{x}_t) - f(x_\star)) + \sum_{k=0}^{T-2} \sum_{t=k+1}^{T-1} \omega_k (f(\hat{x}_t) - f(\hat{x}_{t-1})).
\end{aligned}$$

By performing a telescopic summation, we obtain

$$\begin{aligned}
\sum_{t=0}^{T-1} \omega_t \left\langle \nabla f(\hat{x}_t), x_{t+1} - x_\star \right\rangle &\geq \sum_{t=0}^{T-1} \omega_t (f(\hat{x}_t) - f(x_\star)) + \sum_{t=0}^{T-2} \omega_t (f(\hat{x}_{T-1}) - f(\hat{x}_t)) \\
&= \omega_{T-1} (f(\hat{x}_{T-1}) - f(x_\star)) + \sum_{t=0}^{T-2} \omega_t (f(\hat{x}_t) - f(x_\star) + f(\hat{x}_{T-1}) - f(\hat{x}_t)) \\
&= \sum_{t=0}^{T-1} \omega_t (f(\hat{x}_{T-1}) - f(x_\star)).
\end{aligned}$$

Dividing both sides by $\sum_{t=0}^{T-1} \omega_t$ concludes the proof. ■

C.2. Inductive suboptimality bound (Lemma 15)

Lemma 15 *Let $s_0, s_1, \ldots, s_{T-1}$ and $h_0, h_1, \ldots, h_{T-1}$ be non-negative non-decreasing sequences. Let $b > 1$ such that $\bar{r}_{t+1}/\bar{r}_t \leq b$ for any $t \in \{0, 1, 2 \ldots, T-1\}$. If for all $t \in \{0, 1, 2 \ldots, T-1\}$ there exist $\kappa_t \in \{0, 1, 2 \ldots, t\}$ such that*

$$f(\hat{x}_t) - f(x_\star) \leq \frac{\alpha_{\kappa_t} \sqrt{s_t} \sqrt{f(\hat{x}_{\kappa_t}) - f(x_\star)} + h_t}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2},$$

then for all $t \in \{0, 1, 2 \ldots, T-1\}$ *we have that*

$$f(\hat{x}_t) - f(x_\star) \leq \frac{4b^2(s_t + h_t)}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2}.$$

Proof We prove by induction that

$$f(\hat{x}_t) - f(x_\star) \leq \frac{4b^2(s_t + h_t)}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2}.$$

We will only use the induction assumption for the case where $\kappa_t < t$.

If $\kappa_t = t$**:** We have that

$$\begin{aligned} f(\hat{x}_t) - f(x_\star) &\leq \frac{\alpha_{\kappa_t}\sqrt{s_t}\sqrt{f(\hat{x}_{\kappa_t}) - f(x_\star)} + h_t}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2} \\ &\leq \frac{\frac{\bar{r}_{t+1}}{\bar{r}_t}\sqrt{s_t}\sqrt{f(\hat{x}_{\kappa_t}) - f(x_\star)}}{\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}} + \frac{h_t}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2} \\ &\leq \frac{b\sqrt{s_t}\sqrt{f(\hat{x}_{\kappa_t}) - f(x_\star)}}{\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}} + \frac{h_t}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2}. \end{aligned}$$

Thus,

$$f(\hat{x}_t) - f(x_\star) - \frac{b\sqrt{s_t}\sqrt{f(\hat{x}_{\kappa_t}) - f(x_\star)}}{\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}} \leq \frac{h_t}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2}.$$

If

$$\frac{f(\hat{x}_t) - f(x_\star)}{2} \leq f(\hat{x}_t) - f(x_\star) - \frac{b\sqrt{s_t}\sqrt{f(\hat{x}_{\kappa_t}) - f(x_\star)}}{\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}},$$

then

$$f(\hat{x}_t) - f(x_\star) \leq \frac{2h_t}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2}.$$

Otherwise,

$$\frac{f(\hat{x}_t) - f(x_\star)}{2} \leq \frac{b\sqrt{s_t}\sqrt{f(\hat{x}_{\kappa_t}) - f(x_\star)}}{\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}}.$$

Therefore,

$$\sqrt{f(\hat{x}_t) - f(x_\star)} \leq \frac{2b\sqrt{s_t}}{\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}}.$$

Consequentially,

$$f(\hat{x}_t) - f(x_\star) \leq \frac{4b^2 s_t}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2}.$$

In either case, we obtain that

$$f(\hat{x}_t) - f(x_\star) \leq \frac{4b^2(s_t + h_t)}{\left(\sum_{k=0}^{t} \bar{r}_k / \bar{r}_{t+1}\right)^2}.$$

If $\kappa_t < t$: We assume by induction that

$$f(\hat{x}_{\kappa_t}) - f(x_\star) \leq \frac{4b^2(s_{\kappa_t} + h_{\kappa_t})}{\left(\sum_{k=0}^{\kappa_t} \bar{r}_k/\bar{r}_{\kappa_t+1}\right)^2}.$$

Therefore,

$$\begin{aligned}
\alpha_{\kappa_t}\sqrt{s_t}\sqrt{f(\hat{x}_{\kappa_t}) - f(x_\star)} &\leq 2b\sqrt{s_t}\sqrt{s_{\kappa_t} + h_{\kappa_t}}\frac{\alpha_{\kappa_t}}{\sum_{k=0}^{\kappa_t} \bar{r}_k/\bar{r}_{\kappa_t+1}} \\
&\leq 2b\frac{\bar{r}_{t+1}}{\bar{r}_t}\sqrt{s_t}\sqrt{s_t + h_t} \\
&\leq 2b^2(s_t + h_t).
\end{aligned}$$

Thus,

$$\begin{aligned}
f(\hat{x}_t) - f(x_\star) &\leq \frac{2b^2(s_t + h_t) + h_t}{\left(\sum_{k=0}^{t} \bar{r}_k/\bar{r}_{t+1}\right)^2} \\
&\leq \frac{4b^2(s_t + h_t)}{\left(\sum_{k=0}^{t} \bar{r}_k/\bar{r}_{t+1}\right)^2}.
\end{aligned}$$

Finalizing the induction: For $t = 0$ we have $\kappa_t = 0 = t$. For the case $\kappa_t = t$ we did not use the induction assumption, and therefore we have the base of the induction:

$$f(\hat{x}_0) - f(x_\star) \leq \frac{4b^2(s_0 + h_0)}{\left(\bar{r}_0/\bar{r}_1\right)^2}.$$

Thus, by induction we get that for all $t \in \{0, 1, 2, \ldots, T-1\}$,

$$f(\hat{x}_t) - f(x_\star) \leq \frac{4b^2(s_t + h_t)}{\left(\sum_{k=0}^{t} \bar{r}_k/\bar{r}_{t+1}\right)^2}.$$

■

C.3. General regret bound (Lemma 16)

The following lemma is inspired by the regret analysis of UNIXGRAD [30].

Lemma 16 *Using Algorithm 1, eq. (2) and eq. (3), for any $t \geq 0$, $\rho_t > 0$, we have that*

$$\begin{aligned}
\bar{r}_t\alpha_t \left\langle g_t, x_{t+1} - x_\star \right\rangle \leq{}& \frac{\bar{r}_t^2\alpha_t^2\rho_t}{2}\|g_t - m_t\|^2 - \frac{1}{2\rho_t}\|x_{t+1} - y_t\|^2 \\
&+ \left(\frac{1}{2\rho_t} - \frac{1}{2\tilde{\eta}_{x,t}}\right)\left(\|x_{t+1} - y_t\|^2 + \|x_{t+1} - y_{t+1}\|^2\right) \\
&+ \frac{1}{2\tilde{\eta}_{y,t}}\left(\|y_t - x_\star\|^2 - \|y_{t+1} - x_\star\|^2\right).
\end{aligned}$$

Proof We have

$$\begin{aligned}&\bar{r}_t\alpha_t \langle g_t, x_{t+1} - x_\star\rangle \\ &\quad = \bar{r}_t\alpha_t \langle g_t - m_t, x_{t+1} - y_{t+1}\rangle + \bar{r}_t\alpha_t \langle m_t, x_{t+1} - y_{t+1}\rangle + \bar{r}_t\alpha_t \langle g_t, y_{t+1} - x_\star\rangle . \end{aligned} \tag{32}$$

In addition

$$\begin{aligned}\bar{r}_t\alpha_t \langle g_t - m_t, x_{t+1} - y_{t+1}\rangle &\overset{(i)}{\leq} \bar{r}_t\alpha_t \|g_t - m_t\| \|x_{t+1} - y_{t+1}\| \\ &\overset{(ii)}{\leq} \frac{\rho_t \bar{r}_t \alpha_t^2}{2}\|g_t - m_t\|^2 + \frac{1}{2\rho_t}\|x_{t+1} - y_{t+1}\|,\end{aligned} \tag{33}$$

where (i) is from Holder's Inequality and (ii) is due to Young's Inequality.

For the Euclidean Bregman divergence $\mathcal{D}_\mathcal{R}(x, y) = \frac{1}{2}\|x - y\|^2$ we have that the update rule $x_{t+1} = \mathrm{Proj}_\mathcal{K}(y_t - \alpha_t\eta_{x,t}m_t) = \mathrm{Proj}_\mathcal{K}(y_t - \bar{r}_t\alpha_t\tilde{\eta}_{x,t}m_t)$ is equivalent to the update rule $x_{t+1} = \arg\min_{x\in\mathcal{K}}\left\{\bar{r}_t\alpha_t \langle x, m_t\rangle + \frac{1}{\tilde{\eta}_{x,t}}\mathcal{D}_\mathcal{R}(x, y_t)\right\}$. Therefore, from the optimality of x_{t+1} we get

$$\begin{aligned}\bar{r}_t\alpha_t \langle m_t, x_{t+1} - y_{t+1}\rangle &\leq \frac{1}{\tilde{\eta}_{x,t}} \langle \nabla_x \mathcal{D}_\mathcal{R}(x_{t+1}, y_t), x_{t+1} - y_{t+1}\rangle \\ &= \frac{1}{\tilde{\eta}_{x,t}}(\mathcal{D}_\mathcal{R}(y_{t+1}, y_t) - \mathcal{D}_\mathcal{R}(x_{t+1}, y_t) - \mathcal{D}_\mathcal{R}(y_{t+1}, x_{t+1})).\end{aligned} \tag{34}$$

Similarly, $y_{t+1} = \arg\min_{y\in\mathcal{K}}\left\{\bar{r}_t\alpha_t \langle y, g_t\rangle + \frac{1}{\tilde{\eta}_{y,t}}\mathcal{D}_\mathcal{R}(y, y_t)\right\}$. Therefore, from the optimality of y_{t+1} we get

$$\begin{aligned}\bar{r}_t\alpha_t \langle g_t, y_{t+1} - x_\star\rangle &\leq \frac{1}{\tilde{\eta}_{y,t}} \langle \nabla_x \mathcal{D}_\mathcal{R}(y_{t+1}, y_t), x_\star - y_{t+1}\rangle \\ &= \frac{1}{\tilde{\eta}_{y,t}}(\mathcal{D}_\mathcal{R}(x_\star, y_t) - \mathcal{D}_\mathcal{R}(y_{t+1}, y_t) - \mathcal{D}_\mathcal{R}(x_\star, y_{t+1})).\end{aligned} \tag{35}$$

By combining eqs. (34), (35) and (33) into eq. (32) we obtain that

$$\begin{aligned}\bar{r}_t\alpha_t \langle g_t, x_{t+1} - x_\star\rangle &\leq \frac{\bar{r}_t^2\alpha_t^2\rho_t}{2}\|g_t - m_t\|^2 + \frac{1}{2\rho_t}\|x_{t+1} - y_{t+1}\|^2 \\ &\quad + \frac{1}{2\tilde{\eta}_{x,t}}\left(\|y_{t+1} - y_t\|^2 - \|x_{t+1} - y_t\|^2 - \|y_{t+1} - x_{t+1}\|^2\right) \\ &\quad + \frac{1}{2\tilde{\eta}_{y,t}}\left(\|x_\star - y_t\|^2 - \|y_{t+1} - y_t\|^2 - \|x_\star - y_{t+1}\|^2\right) \\ &= \frac{\bar{r}_t^2\alpha_t^2\rho_t}{2}\|g_t - m_t\|^2 - \frac{1}{2\rho_t}\|x_{t+1} - y_t\|^2 \\ &\quad + \left(\frac{1}{2\rho_t} - \frac{1}{2\tilde{\eta}_{x,t}}\right)\left(\|x_{t+1} - y_t\|^2 + \|x_{t+1} - y_{t+1}\|^2\right) \\ &\quad + \frac{1}{2\tilde{\eta}_{y,t}}\left(\|x_\star - y_t\|^2 - \|x_\star - y_{t+1}\|^2\right) + \left(\frac{1}{2\tilde{\eta}_{x,t}} - \frac{1}{2\tilde{\eta}_{y,t}}\right)\|y_{t+1} - y_t\|^2.\end{aligned}$$

Since $\tilde{\eta}_{y,t} \leq \tilde{\eta}_{x,t}$, we may drop the final term in the above display, completing the proof. ■

Appendix D. Iterate stability lemmas

D.1. A weighted regret bound (Lemma 17)

Lemma 17 *For any sequence of positive numbers $\omega_0, \omega_1, \omega_2, \dots$, define*

$$\hat{x}_t := \frac{\sum_{k=0}^{t} \omega_k x_{k+1}}{\sum_{k=0}^{t} \omega_k}.$$

Let $\tilde{\eta}_0, \tilde{\eta}_1, \tilde{\eta}_2, \dots$ be a non-increasing sequence of positive numbers. We have that for any $T > 0$,

$$\sum_{t=0}^{T-1} \omega_t \tilde{\eta}_t \left\langle \nabla f(\hat{x}_t), x_{t+1} - x_\star \right\rangle \geq 0.$$

Proof Define

$$\tilde{f}(x) = f(x) - f(x_\star).$$

We start from eq. (31) inside the proof of Lemma 14, which says that for all $t \geq 0$

$$\omega_t \left\langle \nabla f(\hat{x}_t), x_{t+1} - x_\star \right\rangle \geq \omega_t (f(\hat{x}_t) - f(x_\star)) + \sum_{k=0}^{t-1} \omega_k (f(\hat{x}_t) - f(\hat{x}_{t-1})).$$

