OPT2021: 13th Annual Workshop on Optimization for Machine Learning

Stochastic Polyak Stepsize with a Moving Target

Robert M. Gower GOWERROBERT@GMAIL.COM
LTCI, Télécom Paris, IPP and Center for Computational Mathematics, Flatiron Institute

Aaron Defazio ADEFAZIO@FB.COM
Michael Rabbat MIKERABBAT@FB.COM
Facebook AI Research

Abstract

We propose a new stochastic gradient method that uses recorded past loss values to compute adaptive stepsizes. Our starting point is to show that the SP (Stochastic Polyak) method directly exploits interpolated models. That is, SP is a subsampled Newton-Raphson method applied to solving certain *interpolation equations*. These interpolation equations only hold for models that interpolate the data. We then use this viewpoint to develop a new variant of the SP method that converges without interpolation called MOTAPS. The MOTAPS method uses n auxiliary variables, one for each data point, that track the loss value for each data point. These auxiliary variables and the loss values are then used to set the step size. We provide a global convergence theory for MOTAPS by showing that it can be interpreted as a special variant of online SGD. We also perform several numerical experiments on convex learning problems, and non-convex learning problem based on image classification and language translation. In all of our tasks we show that MOTAPS is competitive with the relevant baseline method.

1. Introduction

Consider the problem

$$w^* \in \operatorname*{argmin}_{w \in \mathbb{R}^d} f(w) := \frac{1}{n} \sum_{i=1}^{n} f_i(w), \tag{1}$$

where each $f_i(w)$ represents the loss of a model parametrized by $w \in \mathbb{R}^d$ over a given ith *data point*. We assume that there exists a solution $w^* \in \mathbb{R}^d$. Let $\mathcal{W}^*$ denote the set of minimizers of (1).

An ideal method for solving (1) is one that exploits the sum of terms structure, has easy-to-tune hyperparameters, and is guaranteed to converge. *Stochastic gradient descent* (SGD) exploits this sum of terms structure by only using a single stochastic gradient (or a batch) $\nabla f_i(w)$ per iteration. Because of this, SGD is efficient when the *number of data points* n is large, and can even be applied when n is infinite and (1) is an expectation over a continuous random variable.

The downside of SGD is that it is difficult to tune because it requires tuning a sequence of step sizes, otherwise known as a learning rate schedule. Indeed, to make SGD converge, we need a sequence of step sizes that must converge to zero at just the right rate. Here we develop methods with adaptive step sizes that use the loss values to set the stepsize.

We derive our new adaptive methods by first exploiting the *interpolation equations* given by

$$f_i(w) \;=\; 0, \quad \text{for } i = 1, \ldots, n. \tag{2}$$

We say that the *interpolation assumption* holds if there exists $w^* \in \mathcal{W}^*$ that solves (2). Two well known settings where the interpolation assumption holds are 1) for binary classification with a linear model where

the data can be separated by a hyperplane [7] or 2) when we know that each $f_i(w)$ is non-negative, and we have enough parameters in our model so there exists a solution that fits all data points. This second setting is often referred to as the overparametrized regime [32], and it is becoming a common occurrence in several sufficiently overparametrized deep neural networks [3, 36].

Our starting point is to observe that the Stochastic Polyak method (SP) [4, 25] directly exploits and solves the interpolation equations (2). Indeed, the SP method is a subsampled Newton Raphson method [35] as we show next.

The subsampled Newton Raphson method at each iteration samples a single index $i \in \{1, \ldots, n\}$ and focuses on solving the single equation $f_i(w) = 0$. This single equation can still be difficult to solve since $f_i(w)$ can be highly nonlinear. So we linearize f_i around a given $w^t \in \mathbb{R}^d$, and set the linearization of $f_i(w)$ to zero, that is

$$f_i(w^t) + \left\langle \nabla f_i(w^t), w - w^t \right\rangle = 0.$$

This is now a linear equation in $w \in \mathbb{R}^d$ that has d unknowns and thus has infinite solutions. To pick just one solution we use a projection step as follows

$$w^{t+1} = \text{argmin}_{w \in \mathbb{R}^d} \left\| w - w^t \right\|^2 \quad \text{subject to } f_i(w^t) + \left\langle \nabla f_i(w^t), w - w^t \right\rangle = 0. \tag{3}$$

The solution to this projection step (See Lemma 7 for details) is given by

$$w^{t+1} = w^t - \frac{f_i(w^t)}{\|\nabla f_i(w^t)\|^2} \nabla f_i(w^t). \tag{4}$$

This method (6) is known as the Stochastic Polyak method [25] [1]. The SP has many desirable properties: It is incremental, it adapts it step size according to the current loss function, and it enjoy several invariance properties. Thus in many senses the SP is an ideal stochastic method. The obvious downside is that to arrive at (4) we have to assume that the interpolation assumption holds. The main objective of this paper is design methods akin to the SP method that do not rely on the interpolation assumption.

2. The Stochastic Polyak Method

We start by presenting the SP (Stochastic Polyak method) through two different viewpoints. First, we show that SP is a special case of the subsampled Newton-Raphson method [35]. Using this first viewpoint, and leveraging results from [35], we then go on to show that SP can also be viewed as a type of *online SGD method*, which greatly facilitates the analysis of SP.

2.1. The Newton-Raphson viewpoint

As observed in the introduction in Section 1, the SP method is designed for solving interpolation equations. Here we formalize and extend this observation before moving on to our new methods.

We can derive an extended form of the SP method that does not rely on the interpolation assumption. Instead of the interpolation assumptions, let us assume for now that we have access to the loss values $f_i(w^*)$ for each $i = 1, \ldots, n$, where $w^* \in \mathcal{W}$. If we knew the $f_i(w^*)$'s then we can solve the optimization problem (1) by solving instead the nonlinear equations

$$f_i(w) = f_i(w^*), \quad \text{for } i = 1, \ldots, n. \tag{5}$$

1. In [4] the authors also observed that the full batch Polyak stepsize in 1D is a Newton Raphson method.

Applying the subsampled Newton Raphsons method as in Section 1, we arrive at the method

$$w^{t+1} = w^t - \frac{f_i(w^t) - f_i(w^*)}{\|\nabla f_i(w^t)\|^2}\nabla f_i(w^t). \tag{6}$$

This method is a minor extension of (4) where-in we now allow $f_i(w^*) \neq 0$. Despite this minor change, we will also refer to (6) as the Stochastic Polyak method[2].

The issue with the SP method is that we often will not know $f_i(w^*)$ excluding the cases where $f_i(w^*) = 0$. Outside of this setting, it is unlikely that we would have access to each $f_i(w^*)$. Thus we relax this requirement in Sections 3 and 4. But first, we present yet another viewpoint of SP as a type of online SGD method.

2.2. The SGD viewpoint

Fix a given $w^t \in \mathbb{R}^d$ and consider the following *auxiliary* objective function

$$\min_{w\in\mathbb{R}^d} h_t(w) \;\coloneqq\; \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\frac{(f_i(w) - f_i(w^*))^2}{\|\nabla f_i(w^t)\|^2}. \tag{7}$$

Here we use the pseudoinverse convention that if $\left\|\nabla f_i(w^t)\right\| = 0 \in \mathbb{R}$ then $\left\|\nabla f_i(w^t)\right\|^{-1} = 0$. Clearly $w = w^*$ is a minimizer of (7). This suggests that we could try to minimize (7) as a proxy for solving the equations (1). Since (7) is a sum of terms that depends on t, we can use an online SGD to minimize (7). To describe this online SGD method let

$$h_{i,t}(w) \;\coloneqq\; \frac{1}{2}\frac{(f_i(w) - f_i(w^*))^2}{\|\nabla f_i(w^t)\|^2} \quad \text{and thus} \quad \nabla h_{i,t}(w) \;=\; \frac{f_i(w) - f_i(w^*)}{\|\nabla f_i(w^t)\|^2}\nabla f_i(w). \tag{8}$$

The online SGD method is given by sampling $i_t \in \{1, \dots, n\}$ and then iterating

$$w^{t+1} \;=\; w^t - \gamma\nabla h_{i_t,t}(w^t) \;\overset{(8)}{=}\; w^t - \gamma\frac{f_{i_t}(w^t) - f_{i_t}(w^*)}{\|\nabla f_{i_t}(w^t)\|^2}\nabla f_{i_t}(w^t), \tag{9}$$

which is equivalent to the SP method (6) but with addition of a stepsize $\gamma > 0$. This online SGD viewpoint of SP is very useful for proving convergence of SP. Indeed, there exist many convergence results in the literature on online SGD for convex, non-convex, smooth and non-smooth functions that we can now import to analyzing SP. Furthermore, it turns out that (9) enjoys a remarkable growth property that facilitates many SGD proof techniques, as we show in the next lemma.

Lemma 1 (Growth) *The functions $h_{i,t}(w)$ defined in (8) satisfy*

$$\left\|\nabla h_{i,t}(w^t)\right\|^2 \;=\; 2h_{i,t}(w^t). \tag{10}$$

Consequently due to (7) we have that

$$\frac{1}{n}\sum_{i=1}^{n}\left\|\nabla h_{i,t}(w^t)\right\|^2 \;=\; 2h_t(w^t). \tag{11}$$

2. Using $f_i(w^*)$ in the numerator is apparently new, and what we call the Stochastic Polyak method here is not the same as the Stochastic Polyak method proposed in [25]. In Section B we detail these differences. In the more common case where $f_i(w^*) = 0$, there is consensus that (6) is called the Stochastic Polyak method .