Multiplying each side by $\tilde{\eta}_t$ and summing, we obtain

$$\begin{aligned}
\sum_{t=0}^{T-1} \omega_t \tilde{\eta}_t \left\langle \nabla f(\hat{x}_t), x_{t+1} - x_\star \right\rangle &\geq \sum_{t=0}^{T-1} \omega_t \tilde{\eta}_t (f(\hat{x}_t) - f(x_\star)) + \sum_{t=0}^{T-1} \sum_{k=0}^{t-1} \omega_k \tilde{\eta}_t (f(\hat{x}_t) - f(\hat{x}_{t-1})) \\
&= \sum_{t=0}^{T-1} \omega_t \tilde{\eta}_t \tilde{f}(\hat{x}_t) + \sum_{t=0}^{T-1} \sum_{k=0}^{t-1} \omega_k \tilde{\eta}_t \Big(\tilde{f}(\hat{x}_t) - \tilde{f}(\hat{x}_{t-1}) \Big) \\
&\overset{(\star)}{\geq} \sum_{t=0}^{T-1} \omega_t \tilde{\eta}_t \tilde{f}(\hat{x}_t) + \sum_{t=0}^{T-1} \sum_{k=0}^{t-1} \omega_k \Big(\tilde{\eta}_t \tilde{f}(\hat{x}_t) - \tilde{\eta}_{t-1} \tilde{f}(\hat{x}_{t-1}) \Big) \\
&= \sum_{t=0}^{T-1} \omega_t \tilde{\eta}_t \tilde{f}(\hat{x}_t) + \sum_{k=0}^{T-2} \sum_{t=k+1}^{T-1} \omega_k \Big(\tilde{\eta}_t \tilde{f}(\hat{x}_t) - \tilde{\eta}_{t-1} \tilde{f}(\hat{x}_{t-1}) \Big),
\end{aligned}$$

where $(\star)$ is because that $\tilde{f}(\hat{x}_{t-1}) \geq 0$ and $\tilde{\eta}_{t-1} \geq \tilde{\eta}_t > 0$.

We can now perform a telescopic summation and obtain

$$\begin{aligned}
\sum_{t=0}^{T-1} \omega_t \tilde{\eta}_t \left\langle \nabla f(\hat{x}_t), x_{t+1} - x_\star \right\rangle &\geq \sum_{t=0}^{T-1} \omega_t \tilde{\eta}_t \tilde{f}(\hat{x}_t) + \sum_{t=0}^{T-2} \omega_t \Big(\tilde{\eta}_{T-1} \tilde{f}(\hat{x}_{T-1}) - \tilde{\eta}_t \tilde{f}(\hat{x}_t) \Big) \\
&= \omega_{T-1} \tilde{\eta}_{T-1} \tilde{f}(\hat{x}_{T-1}) + \sum_{t=0}^{T-2} \omega_t \Big(\tilde{\eta}_t \tilde{f}(\hat{x}_t) + \tilde{\eta}_{T-1} \tilde{f}(\hat{x}_{T-1}) - \tilde{\eta}_t \tilde{f}(\hat{x}_t) \Big) \\
&= \omega_{T-1} \tilde{\eta}_{T-1} \tilde{f}(\hat{x}_{T-1}) + \sum_{t=1}^{T-1} \omega_t \tilde{\eta}_{T-1} \tilde{f}(\hat{x}_{T-1}).
\end{aligned}$$

Thus, because $\tilde{f}(\hat{x}_{T-1}) \geq 0$, we obtain that

$$\sum_{t=0}^{T-1} \omega_t \tilde{\eta}_t \left\langle \nabla f(\hat{x}_t), x_{t+1} - x_\star \right\rangle \geq 0.$$

■

D.2. Inductive stability bound (Lemma 18)

Lemma 18 *If $r_\epsilon = r_0 \leq d_0$, and for all $t \geq 1$ we have that*

$$\|y_t - x_t\| \leq \frac{\bar{r}_{t-1}}{4} \quad \textit{and}$$
$$d_t^2 \leq \left(d_0 + \frac{1}{4}\bar{r}_{t-1} \right)^2,$$

then for all $t \geq 0$ we get that

$$d_t \leq 2d_0 \quad \textit{and} \quad r_t \leq 4d_0.$$

Proof We prove this lemma by induction. The basis of the induction is that for $t = 0$ we get that $d_0 \leq 2d_0$ and $r_0 \leq d_0 \leq 4d_0$.

For any $t \geq 1$, we assume that $\bar{d}_{t-1} \leq 2d_0$ and $\bar{r}_{t-1} \leq 4d_0$. Thus,

$$d_t \leq d_0 + \frac{1}{4}\bar{r}_{t-1} \leq 2d_0.$$

Also,

$$\|y_t - x_0\| \leq \|y_t - x_\star\| + \|x_0 - x_\star\| = d_t + d_0 \leq 3d_0.$$

In addition,

$$\begin{aligned} \|x_t - x_0\| &\leq \|y_t - x_0\| + \|x_t - y_t\| \\ &\overset{(\star)}{\leq} 3d_0 + \frac{\bar{r}_{t-1}}{4} \\ &\leq 4d_0. \end{aligned}$$

where $(\star)$ is because $\|x_t - y_t\| \leq \frac{\bar{r}_{t-1}}{4}$. As a result,

$$d_t \leq 2d_0 \quad \text{and} \quad r_t \leq 4d_0.$$

Finally, by induction, we get that for all $t \geq 0$

$$d_t \leq 2d_0 \quad \text{and} \quad r_t \leq 4d_0.$$

■

D.3. Single-step iterate stability (Lemma 19)

Lemma 19 *Let c be a positve number. Using Algorithm 1, for any $t \geq 0$, if $\eta_{x,t} \leq \frac{\bar{r}_t}{c\alpha_t\|m_t\|}$, $\eta_{y,t} \leq \frac{\bar{r}_t}{c\alpha_t\|g_t - m_t\|}$ and $\eta_{y,t} \leq \eta_{x,t}$ then*

$$\begin{aligned}\|x_{t+1} - y_t\| &\leq \frac{\bar{r}_t}{c}\\ \|y_{t+1} - y_t\| &\leq \frac{2\bar{r}_t}{c}\\ \|x_{t+1} - y_{t+1}\| &\leq \frac{2\bar{r}_t}{c}\\ \bar{r}_{t+1} &\leq \bar{r}_t\left(1 + \frac{2}{c}\right).\end{aligned}$$

Proof First, by definition of the iterates and the fact that $\mathcal{K}$ is convex (and projection onto a closed convex set is nonexpansive) we have

$$\|x_{t+1} - y_t\| = \|\mathrm{Proj}_{\mathcal{K}}(y_t - \alpha_t\eta_{x,t}m_t) - y_t\| \leq \alpha_t\eta_{x,t}\|m_t\| \leq \frac{\bar{r}_t}{c}. \tag{36}$$

Second, by definition of the iterates and the fact that $\mathcal{K}$ is convex, we also have

$$\begin{aligned}\|y_{t+1} - y_t\| &= \|\mathrm{Proj}_{\mathcal{K}}(y_t - \alpha_t\eta_{y,t}g_t) - y_t\| \leq \alpha_t\eta_{y,t}\|g_t\|\\ &\leq \alpha_t\eta_{y,t}\|g_t - m_t\| + \alpha_t\eta_{y,t}\|m_t\| \leq \frac{2\bar{r}_t}{c}.\end{aligned} \tag{37}$$

Third, by definition of the iterates, the fact that $\mathcal{K}$ is convex, the fact $\eta_{y,t} \leq \eta_{x,t}$, and the assumed upper bounds on $\eta_{y,t}$ and $\eta_{x,t}$ in the premise of this lemma we have

$$\begin{aligned}\|x_{t+1} - y_{t+1}\| &= \|\mathrm{Proj}_{\mathcal{K}}(y_t - \alpha_t\eta_{x,t}m_t) - \mathrm{Proj}_{\mathcal{K}}(y_t - \alpha_t\eta_{y,t}g_t)\|\\ &\leq \alpha_t\|\eta_{x,t}m_t - \eta_{y,t}g_t\| \leq \alpha_t\eta_{y,t}\|g_t - m_t\| + \alpha_t(\eta_{x,t} - \eta_{y,t})\|m_t\|\\ &\leq \alpha_t\eta_{y,t}\|g_t - m_t\| + \alpha_t\eta_{x,t}\|m_t\| \leq \frac{2\bar{r}_t}{c}.\end{aligned}$$

Finally,

$$r_{t+1} \leq r_t + \max(\|x_{t+1} - y_t\|, \|y_{t+1} - y_t\|).$$

Therefore, using eq. (36) and eq. (37) we obtain

$$\bar{r}_{t+1} = \max(\bar{r}_t, r_{t+1}) \leq \bar{r}_t + \max(\|x_{t+1} - y_t\|, \|y_{t+1} - y_t\|) \leq \bar{r}_t\left(1 + \frac{2}{c}\right).$$

■

Appendix E. Concentration bounds

E.1. An empirical-Bernstein-type time uniform concentration bound (Lemma 20)

Lemma 20 (From Ivgi et al. [27]) *Let S be the set of nonnegative and nondecreasing sequences. Let $C_t \in \mathcal{F}_{t-1}$ and let X_t be a martingale difference sequence adapted to $\mathcal{F}_t$ such that $|X_t| \leq C_t$ with probability 1 for all t. Then, for all $\delta \in (0,1)$, $c > 0$, and $\hat{X}_t \in \mathcal{F}_{t-1}$ such that $|\hat{X}_t| \leq C_t$ with probability 1,*

$$\mathbb{P}\left(\exists t \leq T, \exists \{y_i\}_{i=1}^{\infty} \in S : \left|\sum_{i=1}^{t} y_i X_i\right| \geq 8 y_t \sqrt{\theta_{t,\delta} \sum_{i=1}^{t} \left(X_i - \hat{X}_i\right)^2 + c^2 \theta_{t,\delta}^2}\right) \leq \delta + \mathbb{P}(\exists t \leq T : C_t > c).$$

E.2. Concentration bound for suboptimally proof (Lemma 21)

Lemma 21 *Let $\mathfrak{B} > 0$ and $\delta \in (0,1)$. In the bounded noise setting (Assumption 6), using Algorithm 1 and eq. (12), with probability of at least $1 - \delta - \mathbb{P}\left[\bar{\mathfrak{b}}_{T-1} > \mathfrak{B}\right]$ we get that for all $t \in \{0, 1, \ldots, T-1\}$ then*

$$\left|\sum_{k=0}^{t} \bar{r}_t \alpha_k \left\langle \nabla f(\hat{x}_k) - g_k, x_{k+1} - x_\star \right\rangle\right| \leq 8 \alpha_t \bar{r}_t (\bar{r}_{t+1} + d_0) \sqrt{\theta_{t+1,\delta} \sum_{k=0}^{t} \|\nabla f(\hat{x}_k) - g_k\|^2 + (\theta_{t+1,\delta} \mathfrak{B})^2}.$$

Proof For $k \in \{0, 1, \ldots, T\}$ define

$$\tilde{d}_k = \max_{i \leq k} \|x_k - x_\star\|.$$

For $k \in \{0, 1, \ldots, T-1\}$ define the random variables:

$$Y_k = \alpha_k \bar{r}_k \tilde{d}_{k+1}, \quad \text{and} \quad X_k = \left\langle \nabla f(\hat{x}_k) - g_k, \frac{x_{k+1} - x_\star}{\tilde{d}_{k+1}} \right\rangle.$$

From these definitions we get

$$\sum_{k=0}^{t} Y_k X_k = \sum_{k=0}^{t} \bar{r}_t \alpha_k \left\langle \nabla f(\hat{x}_k) - g_k, x_{k+1} - x_\star \right\rangle,$$

and that $\{Y_k\}_{k=0}^{T-1}$ is a non-decreasing sequence of non-negative numbers. In addition, as x_{k+1} and $\tilde{d}_{k+1}$ are independent of the noise of g_k then X_k is a martingale difference sequence. Therefore, as $|X_k| \leq \bar{\mathfrak{b}}_k$ with probability of 1, Lemma 20 gives us that

$$\mathbb{P}\left(\exists t < T \ : \ \left|\sum_{k=0}^{t} Y_k X_k\right| \geq 8 Y_t \sqrt{\theta_{t+1,\delta} \sum_{k=0}^{t} (X_k - 0)^2 + (\theta_{t+1,\delta} \mathfrak{B})^2}\right) \leq \delta + \mathbb{P}\left[\bar{\mathfrak{b}}_{T-1} > \mathfrak{B}\right].$$

Therefore, by using the Cauchy–Schwarz inequality, we obtain that, with a probability of at least $1-\delta-\mathbb{P}\big[\bar{\mathfrak{b}}_{T-1}>\mathfrak{B}\big]$, for all $t\in\{0,1,\dots,T-1\}$

$$\left|\sum_{k=0}^{t}\bar{r}_t\alpha_k\left\langle\nabla f(\hat{x}_k)-g_k,x_{k+1}-x_\star\right\rangle\right|\leq 8\alpha_t\bar{r}_t\tilde{d}_{t+1}\sqrt{\theta_{t+1,\delta}\sum_{k=0}^{t}\|\nabla f(\hat{x}_k)-g_k\|^2+\left(\theta_{t+1,\delta}\mathfrak{B}\right)^2}$$
$$\leq 8\alpha_t\bar{r}_t\Big(\bar{r}_{t+1}+\tilde{d}_0\Big)\sqrt{\theta_{t+1,\delta}\sum_{k=0}^{t}\|\nabla f(\hat{x}_k)-g_k\|^2+\left(\theta_{t+1,\delta}\mathfrak{B}\right)^2}.$$