Proof. The proof follows immediately from (8) and (7) since

$$\left\|\nabla h_{i,t}(w^t)\right\|^2 \overset{(8)}{=} \frac{(f_i(w)-f_i(w^t))^2}{\left\|\nabla f_i(w^t)\right\|^4}\left\|\nabla f_i(w^t)\right\|^2 = \frac{(f_i(w^t)-f_i(w^*))^2}{\left\|\nabla f_i(w^t)\right\|^2} \overset{(8)}{=} 2h_{i,t}(w^t).\square$$

In Section E we will exploit this SGD viewpoint and the growth property in Lemma 1 to prove the convergence of SP. But first we develop new variants of SP that do not require knowing the $f_i(w^*)$'s.

3. Targeted Stochastic Polyak Steps

Now suppose that we do not know $f_i(w^*)$ for $i = 1, \ldots, n$. Instead, we only have a target value for which we would like the *total loss* to reach.

Assumption 1 (Target) *There exists a* target value $\tau \geq 0$ *such that every* $w^* \in \mathcal{W}^*$ *is a solution to the nonlinear equation*

$$f(w) \;=\; \frac{1}{n}\sum_{i=1}^{n} f_i(w) \;=\; \tau. \tag{12}$$

Using Assumption 1 we develop new variants of the SP as follows. First we re-write (12) by introducing auxiliary variables $\alpha_i \in \mathbb{R}$ for $i = 1, \ldots, n$ such that

$$\frac{1}{n}\sum_{i=1}^{n} \alpha_i \;=\; \tau, \tag{13}$$

$$f_i(w) \;=\; \alpha_i, \quad \text{for } i = 1, \ldots, n. \tag{14}$$

This reformulation *exposes* the ith loss function (and thus the ith data point) as a separate equation. Because each loss (and associated data point) is on a separate row, applying a subsampled Newton-Raphson method results in an incremental method, as we show next.

Let $w^t \in \mathbb{R}^d$ and $\alpha^t = (\alpha_1^t, \ldots, \alpha_n^t) \in \mathbb{R}^n$ be the current iterates. At each iteration we can either sample (13) or one of the equations (14). We then apply a Newton-Raphson step using just this sampled equation. For instance, if we sample one of the equations in (14), we first linearize in w and α_i around the current iterate and set this linearization to zero, which gives

$$f_i(w^t) + \left\langle \nabla f_i(w^t), w - w^t \right\rangle = \alpha_i. \tag{15}$$

Projecting the previous iterates onto this linear equation gives

$$w^{t+1}, \alpha_i^{t+1} = \underset{w\in\mathbb{R}^d, \alpha_i \in \mathbb{R}}{\operatorname{argmin}} \left\|w - w^t\right\|^2 + \left\|\alpha_i - \alpha_i^t\right\|^2 \text{ subject to } f_i(w^t) + \left\langle \nabla f_i(w^t), w - w^t \right\rangle = \alpha_i.$$

The solution[3] to the above is given by the updates in lines 8 and 9 in Algorithm 1 when $\gamma = 1$.

Alternatively, if we sample (13), projecting the current iterates onto this constraint gives

$$\alpha^{t+1} \;=\; \operatorname{argmin}_{\alpha\in\mathbb{R}^n} \left\|\alpha - \alpha^t\right\|^2 \quad \text{subject to} \quad \frac{1}{n}\sum_{i=1}^{n} \alpha_i \;=\; \tau. \tag{16}$$

3. Proven in Lemma 6.

The closed form soution to (16) is given in line 6 in Algorithm 1 when $\gamma = 1$. In Algorithm 1 we give the complete pseudocode of the subsampled Newton-Raphson method applied to (14). We refer to this algorithm as the the *Target Stochastic Polyak method*, or TAPS for short.

Algorithm 1 TAPS: TArgetted Stochastic Polyak Step

Inputs: target $\tau \geq 0$ and stepsize $\gamma > 0$

Initialize: $w^0 = 0, \alpha_i^0 = \overline{\alpha}^0 = 0$ for $i = 1, \ldots, n$

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

Sample $i \in \{1, \ldots, n+1\}$ according to some law

if $i = n+1$ **then**

$\alpha_j^{t+1} = \alpha_j^t + \gamma(\tau - \overline{\alpha}^t)$, for $j = 1, \ldots, n$ where $\overline{\alpha}^t = \frac{1}{n}\sum_{i=1}^n \alpha_i^t$.

else

$$\alpha_i^{t+1} = \alpha_i^t + \gamma \frac{f_i(w^t) - \alpha_i^t}{\|\nabla f_i(w^t)\|^2 + 1}$$

$$w^{t+1} = w^t - \gamma \frac{f_i(w^t) - \alpha_i^t}{\|\nabla f_i(w^t)\|^2 + 1} \nabla f_i(w^t).$$

Output: w^T

Remark 2 (TAPS stops at the solution) *Algorithm 2 stops when it reaches the solution. That is, if $w^t = w^*$ and $\alpha_i^t = f_i(w^*)$ for all i, then both lines 8 and 9 have no affect on w or the α_i's. Furthermore $\tau = \overline{\alpha}^t := \frac{1}{n}\sum_{i=1}^n \alpha_i^t$ and consequently the α_i's are no longer updated in line 6. This natural stopping is a sanity check that SGD does not satisfy.*

3.1. The SGD viewpoint

The TAPS method in Algorithm 1 can also be cast as an online SGD method. To see this, first we re-write (7) as the minimization of an auxiliary function

$$\min_{w \in \mathbb{R}^d, \alpha \in \mathbb{R}^n} h_t(w, \alpha) \;:=\; \frac{1}{n+1}\left(\sum_{i=1}^n \frac{1}{2}\frac{(f_i(w) - \alpha_i)^2}{\|\nabla f_i(w^t)\|^2 + 1} + \frac{n}{2}(\overline{\alpha} - \tau)^2\right). \tag{17}$$

In the following lemma we show that minimizing (17) is equivalent to minimizing (1).

Lemma 3 *Let Assumption 1 hold. Every stationary point of (17) is a stationary point of (1). Furthermore, every minimizer $(w^*, \alpha^*) \in \mathbb{R}^{d+n}$ of (17) is such that w^* is a minima of (1) and*

$$\alpha_i^* \;=\; f_i(w^*). \tag{18}$$

Consequently we have that $h_t(w^*, \alpha^*) = 0$.

The proof of this lemma, and all subsequent lemmas are in the appendix in Section G. Due to Lemma 3 we can focus on minimizing (17). Furthermore, note from Lemma 3 we have that the minimizer of (17) does not depend on t, despite the dependence of the objective $h_t(w, \alpha)$ on t.

Since (17) is an average of $(n+1)$ terms we can apply an online SGD method. To simplify notation, for $i = 1, \ldots, n$ let

$$h_{i,t}(w, \alpha) \;:=\; \frac{1}{2}\frac{(f_i(w) - \alpha_i)^2}{\|\nabla f_i(w^t)\|^2 + 1} \quad \text{and} \quad h_{n+1,t}(w, \alpha) \;:=\; \frac{n}{2}(\overline{\alpha} - \tau)^2. \tag{19}$$

Note that despite our notation $h_{n+1,t}(w,\alpha)$ does not in fact depend on t or w. We do this for notational consistency.

Let $\gamma > 0$ be the learning rate. Starting from any w^0 and with $\alpha_i^0 = 0$ for all i, at each iteration t we sample an index $i_t \in \{1,\dots,n+1\}$. If $i_t \neq n+1$ then we sample $h_{i_t,t}$ and update

$$w^{t+1} = w^t - \gamma \nabla_w h_{i_t,t}(w^t,\alpha^t) \overset{(19)}{=} w^t - \gamma \frac{f_{i_t}(w^t) - \alpha_i^t}{\|\nabla f_i(w^t)\|^2 + 1} \nabla f_{i_t}(w^t) \tag{20}$$

$$\alpha_i^{t+1} = \alpha_i^t - \gamma \nabla_{\alpha_i} h_{i_t,t}(w^t,\alpha^t) \overset{(19)}{=} \alpha_i^t + \gamma \frac{f_{i_t}(w^t) - \alpha_i^t}{\|\nabla f_{i_t}(w^t)\|^2 + 1}. \tag{21}$$

Thus we have that (20) and (21) are equal to lines 9 and 8 in Algorithm 1, respectively. Alternatively if $i_t = n+1$ then we sample $h_{n+1,t}$ and our SGD step is given by

$$\alpha^{t+1} = \alpha^t - \gamma \nabla_\alpha h_{n+1,t}(w^t,\alpha^t) \overset{(19)}{=} \alpha^t - \gamma(\overline{\alpha}^t - \tau) \tag{22}$$

which is equal to line 6 in Algorithm 1.

We rely on this SGD interpretation of the TAPS method to provide a convergence analysis in Section E (specialized to TAPS in Section I). Key to this forthcoming analysis, is the following property.

Lemma 4 (Growth) *The functions $h_{i,t}(w)$ defined in (19) satisfy*

$$\left\|\nabla h_{i,t}(w^t,\alpha)\right\|^2 = 2h_{i,t}(w^t,\alpha), \quad \text{for } i = 1,\dots,n+1. \tag{23}$$

Consequently the function $h_t(w,\alpha)$ in (17) satisfies

$$\frac{1}{n+1}\sum_{i=1}^{n} \left\|\nabla h_{i,t}(w^t,\alpha)\right\|^2 = 2h_t(w^t,\alpha). \tag{24}$$

In the next section we completely remove Assumption 1 to develop a stochastic method that records only function values and needs no prior information on $f_i(w^*)$ or $f(w^*)$.