Thus,

$$\left|\sum_{k=0}^{t}\bar{r}_t\alpha_k\left\langle\nabla f(\hat{x}_k)-g_k,x_{k+1}-x_\star\right\rangle\right|\leq 8\alpha_t\bar{r}_t(\bar{r}_{t+1}+d_0)\sqrt{\theta_{t+1,\delta}\sum_{k=0}^{t}\|\nabla f(\hat{x}_k)-g_k\|^2+\left(\theta_{t+1,\delta}\mathfrak{B}\right)^2}.$$

■

E.3. Concentration bound for iterate stability proof (Lemma 22)

Lemma 22 *Let $\tilde{\eta}_{y,t}$ be such that, for some $c,s>0$ we have*

$$\frac{1}{\tilde{\eta}_{y,t}}\geq c\max\left\{\sqrt{s+Q_t}\log_+\left(\frac{s+Q_t}{s}\right),\alpha_t\|\nabla f(\hat{x}_t)-m_t\|,\alpha_t\|\nabla f(\hat{x}_t)-g_t\|\right\}.$$

If for all $t\geq 0$ we have that $\eta_{y,t}=\bar{r}_t\tilde{\eta}_{y,t}$ is independent of g_t given $x_0,\dots,x_t$, then, with probability of at least $1-\delta$, for all $t\geq 0$,

$$\left|\sum_{k=0}^{t}\alpha_k\eta_{y,k}\left\langle g_k-\nabla f(\hat{x}_k),x_{k+1}-x_\star\right\rangle\right|\leq\frac{12\theta_{t+1,\delta}}{c}\bar{r}_t(\bar{r}_{t+1}+d_0).$$

Proof For $t\in\{0,1,\dots,T\}$ define

$$\tilde{d}_t=\max_{k\leq t}\|x_t-x_\star\|.$$

For $t\in\{0,1,\dots,T-1\}$ define

$$X_t=\alpha_t\tilde{\eta}_{y,t}\left\langle g_t-\nabla f(\hat{x}_t),\frac{x_{t+1}-x_\star}{\tilde{d}_{t+1}}\right\rangle\ ,$$
$$\hat{X}_t=\alpha_t\tilde{\eta}_{y,t}\left\langle\nabla f(\hat{x}_t)-m_t,\frac{x_{t+1}-x_\star}{\tilde{d}_{t+1}}\right\rangle\ \text{ and}$$
$$Y_t=\bar{r}_t\tilde{d}_{t+1}.$$

The assumption $\frac{1}{\tilde{\eta}_{y,t}}\geq c\alpha_t\max\{\|\nabla f(\hat{x}_t)-m_t\|,\|g_t-\nabla f(\hat{x}_t)\|\}$ implies that $\max\Big\{|X_t|,|\hat{X}_t|\Big\}\leq\frac{1}{c}$. In addition, as m_t ,x_{t+1} and $\tilde{d}_{t+1}$ are independent of the noise of g_t then $\hat{X}_t$ is independent of the noise of g_t and X_t is a martingale difference sequence. Thus, Lemma 20 gives us that

$$\mathbb{P}\left(\forall t\in\{0,1,\dots\}\ :\ \left|\sum_{k=0}^{t}Y_kX_k\right|<8\bar{r}_t\tilde{d}_{t+1}\sqrt{\theta_{t+1,\delta}\sum_{k=0}^{t}\Big(X_k-\hat{X}_k\Big)^2+\frac{1}{c^2}\theta_{t+1,\delta}^2}\right)\geq 1-\delta.$$

Furthermore, we have

$$\sum_{k=0}^{t}\Big(X_t-\hat{X}_t\Big)^2=\sum_{k=0}^{t}\left(\alpha_k\tilde{\eta}_{y,k}\left\langle g_k-m_k,\frac{x_{k+1}-x_\star}{\tilde{d}_{t+1}}\right\rangle\right)^2\le\sum_{k=0}^{t}\alpha_t^2\tilde{\eta}_{y,k}^2\|g_k-m_k\|^2$$
$$\overset{(i)}{\le}\frac{1}{c^2}\sum_{k=0}^{t}\frac{\alpha_t^2\|g_k-m_k\|^2}{\left(s+\sum_{k=0}^{t}\alpha_k^2\|g_k-m_k\|^2\right)\log_+^2\left(\frac{s+\sum_{k=0}^{t}\alpha_k^2\|g_k-m_k\|^2}{s}\right)}\overset{(ii)}{\le}\frac{1}{c^2},$$

where (i) follows from the assumption that $\frac{1}{\tilde{\eta}_{y,t}}\ge c\sqrt{s+Q_t}\log_+\left(\frac{s+Q_t}{s}\right)$ and the definition of Q_t, and (ii) is a direct result of Lemma 30 with $a_k=s+\sum_{k=0}^{t}\alpha_k^2\|g_k-m_k\|^2$. In addition, we have that

$$Y_tX_t=\alpha_t\eta_{y,t}\left\langle g_t-\nabla f(\hat{x}_t),x_{t+1}-x_\star\right\rangle.$$

Therefore, with probability of at least $1-\delta$, for all $t\ge0$ we have that

$$\left|\sum_{k=0}^{t}\alpha_k\eta_{y,k}\left\langle g_k-\nabla f(\hat{x}_k),x_{k+1}-x_\star\right\rangle\right|\le8\bar{r}_t\tilde{d}_{t+1}\sqrt{\frac{\theta_{t+1,\delta}}{c^2}+\frac{\theta_{t+1,\delta}^2}{c^2}}$$
$$\le\frac{12\theta_{t+1,\delta}}{c}\bar{r}_t\Big(\bar{r}_{t+1}+\tilde{d}_0\Big)$$
$$\le\frac{12\theta_{t+1,\delta}}{c}\bar{r}_t(\bar{r}_{t+1}+d_0).$$

■

E.4. Relating $\bar{Q}_t$ to Q_t (Lemma 23)

Lemma 23 *Let $\mathfrak{B}>0$ and $\delta\in(0,1)$. In the bounded noise setting (Assumption 6), using Algorithm 1 and the step sizes (14), with probability of at least $1-\delta-\mathbb{P}\left[\bar{\mathfrak{b}}_{T-1}>\mathfrak{B}\right]$ we get that, for all $t\in\{0,1,\dots,T-1\}$,*

$$\bar{Q}_t\le5Q_t+80(t+1)^3\sqrt{\theta_{t+1,\delta}}V_t+2(t+1)^2\theta_{t+1,\delta}\mathfrak{B}^2.$$

Proof For all $k\ge0$ we have

$$\|\tilde{g}_k-m_k\|^2\le2\|g_k-m_k\|^2+2\|g_k-\tilde{g}_t\|^2$$
$$\le2\|g_k-m_k\|^2+4\|g_k-\nabla f(\hat{x}_k)\|^2+4\|\tilde{g}_k-\nabla f(\hat{x}_k)\|^2.$$

Therefore, since $\alpha_k\le k+1$,

$$\sum_{k=0}^{t}\alpha_k^2\|\tilde{g}_k-m_k\|^2\le2\sum_{k=0}^{t}\alpha_k^2\|g_k-m_k\|^2+8\sum_{k=0}^{t}(k+1)^2\|g_k-\nabla f(\hat{x}_k)\|^2$$
$$+4\sum_{k=0}^{t}(k+1)^2(\|\tilde{g}_k-\nabla f(\hat{x}_k)\|^2-\|g_k-\nabla f(\hat{x}_k)\|^2).\qquad(38)$$

We now bound $\sum_{k=0}^{t}(k+1)^2(\|\tilde{g}_k - \nabla f(\hat{x}_k)\|^2 - \|g_k - \nabla f(\hat{x}_k)\|^2)$. Define

$$\begin{aligned} X_t &= (\|\tilde{g}_t - \nabla f(\hat{x}_t)\|^2 - \|g_t - \nabla f(\hat{x}_t)\|^2) \ , \\ \hat{X}_t &= \|\tilde{g}_t - \nabla f(\hat{x}_t)\|^2 \ \text{ and} \\ Y_t &= (t+1)^2. \end{aligned}$$

We have that for all $t \geq 0$ then $|X_t| \leq \bar{\mathfrak{b}}_t^2$ and $|\hat{X}_t| \leq \bar{\mathfrak{b}}_t^2$ with probability 1. Therefore, Lemma 20 gives us that

$$\begin{aligned} \mathbb{P}\left(\forall t \in \{0,1,\dots,T-1\} \ : \ \left|\sum_{k=0}^{t} Y_k X_k\right| < 8Y_t\sqrt{\theta_{t+1,\delta}\sum_{k=0}^{t}\left(X_k - \hat{X}_k\right)^2 + \theta_{t+1,\delta}^2\mathfrak{B}^4}\right) \\ \geq 1 - \delta - \mathbb{P}\big(\bar{\mathfrak{b}}_{T-1} > \mathfrak{B}\big). \end{aligned}$$

Consequentially, by combining this result with eq. (38), we get that with probability at least $1 - \delta - \mathbb{P}\big(\bar{\mathfrak{b}}_{T-1} > \mathfrak{B}\big)$ that for all $t \in \{0,1,\dots,T-1\}$ we have that

$$\sum_{k=0}^{t}\alpha_k^2\|\tilde{g}_k - m_k\|^2 \leq 2\sum_{k=0}^{t}\alpha_k^2\|g_k - m_k\|^2 + 40(t+1)^2\sqrt{\theta_{t+1,\delta}}\sum_{k=0}^{t}\|g_k - \nabla f(\hat{x}_k)\|^2 + (t+1)^2\theta_{t+1,\delta}\mathfrak{B}^2.$$

Substituting into the above equation the definition of Q_t and V_t given in eq. (1) and eq. (11), respectively, and recalling the definition of $\bar{Q}_t$ given in eq. (13)

$$\bar{Q}_t = \sum_{k=0}^{t}\alpha_k^2\max\{\|g_k - m_k\|^2, 2\|\tilde{g}_k - m_k\|^2\} \leq Q_t + 2\sum_{k=0}^{t}\alpha_k^2\|\tilde{g}_k - m_k\|^2$$

completes the proof. ■

E.5. Concentration inequality for bounded random vectors (Lemma 24)

Lemma 24 (Howard et al. [25]) *For $T \in \mathbb{N}$, let $\{U_t\}_{t\in[T]}$ be a sequence of mean zero random vectors in $\mathbb{R}^d$ with $\|U_t\| \leq c$ almost surely. Then*

$$\mathbb{P}\left(\left\|\sum_{t=1}^{T} U_t\right\| \geq x\right) \leq 2\exp\left(-\frac{x^2}{2c^2T}\right).$$

Proof This result follows from Howard et al. [25, Corollary 10.a] with $Y_t = \sum_{k=1}^{t} U_k$, $\Psi(\cdot) = \|\cdot\|$, $c_t = c$ and $m = c^2T$. The selection of $\Psi(\cdot) = \|\cdot\|$ yields $D_\star = 1$ (see discussion preceding [25, Corollary 10.a]). Setting $c_t = c$ yields $V_t = c^2t$. Hence $\frac{D_\star^2}{2m}(V_T - m) \leq 0$ and Howard et al. [25, eq. (4.28)] gives the desired result. ■

Appendix F. Auxiliary lemmas

F.1. The growth rate of $\sum_k \bar{r}_k \alpha_k$ (Lemma 25)

We note that in accelerated optimization algorithms we normally have that $\alpha_t = \Theta(t)$. Even though this is not the case for U-DoG, α_t is roughly similar to t. First, it is easy to see that $1 \leq \alpha_t \leq t$. Secondly, the running sum of $\bar{r}_t \alpha_t$ grows roughly quadratically. This is shown in the following lemma, in which we replace α_t and $\bar{r}_t$ with a_t and s_t, respectively

Lemma 25 *Let $s_0, s_1, \ldots, s_t$ be a non-decreasing sequence of positive numbers. Define $a_k := \sum_{i=0}^{k} \frac{s_i}{s_k}$, then*

$$s_t a_t^2 \leq 2 \sum_{k=0}^{t} s_k a_k.$$

Proof We have

$$\sum_{k=0}^{t} s_k a_k = \sum_{k=0}^{t} \sum_{i=0}^{k} s_i = \sum_{k=0}^{t} (t-k+1) s_k.$$

And,

$$s_t a_t^2 = \frac{1}{s_t} \sum_{k=0}^{t} \sum_{i=0}^{t} s_k s_i = \frac{2}{s_t} \sum_{k=0}^{t} s_k \sum_{i=k}^{t} s_i - \frac{1}{s_t} \sum_{k=0}^{t} s_k^2 \leq 2 \sum_{k=0}^{t} s_k \sum_{i=k}^{t} \frac{s_i}{s_t} \leq 2 \sum_{k=0}^{t} (t-k+1) s_k.$$

Thus,

$$s_t a_t^2 \leq 2 \sum_{k=0}^{t} s_k a_k.$$

■

F.2. Discrete derivative lemma (Lemma 26)

Lemma 26 *Let c be a positive number, and let $s_0, s_1, s_2, \ldots$ be a sequence of positive numbers. For every $t \geq 0$ define*

$$\rho_t = \frac{1}{c \sqrt{\sum_{k=0}^{t} s_k}}.$$

We have that for every $t \geq 0$

$$\frac{1}{\rho_{t+1}} - \frac{1}{\rho_t} \leq c^2 \rho_{t+1} s_{t+1}.$$

Proof For every $t \geq 0$ we have that

$$s_{t+1} = \sum_{k=0}^{t+1} s_k - \sum_{k=0}^{t} s_k \geq \sqrt{\sum_{k=0}^{t+1} s_k}\left(\sqrt{\sum_{k=0}^{t+1} s_k} - \sqrt{\sum_{k=0}^{t} s_k}\right) = \frac{1}{c^2\rho_{t+1}}\left(\frac{1}{\rho_{t+1}} - \frac{1}{\rho_t}\right).$$

Thus,

$$\frac{1}{\rho_{t+1}} - \frac{1}{\rho_t} \leq c^2\rho_{t+1}s_{t+1}.$$

■

F.3. Discrete integral lemma (Lemma 27)