4. Moving Targeted Stochastic Polyak Steps

Here we dispense of Assumption 1 and instead introduce τ as a variable. Our objective is to design a *moving target* variant of the TAPS method that updates the target τ in a such a way that guarantees convergence. To design this moving target variant, we rely on the SGD online viewpoint. Consider the auxiliary objective function

$$\min_{w\in\mathbb{R}^d,\alpha\in\mathbb{R}^n,\tau\in\mathbb{R}} h_t(w,\alpha,\tau) := \frac{1}{n+1}\left(\sum_{i=1}^{n} \frac{1-\lambda}{2}\frac{(f_i(w)-\alpha_i)^2}{\|\nabla f_i(w^t)\|^2+1} + \frac{1-\lambda}{2}n(\overline{\alpha}-\tau)^2 + \frac{\lambda}{2}\tau^2\right), \tag{25}$$

where $\lambda > 0$ is a dampening parameter. Note that for $\lambda = 0$ we recover the same auxiliary function of the TAPS method in (17).

Lemma 5 *Every stationary point of (25) is a stationary point of (1). Finally if $f(w) \geq 0$ and $(w^*,\hat{\alpha},\hat{\tau})$ is a minima of (25) then w^* is a minima of (1).*

Since minimizing (25) is equivalent to minimizing (1), we can focus on solving (25). Following the same pattern from the previous sections, we will minimize the sum of $(n+1)$ terms in (25) using SGD. We refer to the resulting algorithms as MOTAPS, and provide it's detailed derivation in Section D. MOTAPS is similar to TAPS but with an additional of target variable τ. The complete pseudocode of MOTAPS can be found in the appendix in Algorithm 2.

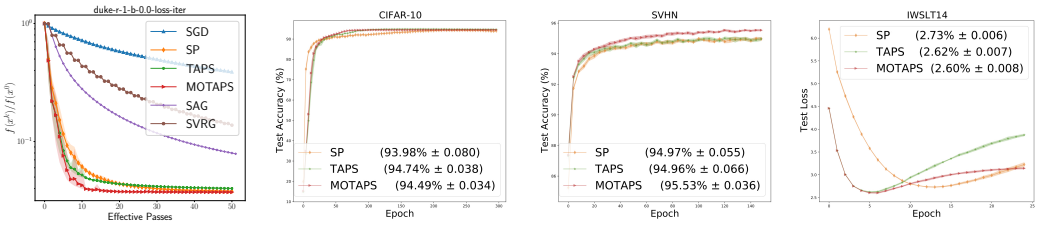


Figure 1: Comparison of SP, TAPS and MOTAPS on, from left to right, duke logistic regression, CIFAR [22], SVHN [27] and IWSLT14 [5]. The benchmark method and test error for

5. Conclusions and Contributions

We conclude by highlighting some of our contributions.

New perspectives and analysis of Stochastic Polyak. We provide new viewpoints of the SP method: 1) as a subsampled Newton method, 2) as a type of online SGD method.

Moving Targetted Stochastic Polyak. By leveraging the subsampled Newton viewpoint, we develop a new variant of the SP method that does not rely on the interpolation assumption. Instead, our new *TArgetted stochastic Polyak Stepsize* (TAPS) method assumes that $f(w^*)$ is known. TAPS uses n auxiliary scalar variables that track the evolution of the individual function values $f_i(w)$. Using the SGD viewpoint of SP, we then propose the *Moving* Targetted Stochastic Polyak (MOTAPS), that does not even require knowledge of $f(w^*)$. MOTAPS has the same n auxiliary scalars as TAPS plus one additional variable that tracks the global loss $f(w)$.

Unifying Convergence Theory. We prove that all three of our methods SP, TAPS and MOTAPS can be interpreted as variants of online SGD, and we use this to establish a unifying convergence theorem for all three of these methods. Furthermore, we show how theses variants of online SGD enjoy a remarkable growth property that greatly facilitates a proof of convergence. Indeed, we present a single convergence theorem (Theorem 10) that holds for these three methods by using this online SGD viewpoint and a star-convexity assumption [17, 23]. Star convex functions are a class of non-convex functions that include the loss function of some neural networks along the path of SGD [21, 38], several non-convex loss functions for generalized linear models [23], and learning linear dynamical systems [15].[4]

Competitive experimental results. We also show on several convex and non-convex learning problems how MOTAPS is competitive with the relevant benchmarks. See Figure 1 for a sub-selection of these experiments, and see Section F for the details on these experiments. Both TAPS and MOTAPS show favorable results compared to SP on all three problems. On the computer vision datasets, neither method reaches the generalization performance of SGD with a highly tuned step-wise learning rate schedule (95.2% for CIFAR10, 95.9% on SVHN). On the IWSLT14 problem, both TAPS and MOTAPS out-perform Adam [19] which achieved a 2.69 test loss and is the gold-standard for this task.