Lemma 27 *For any positive numbers c_1, c_2, for any $t \geq 0$, and for any sequence of non-negative numbers $B_0, B_1, B_2, \ldots, B_t$ we have that*

$$c_1\sqrt{\sum_{k=0}^{t} B_k^2} - \sum_{k=0}^{t} \frac{B_k^2}{c_2}\sqrt{\sum_{j=0}^{k} B_j^2} \leq 2c_1^{3/2}c_2^{1/2}.$$

Proof Define

$$\eta_{B,k} = \frac{1}{\sqrt{\sum_{j=1}^{k} B_j^2}}.$$

Lemma 28 gives us that

$$\sqrt{\sum_{k=0}^{t} B_k^2} \leq \sum_{k=0}^{t} \frac{B_k^2}{\sqrt{\sum_{j=0}^{k} B_j^2}}.$$

Therefore, we obtain

$$\begin{aligned} c_1\sqrt{\sum_{k=0}^{t} B_k^2} - \sum_{k=0}^{t} \frac{B_k^2}{c_2}\sqrt{\sum_{j=0}^{k} B_j^2} &\leq c_1\sum_{k=0}^{t} \frac{B_k^2}{\sqrt{\sum_{j=0}^{k} B_j^2}} - \sum_{k=0}^{t} \frac{B_k^2}{c_2}\sqrt{\sum_{j=0}^{k} B_j^2} \\ &= \sum_{k=0}^{t}\left(c_1\eta_{B,k} - \frac{1}{c_2\eta_{B,k}}\right)B_k^2. \end{aligned}$$

Define

$$\kappa = \max\left[\left\{t \in \{0,1,\ldots,t\} \;:\; 2c_1\eta_{B,t} - \frac{1}{c_2\eta_{B,t}} > 0\right\} \cup \{-1\}\right].$$

We have,

$$c_1\sqrt{\sum_{k=0}^{t} B_k^2} - \sum_{k=0}^{t} \frac{B_k^2}{c_2}\sqrt{\sum_{j=0}^{k} B_j^2} \leq \sum_{k=0}^{\kappa} c_1\eta_{B,k}B_k^2 = c_1\sum_{k=0}^{\kappa} \frac{B_k^2}{\sqrt{\sum_{j=0}^{k} B_j^2}} \overset{(\star)}{\leq} 2c_1\sqrt{\sum_{k=0}^{\kappa} B_k^2} = \frac{2c_1}{\eta_{B,\kappa}}\mathbb{1}_{\{\kappa\geq 0\}},$$

where $(\star)$ is because of Lemma 28. From the definition of κ, we obtain that

$$c_1 \eta_{B,\kappa} > \frac{1}{c_2 \eta_{B,\kappa}}.$$

Thus,

$$c_1 \sqrt{\sum_{k=0}^{t} B_k^2} - \sum_{k=0}^{t} \frac{B_k^2}{c_2} \sqrt{\sum_{j=0}^{k} B_j^2} \leq \frac{2c_1}{\eta_{B,\kappa}} \mathbb{1}_{\{\kappa \geq 0\}} \leq 2 c_1^{3/2} c_2^{1/2}.$$

■

F.4. Additional lemmas from prior work

Lemma 28 (e.g., Levy et al. [36]) *For any $k \geq 0$ and for any sequence on non-negative numbers $s_0, s_1, s_2, \ldots, s_k$ the following holds:*

$$\sqrt{\sum_{i=0}^{k} s_i} \leq \sum_{i=0}^{k} \frac{s_i}{\sqrt{\sum_{j=0}^{i} s_j}} \leq 2 \sqrt{\sum_{i=0}^{k} s_i}.$$

Lemma 29 (Ivgi et al. [27, Lemma 3]) *Let $s_0, s_1, \ldots, s_T$ be a positive nondecreasing sequence. Then*

$$\max_{t \leq T} \sum_{i<t} \frac{s_i}{s_t} \geq \frac{1}{e} \left(\frac{T}{\log_+(s_T/s_0)} - 1 \right).$$

Lemma 30 (Ivgi et al. [27, Lemma 6]) *Let $a_{-1}, a_0, a_1, \ldots, a_t$ be a non-decreasing sequence of non-negative numbers, then*

$$\sum_{k=0}^{t} \frac{a_k - a_{k-1}}{a_k \log_+^2(a_k/a_{-1})} \leq 1.$$

Lemma 31 (Attia and Koren [3, Lemma 15]) *Let X be a $\sigma^2(x)$-sub-Gaussian. For and $\delta \in (0,1)$ here exist a random variable $\bar{X}$ such that:*

1. *$\bar{X}$ is zero-mean: $\mathbb{E}\bar{X} = 0$.*
2. *$\bar{X}$ is equal to X w.h.p: $\mathbb{P}\big(\bar{X} = X\big) \geq 1 - \delta$.*
3. *$\bar{X}$ is bounded with probability 1: $\mathbb{P}\Big(\|\bar{X}\| = 3\sigma\sqrt{\log(4/\delta)}\Big) = 1$.*

Appendix G. Experimental details

G.1. U-DoG step sizes

In the experiments, we use the following step sizes for U-DoG

$$\eta_{x,t} = \frac{\bar{r}_t}{\sqrt{\max\{Q_{t-1}, M_t\}}} \text{ and } \eta_{y,t} = \frac{\bar{r}_t}{\sqrt{\max\{Q_t, M_t\}}},$$

with $\bar{r}_t$, Q_t, and M_t as defined in Section 2. This step size is similar to the choice in eq. (10), which enjoys proven stability in the noiseless case, except we replace the logarithmic factor in the denominator with 1; preliminary experiments indicated 1 was the smallest value for which the algorithm was stable in practice. This difference between practical and theoretical algorithms is analogous to the difference between DoG and its theoretically stable variant T-DoG [27]. However, we maintain the maximization with M_t in the denominator, mainly in order to ensure that $\eta_{x,t}$ and $\eta_{y,t}$ are not too large early in the training. As with DoG, the additional step size adjustments necessary for the stochastic setting (given in eq. (14)) do not appear to be useful in practical settings.

G.2. ACCELEGRAD-DOG (A-DOG)

While U-DoG enjoys strong theoretical guarantees, it requires an extra-gradient computation at each step, which can be expensive in practice. To address this, we propose an alternative algorithm, A-DoG, which combines AcceleGrad [36] and DoG. To complete the combination we set α_t in the same way as it is calculated in U-DoG (algorithm 1). A-DoG is a simple algorithm that does not require an extra-gradient computation at each step and is presented in Algorithm 2. While we do not provide theoretical guarantees for A-DoG, our experiments demonstrate its efficacy in practice. The main challenge in proving guarantees for A-DoG appears to lie in deriving a suboptimality bound akin to Proposition 3, whose proof strongly leverages U-DoG's extra-gradient structure.

Algorithm 2: AcceleGrad-DoG (A-DoG)

Input: Initialization $z^{(0)} \in \mathcal{K}$, positive constant r_ϵ and number of iterations T.

Set $y_0 = x_0 = z_0$ and $\bar{r}_0 = r_\epsilon$

for $t = 0, 1, , \ldots, T-1$ **do**

$\alpha_t = \sum_{k=0}^{t} \bar{r}_k / \bar{r}_t$

$g_t \sim \mathcal{G}(x_{t+1})$

$\eta_t = \frac{\bar{r}_t}{\sqrt{\sum_{k=0}^{t} \alpha_k^2 \|g_k\|^2}}$

$x_{t+1} = \frac{\alpha_t}{\sum_{k=0}^{t} \alpha_k} z_t + \left(1 - \frac{\alpha_t}{\sum_{k=0}^{t} \alpha_k}\right) y_t$

$y_{t+1} = x_{t+1} - \eta_t g_t$

$z_{t+1} = \Pi_{\mathcal{K}}(z_t - \alpha_t \eta_t g_t)$

$\bar{r}_{t+1} = \max\{\bar{r}_t, \|z_{t+1} - z_0\|\}$

end

return x_T ▷ returning y_T gives similar results in practice

G.3. Convex experiments

The bulk of our experiments focus on smooth stochastic convex optimization problems, matching our theoretical assumptions.

Multiclass logistic regression. We experiment with multi-class logistic regression on multiple tasks from the VTAB benchmark and the LIBSVM [12] suite (a full list is given in Section G.5). For VTAB tasks we use features obtained from a pretrained ViT-B/32 [20] model (i.e., perform linear probes), and for LIBSVM tasks we use apply logistic regression directly on the features provided.

Figures 2, 4, 6, 8, 10, 12, 14 and 16 show a view of the results for different datasets analogous to Figure 1. Figures 3, 5, 7, 9, 11, 13, 15 and 17 give a complementary view by providing training curves at different batch sizes. As discussed in Section 5, we find that both U-DoG and A-DoG are competitive with well-tuned accelerated SGD (ASGD) and often significantly outperform DoG and tuned SGD. This is especially true for the training loss (for which our theory directly holds) and at large batch sizes, with A-DoG outperforming U-DoG in most cases, as both algorithms take advantage of the reduced variance in the gradient estimates to scale effectively with the batch size, as the theory suggests. In most experiments A-DoG attain and tuned ASGD attain superior convergence rate in terms of test accuracy as well as train loss; the only exception is CIFAR-100 (Figures 4 and 5, bottom rows) where the test accuracy does not closely track the train loss.

Least-squares. We modify the loss on a subset of the previous experiments to least squares, learned over a one-hot encoding of the features. We use features obtained from a pretrained ViT-B/32, similar to what we used for the multiclass logistic regression. We find that our algorithms perform well in this setting as well. In comparison, while SGD and ASGD can perform well when tuned correctly, they become more sensitive to the choice of step size and momentum, performing poorly when not properly tuned and sometimes diverging completely. Similar to the other experiments, the results are given in Figures 18 to 21.

Noiseless quadratic experiments. As a final experiment, we compare the performance of the different algorithms on the quadratic function $f(x) = \sum_{i=1}^{n} \left(\frac{i}{2n} x_i^2 + x_i\right)$ with $n = 10^4$. The results agree with the theoretical analysis, with all algorithms reaching the optimal solution or very close to it, barring GD and AGD with excessively high momentum and learning rate. Results are depicted in Figure 22.

G.4. Non-convex experiments

While we mainly focus on demonstrating the effectiveness of U-DoG and A-DoG in settings that match our theoretical analysis, we also perform preliminary experimentation in practical scenarios, namely training neural networks on datasets of moderate scales. In particular, we train a ResNet-50 [24] from scratch on a subset of the VTAB benchmark (Figures 23 to 27). Additionally, we repeat two experiments from [27]: fine-tuning a CLIP model [53] on ImageNet (Figure 28), and training a WideResnet-28-10 [66] model from scratch on CIFAR-10 (Figure 29). We observe that U-DoG often fails to converge to competitive results, while A-DoG is quite competitive with DoG on the VTAB tasks, but under-performs it for CIFAR-10 and ImageNet fine-tuning, indicating that it is not a yet a viable general-purpose neural network optimizer.

G.5. Implementation details

Environment settings. All of our experiments were based on `PyTorch` [51] (version 1.12.0). For DoG and the implementation of polynomial-decay model averaging [57], we used the `dog-optimizer` package (version 1.0.3) [27]. For ASGD, we used the native `PyTorch` SGD[5] with the Nesterov option enabled.

VTAB experiments were based on the PyTorch Image Models (`timm`, version0.7.0dev0) repository [63], with `TensorFlow datasets` (version 4.6.0) as a dataset backend [1]. LIBSVM [12] experiments were based on the `libsvmdata` (version 0.4.1) package.

5. https://pytorch.org/docs/stable/generated/torch.optim.SGD.html

To support the training and analysis of the results, we used `numpy` [23], `scipy` [61], `pandas` [62] and `scikit-learn` [52].

As much as possible, we leveraged existing recipes as provided by `timm` to train the models.

Datasets. The subset of datasets used in our VTAB experiments are: **CIFAR-100** [34], **CLEVR-Dist** [29], **DMLab** [5], **Resisc45** [13], **Sun397** [64, 65], and **SVHN** [46]. From LIBSVM, we used the **Pendigits** [2] and **Covertype** [8] datasets, where cover covertype we used the scaled features version (i.e., `covtype.scale`). We also experiment with **CIFAR-10** [34] and **ImageNet** [18].

Models. The computer vision pre-trained models were accessed via `timm`. The strings used to load the models were: 'resnet50', 'vit_base_patch32_224_in21k'.

Complexity measure. To fairly compare all algorithms, we measure complexity by the number of batches evaluated, i.e., the number of stochastic gradient queries performed by the algorithm. U-DoG requires two batches per iteration while the rest of the algorithms we consider require only one. We note that the algorithms we compare also have different memory footprints and runtimes per iteration (by constant factors). We focus on the number of batches as our complexity metric since it is most relevant to our theory. Memory and per-iteration runtime optimizations are potentially possible for U-DoG and A-DoG; we leave investigating those to future work.

ASGD model selection. In the convex optimization experiments, we run (A)SGD over a wide range of momentum and learning rate parameters. For the batch size scaling figures (e.g., the left panels in Figure 1), we pick the parameters that reach the target metric in the smallest number of batches, providing a conservative upper bound on the performance obtainable with a very carefully tuned algorithm. The learning curve figures adjacent to the batch size scaling figures (e.g., the middle panels in Figure 1) show the learning curve for the (A)SGD run attaining the best target performance at the batch size indicated. For plots of learning curves at different batch sizes (e.g., Figure 19), we select the (A)SGD parameters that are the first to reach 95% of the best metric attained by A-DoG. If no such parameters exist, we take the parameters that reach the best performance within the iteration budget.

Iterate averaging. When evaluating test accuracy, we follow Ivgi et al. [27] and apply polynomial-decay weight averaging [57] with parameter 8. We did not tune this parameter or comprehensively check how beneficial the averaging is. Nevertheless, a cursory examination of our data suggests that averaging is mostly helpful across the board, but much more so for DoG and SGD than their accelerated counterparts. This is in line with the theory, which provides guarantees on (essentially) the last iterate of U-DoG, but only the averaged iterate of DoG.

Learning rate schedule. We use a constant learning rate schedule for (A)SGD. We do not use a decaying schedule such as cosine decay [37] as it would complicate comparing the smallest number of steps required to reach a target metric, since a decaying schedule requires knowing the number of steps in advance. Preliminary experiments indicate that, in the settings we study, cosine decay is not significantly better than a constant schedule combined with iterate averaging.