4. To be precise, the proof in [15] relies on a quasi-convex assumption, which is a slight relaxation over star-convex functions be introducing a relaxing parameter.

Acknowledgments

We would like to thank Konstantin Mischenko for useful discussions and the observationthat the SP method is essentially invariant to power transformations of the loss functions.

References

[1] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745–6750, 1999. ISSN 0027-8424. doi: 10.1073/pnas.96.12.6745.

[2] Hilal Asi and John C. Duchi. Stochastic (approximate) proximal point methods: Convergence, optimality, and adaptivity. *SIAM J. Optim.*, 29(3):2257–2290, 2019.

[3] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning practice and the bias-variance trade-off, 2019.

[4] Leonard Berrada, Andrew Zisserman, and M. Pawan Kumar. Training neural networks for and by interpolation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 799–809, 13–18 Jul 2020.

[5] Mauro Cettolo, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. Report on the 11th IWSLT evaluation campaign, IWSLT 2014. 2014.

[6] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[7] Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. Online passive-aggressive algorithms. *J. Mach. Learn. Res.*, 7:551–585, 2006.

[8] Aaron Defazio, Francis Bach, and Simon Lacoste-julien. Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems 27*, pages 1646–1654. 2014.

[9] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

[10] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *J. Mach. Learn. Res.*, 12:2121–2159, July 2011.

[11] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and E S Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, October 1999.

[12] Robert M. Gower. *Sketch and Project: Randomized Iterative Methods for Linear Systems and Inverting Matrices*. PhD thesis, University of Edinburgh, 2016.

[13] Robert M. Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *Proceedings of The 24th International Conference on Artificial Intelligence and Statistics*, volume 130, pages 1315–1323. PMLR, 2021.

[14] Robert Mansel Gower, Nicolas Loizou, Xun Qian, Alibek Sailanbayev, Egor Shulgin, and Peter Richtárik. Sgd: General analysis and improved rates. In *International Conference on Machine Learning*, pages 5200–5209, 2019.

[15] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Identity mappings in deep residual networks. In *Computer Vision – ECCV 2016*, 2016.

[17] Oliver Hinder, Aaron Sidford, and Nimit Sharad Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. *arXiv preprint arXiv:1906.11985*, 2019.

[18] Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323. Curran Associates, Inc., 2013.

[19] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[20] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015*, 2015.

[21] Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International Conference on Machine Learning*, pages 2698–2707, 2018.

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

[23] Jasper C. H. Lee and Paul Valiant. Optimizing star-convex functions. In Irit Dinur, editor, *IEEE 57th Annual Symposium on Foundations of Computer Science, FOCS*, pages 603–614, 2016.

[24] M. Pawan Kumar Leonard Berrada, Andrew Zisserman. Comment on Stochastic Polyak Step-Size: Performance of ALI-G. *arXiv:2105.10011*, May:1–2, 2021.

[25] Nicolas Loizou, Sharan Vaswani, Issam Laradji, and Simon Lacoste-Julien. Stochastic polyak step-size for sgd: An adaptive learning rate for fast convergence. *arXiv:2002.10542*, 2020.

[26] Eric Moulines and Francis R Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

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

[28] Adam M. Oberman and Mariana Prazeres. Stochastic gradient descent with Polyak's learning rate. *arXiv preprint arXiv:1903.08688*, 2019.

[29] B.T. Polyak. *Introduction to Optimization*. Optimization Software, New York, 1987.

[30] Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1):83–112, Mar 2017.

[31] Othmane Sebbouh, Robert M. Gower, and Aaron Defazio. Almost sure convergence rates for stochastic gradient descent and stochastic heavy ball. *COLT*, 2021.

[32] Sharan Vaswani, Francis Bach, and Mark Schmidt. Fast and faster convergence of SGD for over-parameterized models and an accelerated perceptron. *arXiv preprint arXiv:1810.07288*, 2018.

[33] Sharan Vaswani, Aaron Mishkin, Issam Laradji, Mark Schmidt, Gauthier Gidel, and Simon Lacoste-Julien. Painless stochastic gradient: Interpolation, line-search, and convergence rates. 32, 2019.

[34] Mike West, Carrie Blanchette, Holly Dressman, Erich Huang, Seiichi Ishida, Rainer Spang, Harry Zuzan, John A. Olson, Jeffrey R. Marks, and Joseph R. Nevins. Predicting the clinical status of human breast cancer by using gene expression profiles. *Proceedings of the National Academy of Sciences*, 98 (20):11462–11467, 2001.