Setting r_ϵ. Similarly to Ivgi et al. [27] we set $r_\epsilon = \gamma(1 + \|x_0\|)$ with $\gamma = 10^{-6}$. Our theoretical analysis suggests that the particular choice of r_ϵ does not matter as long as it is sufficiently small relative to the distance between the weight initialization x0 and the optimum.

Weight decay. We do not use weight decay in most experiments, except for training from scratch on CIFAR-10 (Figure 29), where we use a weight decay of $5 \cdot 10^{-4}$. For DoG we decay the parameters toward zero, while for U-DoG and A-DoG we decay the parameters toward the initial point x_0. That is, for DoG we add $5 \cdot 10^{-4} x$ to the stochastic gradient evaluated at x, while for U-DoG and A-DoG we add $5 \cdot 10^{-4}(x - x_0)$.

Gradient accumulation. Due to GPU memory limitations, in the non-convex experiments, for large batch sizes we divide each batch into smaller sub-batches of size of either 128 or 256 samples. We calculate the gradient for each sub-batch and average those into a single gradient which we then use to perform a single step. When batch normalization is used (that is, for ResNet50), this is not mathematically identical to computing the gradient in one large batch.

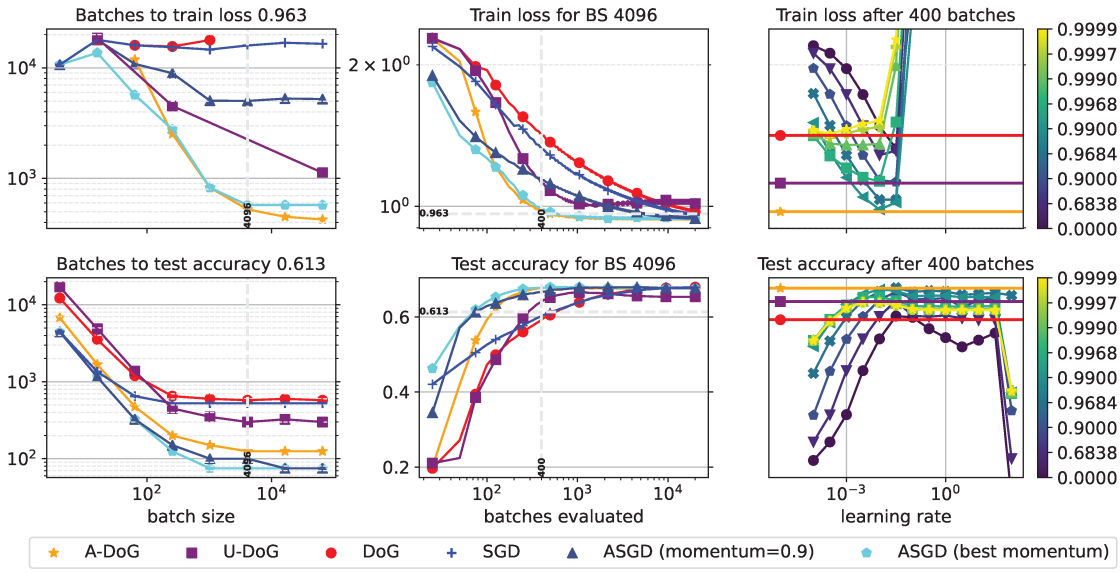

Figure 2: Training a linear model with ViT-32 features and log loss on SVHN. Top: Train loss. Bottom: Test accuracy after iterate averaging. First column: Batch size scaling of complexity to reach target performance. Second column: Learning curves. Third column: ASGD performance at all learning rates and momenta, contrasted with DoG variants.

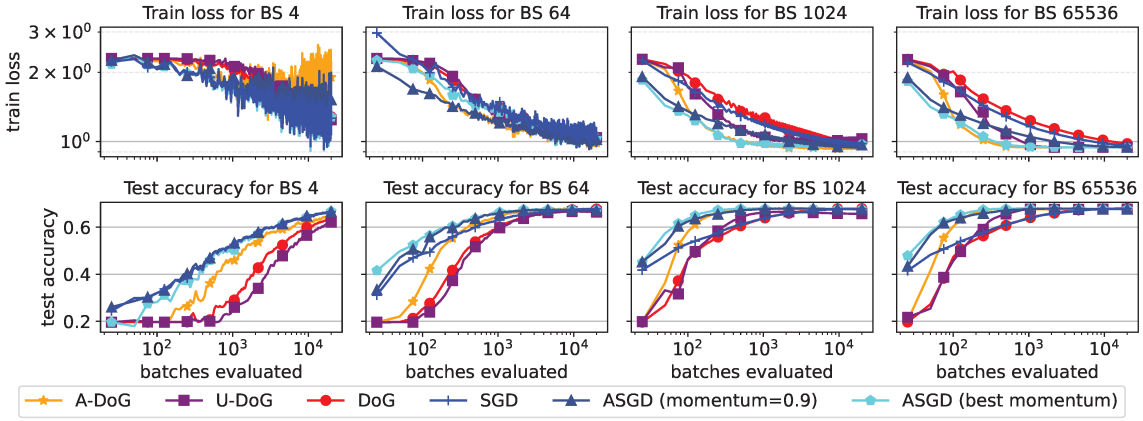

Figure 3: Training a linear model with ViT-32 features and log loss on SVHN. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy of averaged model vs. batches processed for different batch sizes.

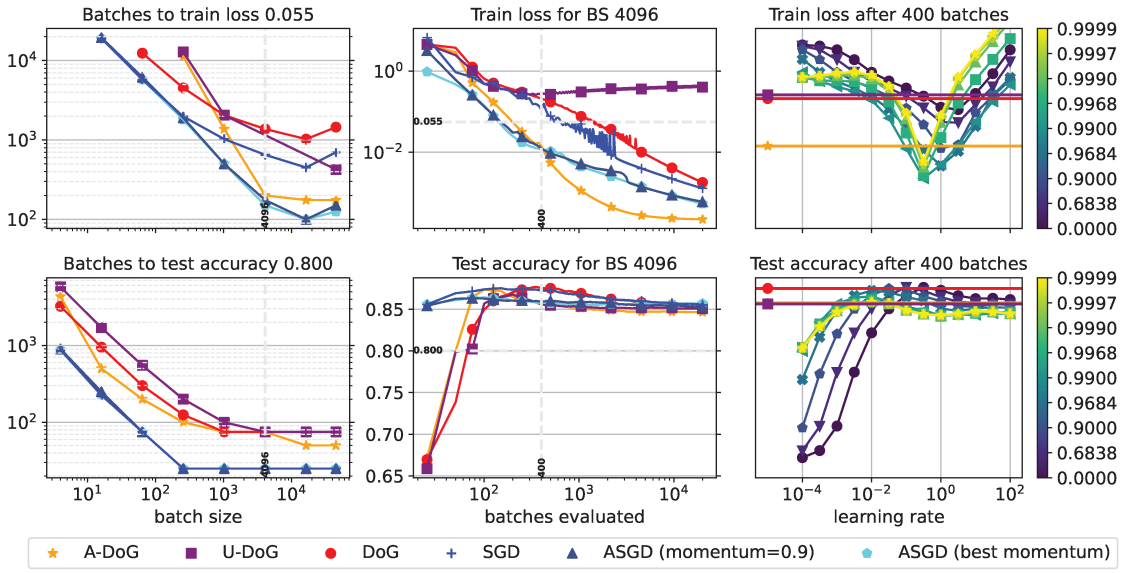

Figure 4: Training a linear model with ViT-32 features and log loss on CIFAR-100. Top: Train loss. Bottom: Test accuracy after iterate averaging. First column: Batch size scaling of complexity to reach target performance. Second column: Learning curves. Third column: ASGD performance at all learning rates and momenta, contrasted with DoG variants.

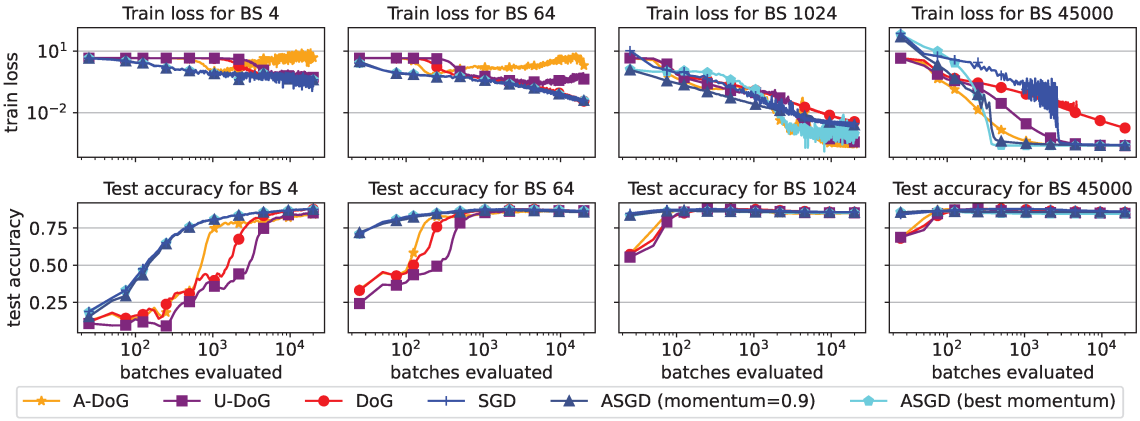

Figure 5: Training a linear model with ViT-32 features and log loss on CIFAR-100. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy of averaged model vs. batches processed for different batch sizes.

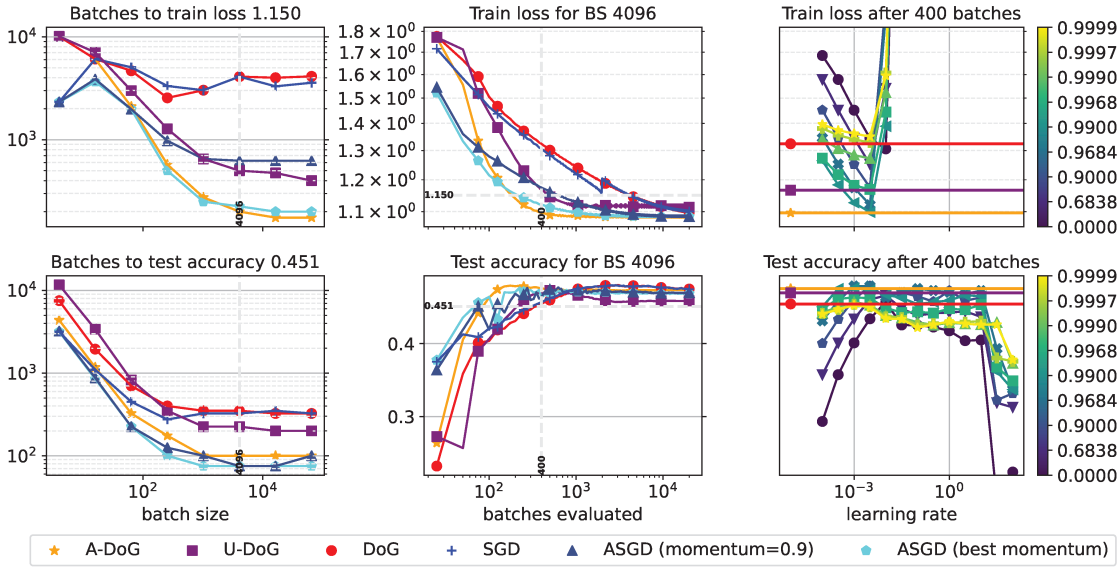

Figure 6: Training a linear model with ViT-32 features and log loss on DMLab. Top: Train loss. Bottom: Test accuracy after iterate averaging. First column: Batch size scaling of complexity to reach target performance. Second column: Learning curves. Third column: ASGD performance at all learning rates and momenta, contrasted with DoG variants.

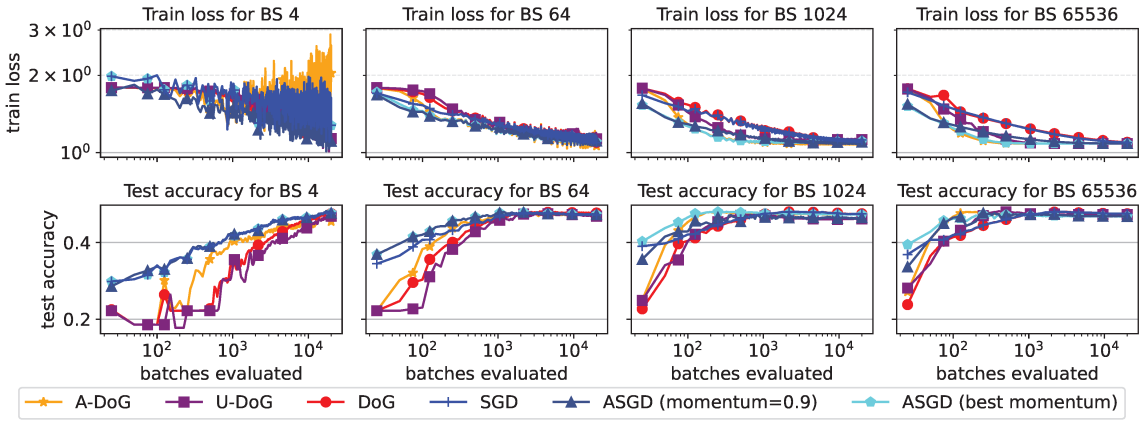

Figure 7: Training a linear model with ViT-32 features and log loss on DMLab. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy of averaged model vs. batches processed for different batch sizes.

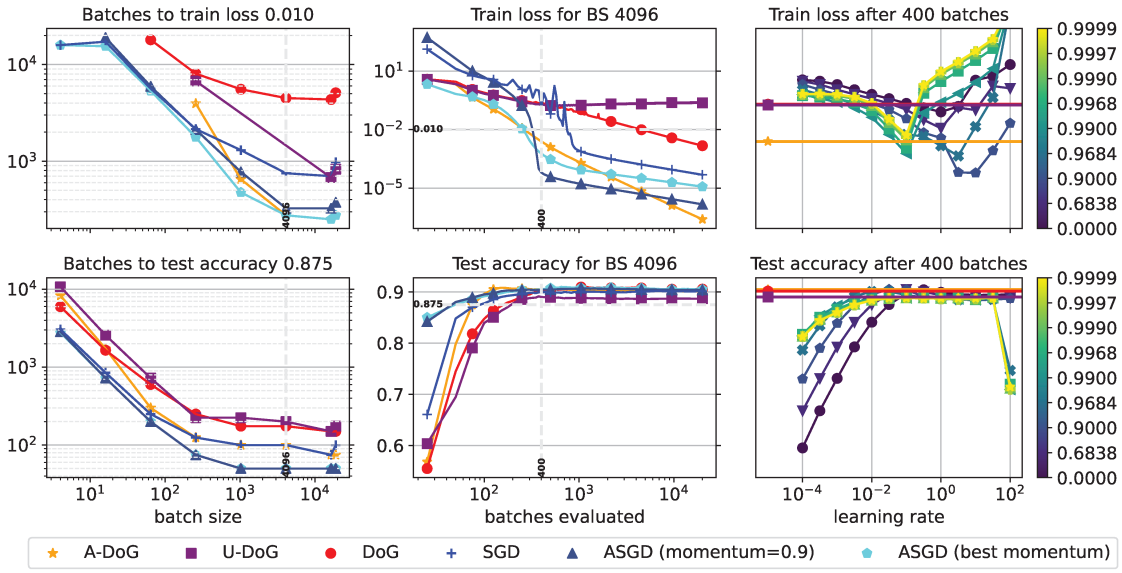

Figure 8: Training a linear model with ViT-32 features and log loss on Resisc45. Top: Train loss. Bottom: Test accuracy after iterate averaging. First column: Batch size scaling of complexity to reach target performance. Second column: Learning curves. Third column: ASGD performance at all learning rates and momenta, contrasted with DoG variants.

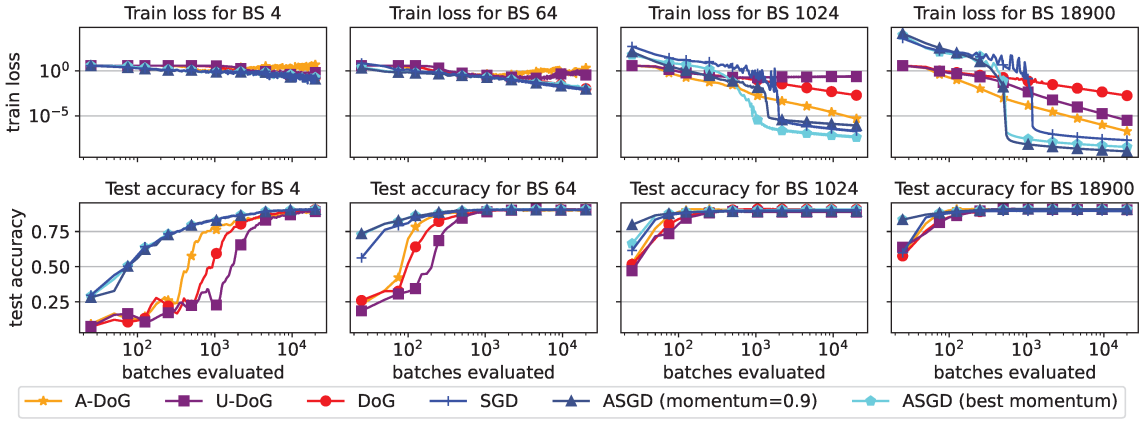

Figure 9: Training a linear model with ViT-32 features and log loss on Resisc45. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy of averaged model vs. batches processed for different batch sizes.

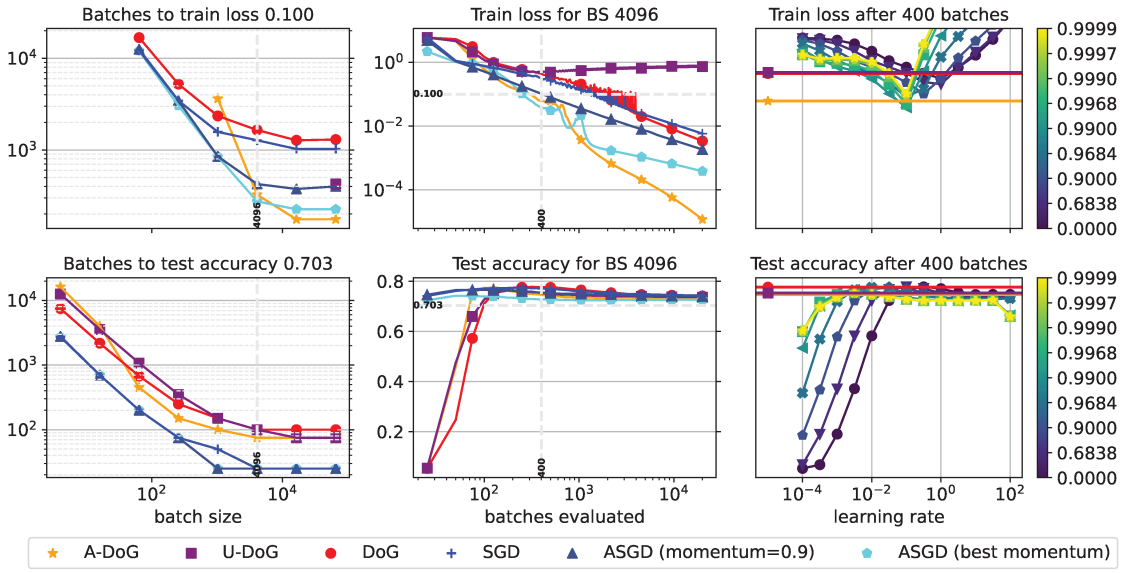

Figure 10: Training a linear model with ViT-32 features and log loss on Sun397. Top: Train loss. Bottom: Test accuracy after iterate averaging. First column: Batch size scaling of complexity to reach target performance. Second column: Learning curves. Third column: ASGD performance at all learning rates and momenta, contrasted with DoG variants.

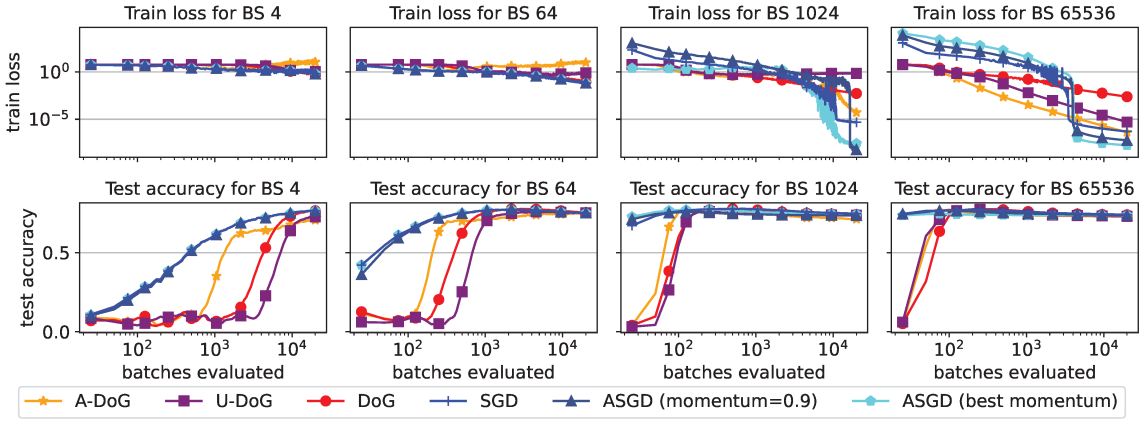

Figure 11: Training a linear model with ViT-32 features and log loss on Sun397. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy of averaged model vs. batches processed for different batch sizes.

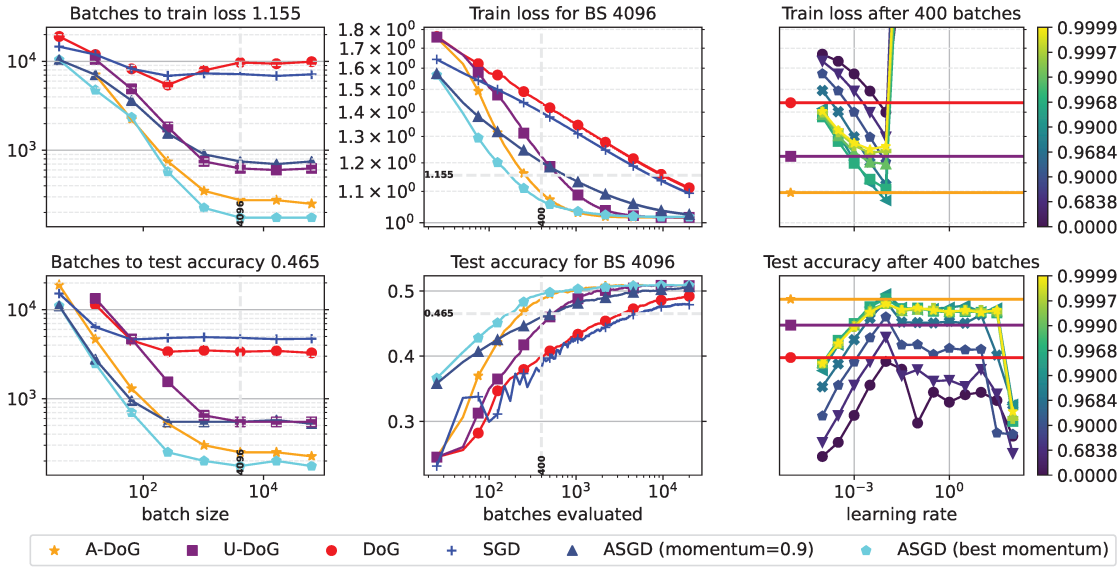

Figure 12: Training a linear model with ViT-32 features and log loss on CLEVR-Dist. Top: Train loss. Bottom: Test accuracy after iterate averaging. First column: Batch size scaling of complexity to reach target performance. Second column: Learning curves. Third column: ASGD performance at all learning rates and momenta, contrasted with DoG variants.

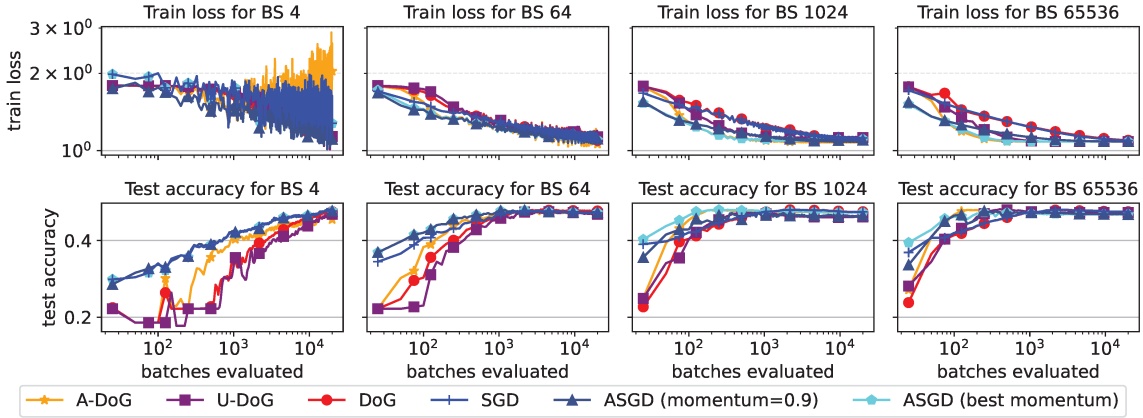

Figure 13: Training a linear model with ViT-32 features and log loss on CLEVR-Dist. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy of averaged model vs. batches processed for different batch sizes.

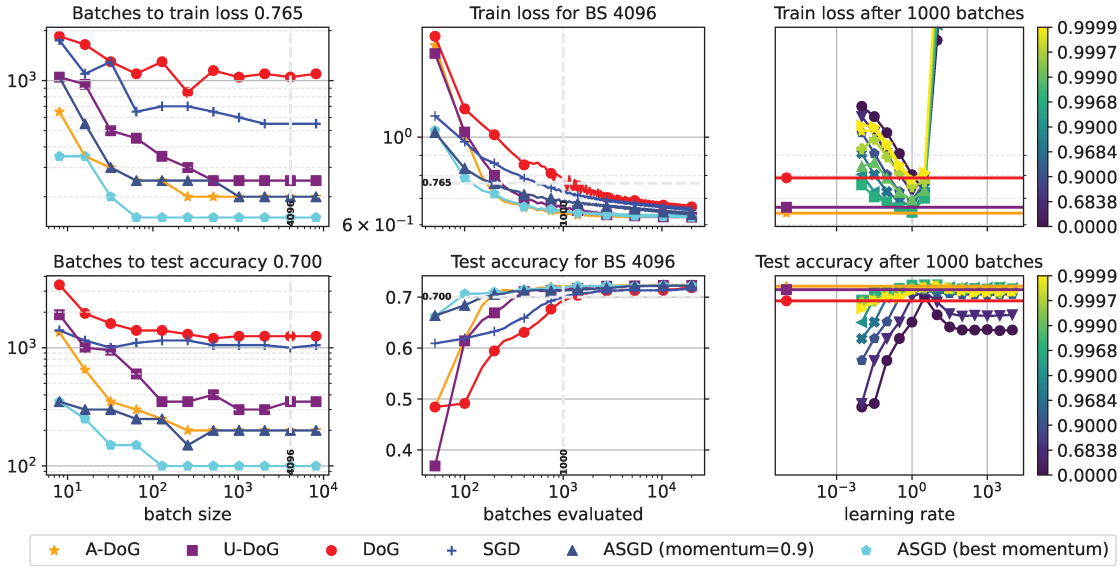

Figure 14: Training a linear model with log loss on LIBSVM/CovertypeScale. Top: Train loss. Bottom: Test accuracy after iterate averaging. First column: Batch size scaling of complexity to reach target performance. Second column: Learning curves. Third column: ASGD performance at all learning rates and momenta, contrasted with DoG variants.