[35] Rui Yuan, Alessandro Lazaric, and Robert M. Gower. Sketched newton-raphson. *arXiv:2006.12120, ICML workshop "Beyond first order methods in ML systems"*, 2020.

[36] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.

[37] Pengchuan Zhang, Hunger Lang, Qiang Liu, and Lin Xiao. Statistical adaptive gradient methods. *arXiv preprint arXiv:2002.10597*, 2020.

[38] Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. SGD converges to global minimum in deep learning via star-convex path. In *International Conference on Learning Representations*, 2019.

Appendix

The Appendix is organized as follows: In Section A we provide the related background to our methods. In Section C, we give some additional lemmas used to establish the closed form update of the methods. In Section D we present the detailed derivation of `MOTAPS` which was omitted from the main paper due. In Section G we present the missing proofs of the lemmas and theorems. We then present our unifying convergence theorem for `SP`, `TAPS` and `MOTAPS` in Section E. In Sections H, I and J we then discuss the consequences of our unifying convergence theorem to the `SP`, `TAPS` and `MOTAPS` method, respectively. We present detailed numerical experiments in Section F. Finally, in Section K and L we give further details on our implementations of the methods and the numerical experiments.

Contents

Appendix A. Background

Developing methods that adapt the stepsize using information collected during the iterative process is now a very active area of research. Adaptive methods such as AdaGrad [10] and Adam [20] have a step size that adapts to the scaling of the gradient, and thus are generally easier to tune than SGD, and have now become staples in training DNNs (deep neural networks). While the practical success of AdaGrad and Adam are undeniable, there lacks a fundamental understanding of why these methods work so well, particularly on models that interpolate data such as DNNs.

Recently a new family of adaptive methods based on the Polyak step size [29] has emerged, including the *stochastic Polyak step size* (SPS) method [25, 28] and ALI-G [4]. The Stochastic Polyak method (`SP`) is also an adaptive method, since it adjusts its step size depending on both the current loss value and magnitude of the stochastic gradient. Under the interpolation condition, the SPS method converges sublinearly under convexity [25] and star-convexity [13], and linearly under strong convexity and the PL condition [13, 25]. Recently in [25] the authors proposed the $\text{SPS}_{\max}$ method, which is a variant of `SP` that caps large stepsizes which greatly helps to stablize the convergence of `SP`. Prior to this, the ALI-G method [4] can be interpreted as dampened version of $\text{SPS}_{\max}$ method with follow-up work highlighting the importance of momentum in accelerating these methods in practice [24].

Our derivation of `SP` as a projection in (3) shows that `SP` can be interpreted as a extension of the passive-aggressive methods to nonlinear models [7]. Indeed, the passive aggressive methods apply the same projection in (3) but with the constraint $f_i(w) = 0$. This projection has a closed form solution when f_i is a hinge loss over a linear model, which was the setting where passive-aggressive models were first developed and most applied.

Another related set of methods are the model based methods in [2], where each new iteration is the result of minimizing the sum of a model of $f_i(w)$ and the norm squared distance to a prior point, that is

$$w^{t+1} = \text{argmin}_{w\in\mathbb{R}^d} \left\{ f_{t,i}(w) + \frac{1}{2}\left\|w - w^t\right\|^2 \right\}, \tag{26}$$

where $f_{t,i}(w)$ is some local model of $f_i(w)$ such that $f_{t,i}(w^t) = f_i(w^t)$. This model includes linearizations of $f_i(w)$ as a special case. The $SPS_{\max}$ method [25] is in fact a special case of the model based methods (26), where-in the model is given by the positive part of a local linearization, that is

$$f_{t,i}(w) = \max\{f_i(w^t) + \left\langle \nabla f_i(w^t), w - w^t \right\rangle, 0\}. \tag{27}$$

Using the positive part is justified for non-negative loss functions. This connection was first noted in [4].

Another promising direction for adaptive methods is to use a line search that works with stochastic gradient type methods [33, 37].

Appendix B. Comparing SP to SPS given in [25, 28]

The SP method given in (6) is closely related to the SPS method [25, 28] given by

$$w^{t+1} = w^t - \frac{f_i(w^t) - f_i^*}{\|\nabla f_i(w^t)\|^2}\nabla f_i(w^t), \tag{28}$$

where $f_i^* := \inf_w f_i(w)$ for $i = 1, \dots, n$. Note that the only difference between (28) and the SP method (6) is that $f_i(w^*)$ has been replaced by f_i^*. If the Interpolation Assumption 3 holds then $f_i^* = f_i(w^*)$ and the two methods are equal. Outside of the interpolation regime, these two methods are not necessarily the same, and this is where their difference lies.