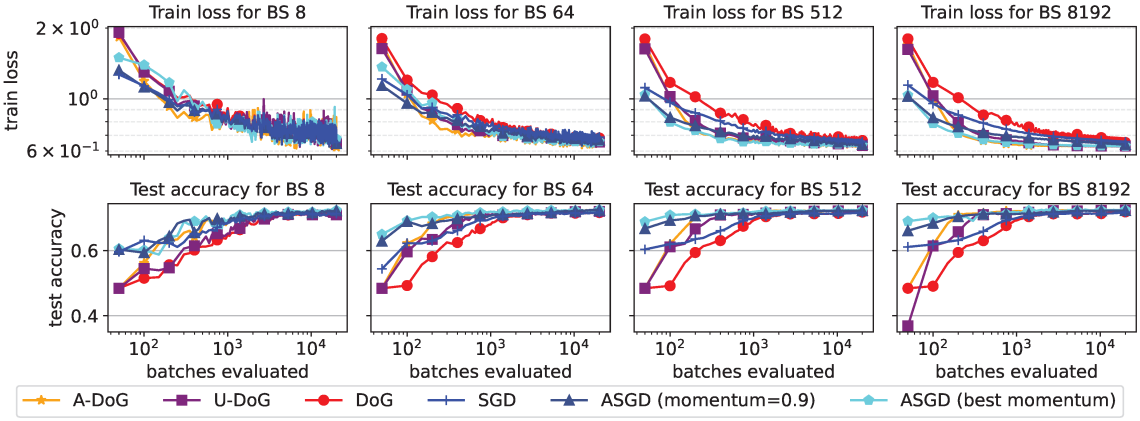

Figure 15: Training a linear model with log loss on LIBSVM/CovertypeScale. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy of averaged model vs. batches processed for different batch sizes.

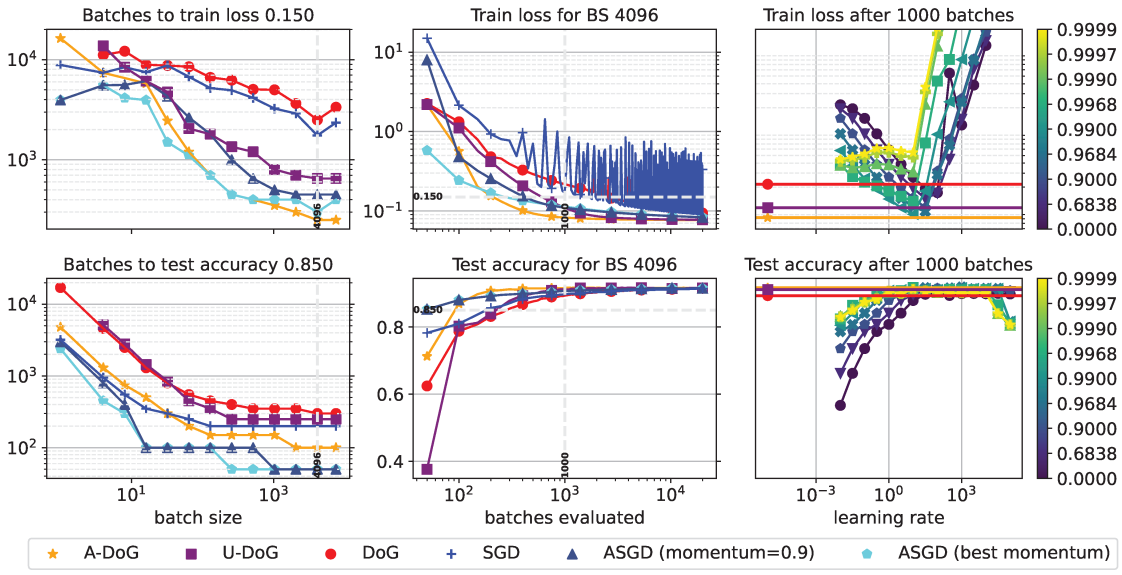

Figure 16: Training a linear model with log loss on LIBSVM/Pendigits. Top: Train loss. Bottom: Test accuracy after iterate averaging. First column: Batch size scaling of complexity to reach target performance. Second column: Learning curves. Third column: ASGD performance at all learning rates and momenta, contrasted with DoG variants.

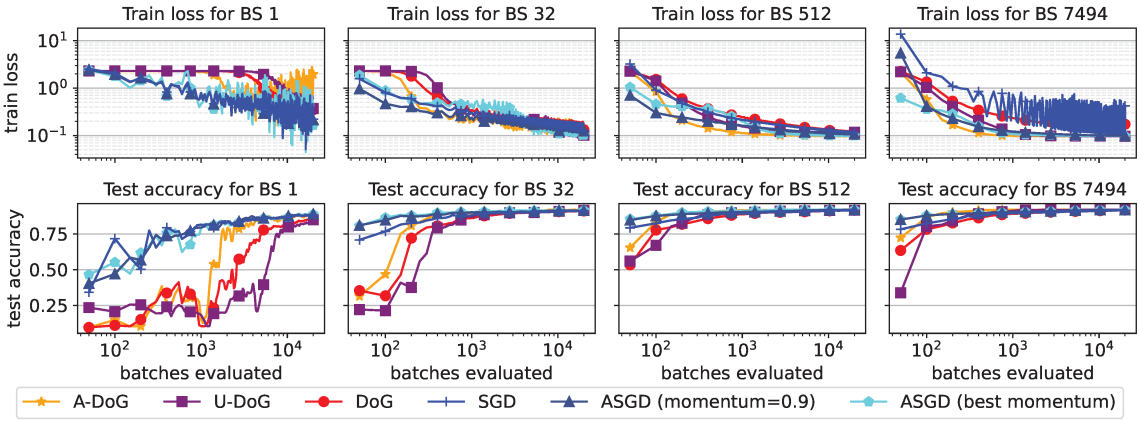

Figure 17: Training a linear model with log loss on LIBSVM/Pendigits. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy of averaged model vs. batches processed for different batch sizes. As most algorithms here fail to converge at reasonable rate, we use significantly lower targets to choose hyper-parameters.

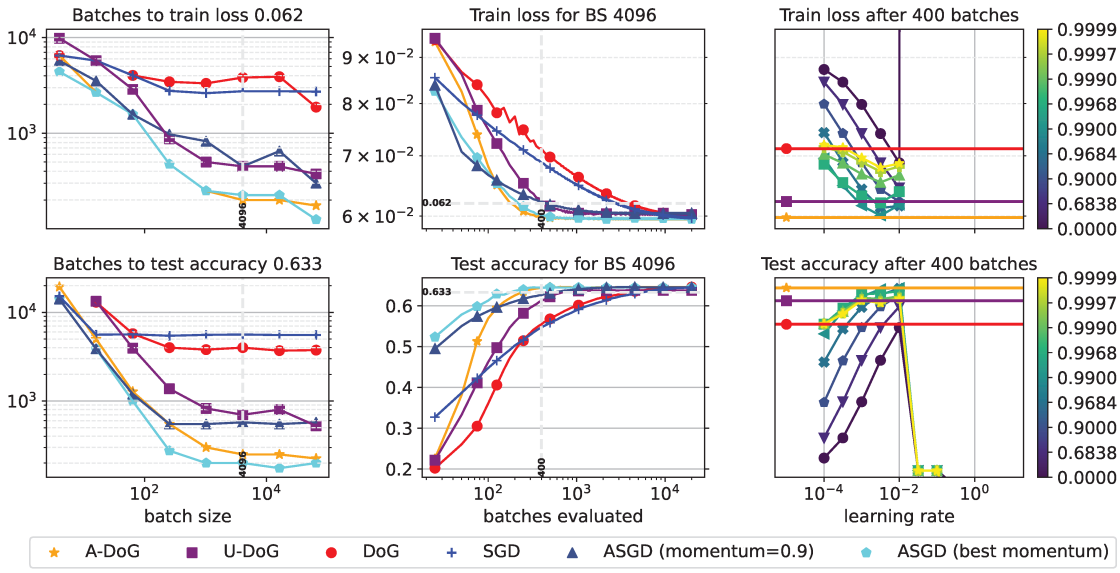

Figure 18: Training a linear model with ViT-32 features and least-squares loss on SVHN. Top: Train loss. Bottom: Test accuracy after iterate averaging. First column: Batch size scaling of complexity to reach target performance. Second column: Learning curves. Third column: ASGD performance at all learning rates and momenta, contrasted with DoG variants. This is the same as Figure 1.

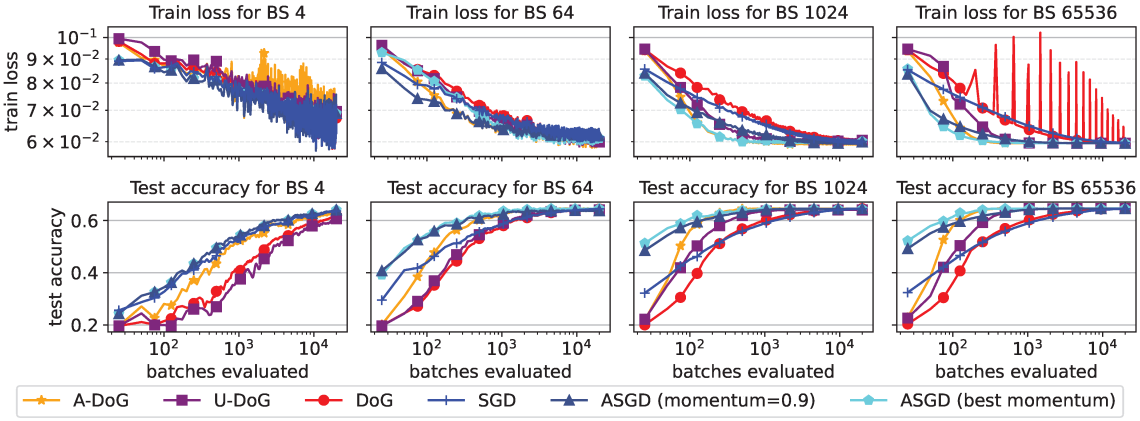

Figure 19: Training a linear model with ResNet50 features and least-squares loss on SVHN. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy of averaged model vs. batches processed for different batch sizes.

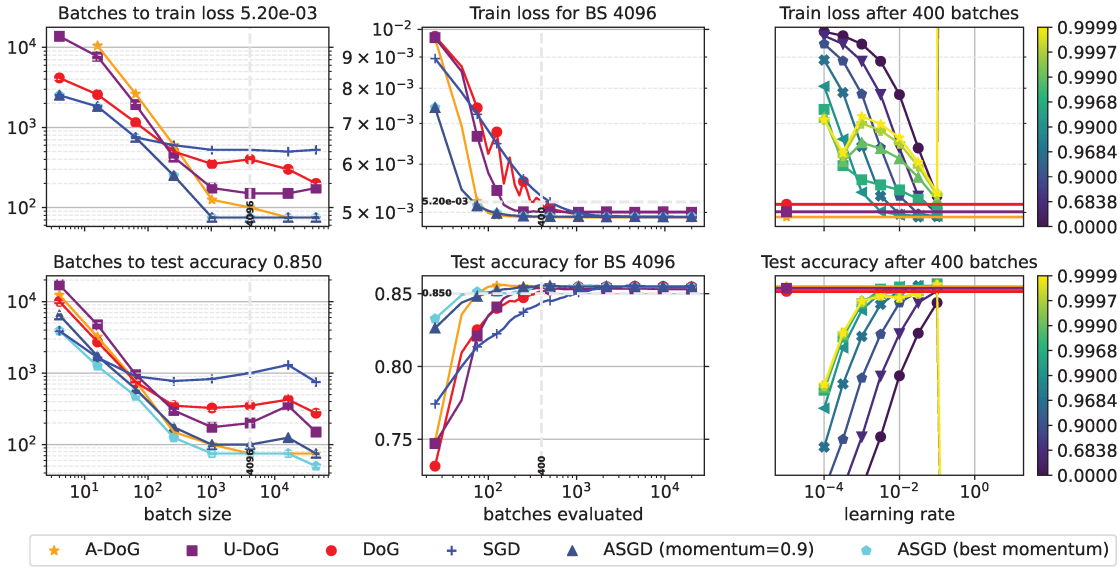

Figure 20: Training a linear model with ViT-32 features and least-squares loss on CIFAR-100. Top: Train loss. Bottom: Test accuracy after iterate averaging. First column: Batch size scaling of complexity to reach target performance. Second column: Learning curves. Third column: ASGD performance at all learning rates and momenta, contrasted with DoG variants.

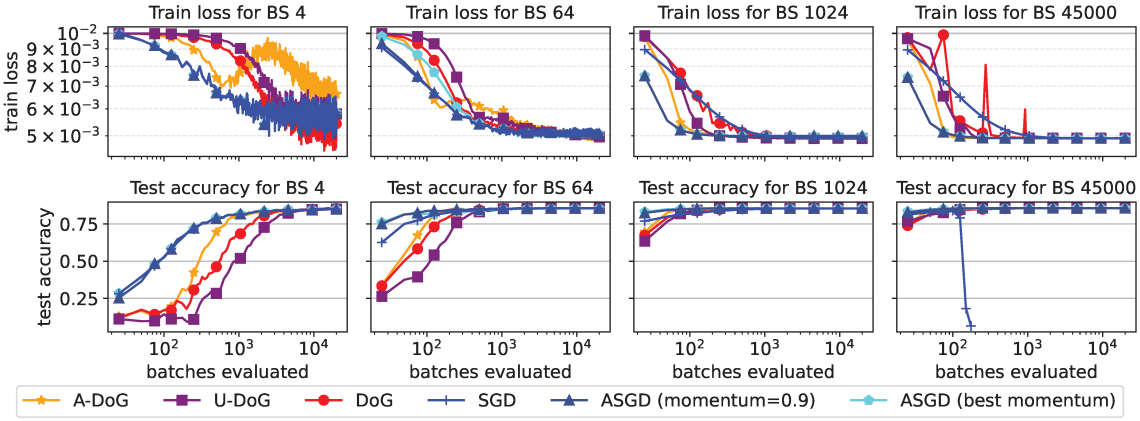

Figure 21: Training a linear model with ResNet50 features and least-squares loss on CIFAR-100. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy of averaged model vs. batches processed for different batch sizes.

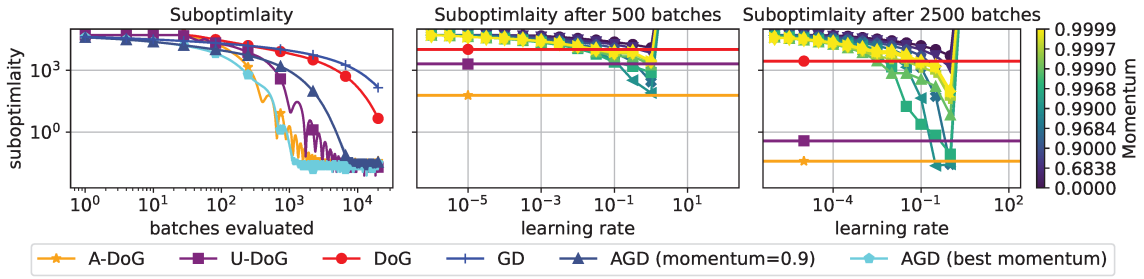

Figure 22: Training a model on a noiseless quadratic problem. At larger base learning rates, all AGD variants diverge while DoG variants remain stable, and U-DoG and A-DoG perform especially well.

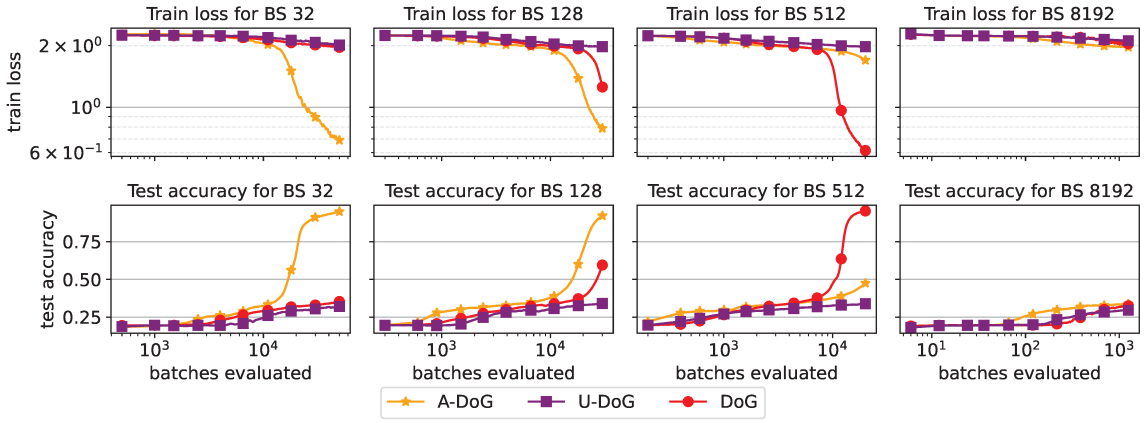

Figure 23: Training a ResNet50 model from scratch on SVHN. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy vs. batches processed for averaged iterates at varied batch sizes.

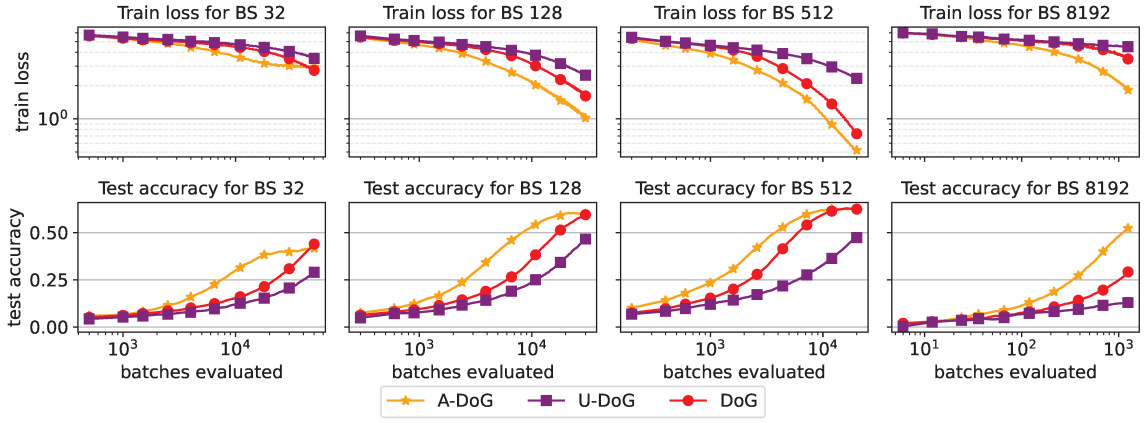

Figure 24: Training a ResNet50 model from scratch on Sun397. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy vs. batches processed for averaged iterates at varied batch sizes.

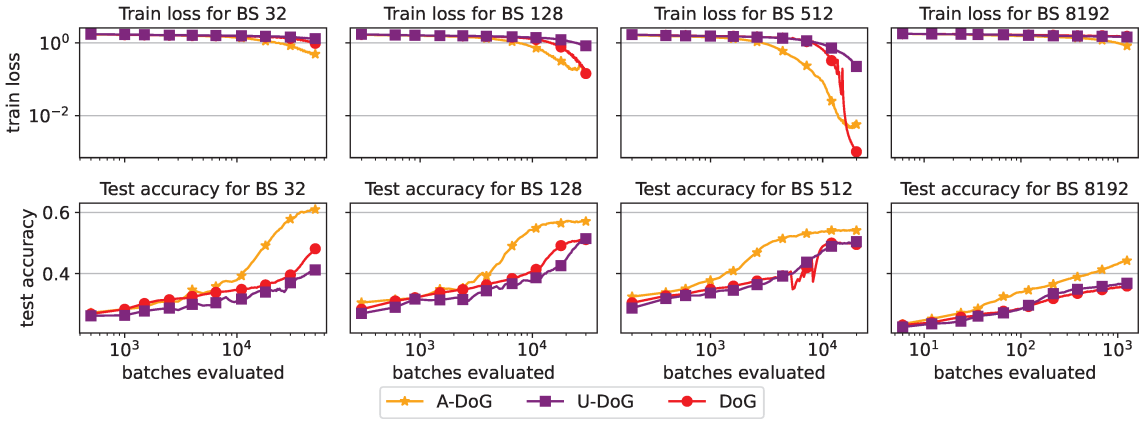

Figure 25: Training a ResNet50 model from scratch on DMLab. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy vs. batches processed for averaged iterates at varied batch sizes.

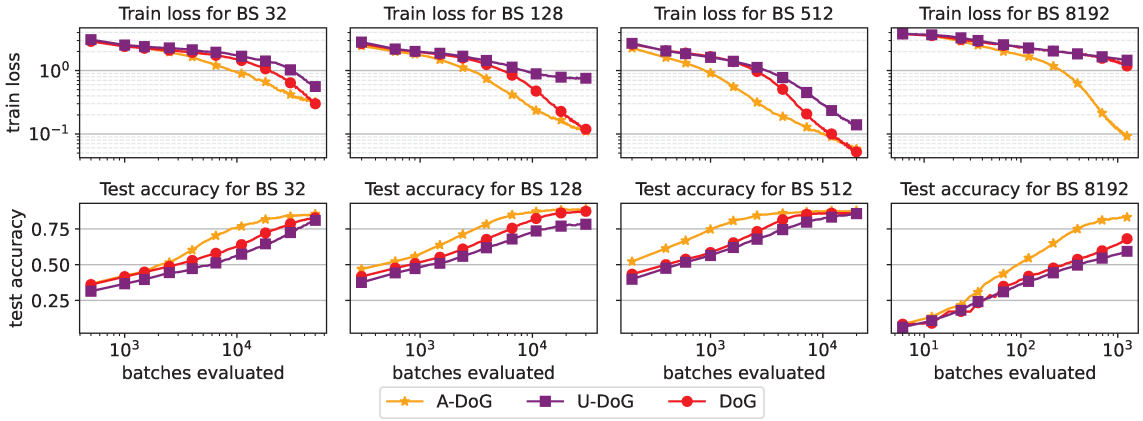

Figure 26: Training a ResNet50 model from scratch on Resisc45. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy vs. batches processed for averaged iterates at varied batch sizes.

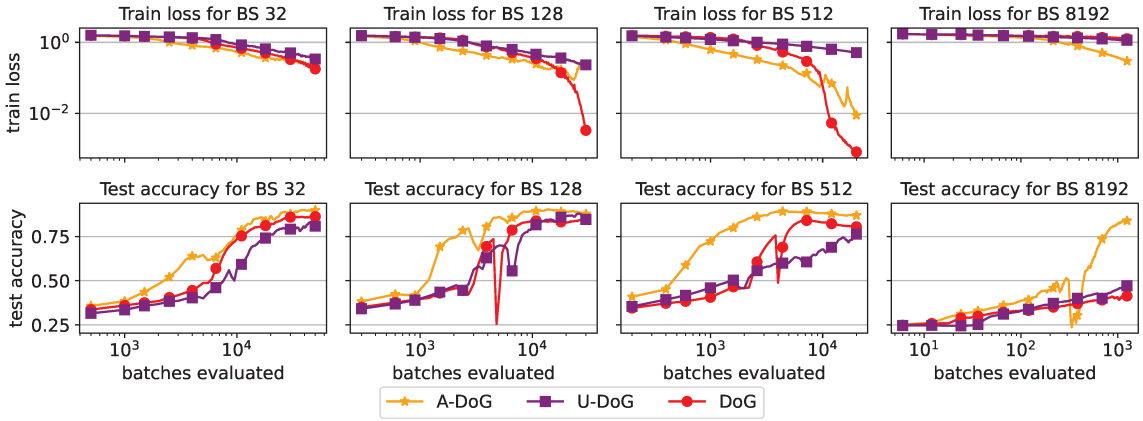

Figure 27: Training a ResNet50 model from scratch on CLEVR-Dist. Top: Loss vs. batches processed training for different batch sizes. Bottom: Test accuracy vs. batches processed for averaged iterates at varied batch sizes.

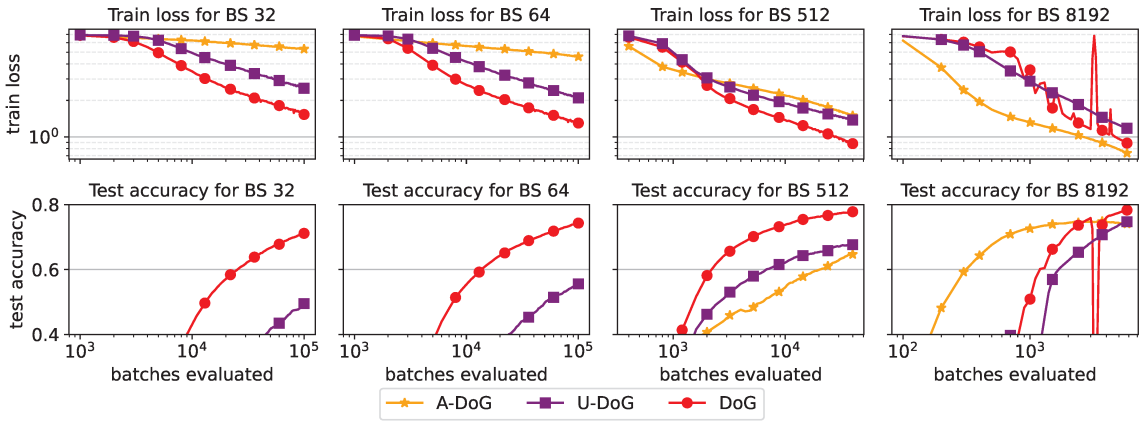

Figure 28: Fine-tuning a Clip-ViT-B/32 model on ImageNet, at different batch sizes. Top: Loss vs. step training curve for different batch sizes. Bottom: Test accuracy vs. step curve for averaged iterates at varied batch sizes.

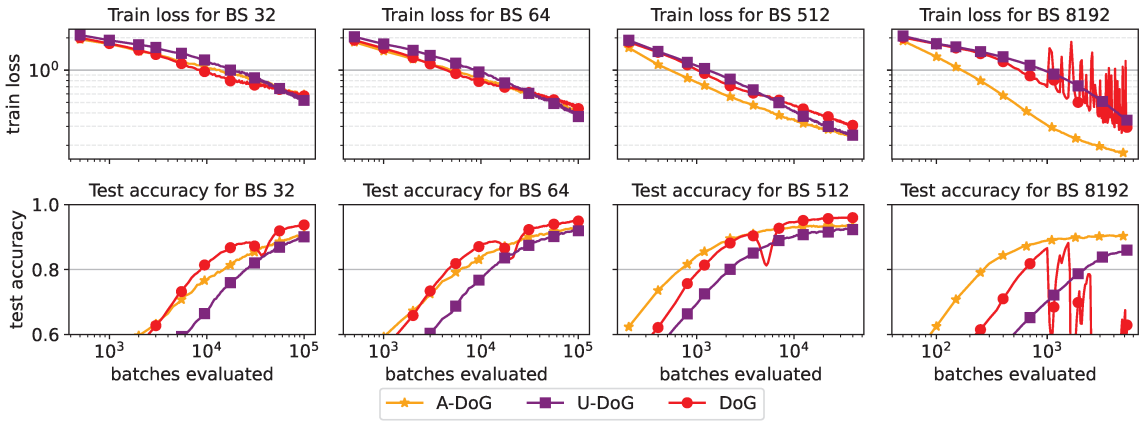

Figure 29: Training a Wide-ResNet-28-10 model on CIFAR-10 from scratch, at different batch sizes. Top: Loss vs. step training curve for different batch sizes. Bottom: Test accuracy vs. step curve for averaged iterates at varied batch sizes.