In terms of convergence theory, the difference is only cosmetic, since the SPS (28) method only converges when $f_i^* = f_i(w^*)$, that is, when the two methods are equal. Indeed, let

$$\sigma := \frac{1}{n}\sum_{i=1}^{n}(f_i(w^*) - f_i^*).$$

Note that $\sigma \geq 0$ by the definition of f_i^*. According to Theorems 3.1 and 3.4 in [25] the SPS method (28) converges to a neighborhood of the solution with a diameter that depends on σ. Thus SPS converges to the solution when $\sigma = 0$. This only happens when the interpolation Assumption 3 holds. Putting convergence aside, the SPS method (28) has the advantage that for many machine learning f_i^* is known. This is in contrast to the SP method (6), where $f_i(w^*)$ is not known for most applications.

Appendix C. Auxiliary Lemmas

Lemma 6 *The solution to*

$$\begin{aligned} w^{t+1}, \alpha_i^{t+1} &= \underset{w\in\mathbb{R}^d, \alpha_i\in\mathbb{R}}{\operatorname{argmin}} \left\|w - w^t\right\|^2 + \left\|\alpha_i - \alpha_i^t\right\|^2 \\ &\qquad \textit{subject to } f_i(w^t) + \left\langle \nabla f_i(w^t), w - w^t\right\rangle = \alpha_i \end{aligned} \tag{29}$$

is given by

$$\begin{aligned} \alpha_i^{t+1} &= \alpha_i^t + \frac{f_i(w^t) - \alpha_i^t}{\|\nabla f_i(w^t)\|^2 + 1} \\ w^{t+1} &= w^t - \frac{f_i(w^t) - \alpha_i^t}{\|\nabla f_i(w^t)\|^2 + 1}\nabla f_i(w^t). \end{aligned} \tag{30}$$

Proof Introducing the variable $x = [w,\ \alpha_i] \in \mathbb{R}^{d+1}$ we can re-write (31) as

$$\begin{aligned} x^{t+1} &= \operatorname{argmin}_x \left\|x - x^t\right\|^2 \\ &\qquad \text{subject to } \begin{bmatrix}\nabla f_i(w^t) \\ -1\end{bmatrix}^\top x = -f_i(w^t) + \left\langle \nabla f_i(w^t), w^t\right\rangle. \end{aligned} \tag{31}$$

Using Lemma 7 (just below) we have that the solution to the above is given by

$$x^{t+1} \;=\; x^t + \begin{bmatrix}\nabla f_i(w^t)\\ -1\end{bmatrix}\frac{1}{\|\nabla f_i(w^t)\|^2+1}(-f_i(w^t)+\langle\nabla f_i(w^t),w^t\rangle-(\langle\nabla f_i(w^t),w^t\rangle-\alpha_i^t))$$

Substituting out $x=[w,\,\alpha_i]$ and simplifying we have

$$\begin{bmatrix}w^{t+1}\\ \alpha_i^{t+1}\end{bmatrix} \;=\; \begin{bmatrix}w^t\\ \alpha_i^t\end{bmatrix}+\begin{bmatrix}\nabla f_i(w^t)\\ -1\end{bmatrix}\frac{\alpha_i^t-f_i(w^t)}{\|\nabla f_i(w^t)\|^2+1},$$

which is equal to (30). ■

Lemma 7 *The solution to*

$$\begin{aligned} x^+ \;&=\; \operatorname{argmin}_{x\in\mathbb{R}^d}\|x-x^0\|^2 \\ &\quad\; \textit{subject to } a^\top x=b \end{aligned} \tag{32}$$

is given by

$$x^+ \;=\; x^0+\frac{a}{\|a\|^2}(b-a^\top x^0) \tag{33}$$

Proof Substitute $z=x-x^0$ and consider the resulting problem

$$\begin{aligned} z^+ \;&=\; \operatorname{argmin}_{z\in\mathbb{R}^d}\|z\|^2 \\ &\quad\; \text{subject to } a^\top z=b-a^\top x^0 \end{aligned} \tag{34}$$

One of the properties of the pseudo-inverse is that the least norm solution to the linear equation in (34) is given by

$$z^+=a^{+\top}(b-a^\top x^0), \tag{35}$$

where $a^{+\top}$ is the pseudo-inverse of $a^\top$. It is now easy to show that $a^{+\top}=\frac{a}{\|a\|^2}$ is the pseudo-inverse [5] of a. Substituting back x and the definition of $a^{+\top}$ in (35) gives (33). ■

C.1. Linear Algebra

Lemma 8 *For any matrices A, B, and C of appropriate dimensions we have that*

$$\left\|\begin{bmatrix}A & C\\ C^\top & B\end{bmatrix}\right\| \;\le\; \|A\|+2\|C\|+\|D\| \tag{36}$$

Proof Let $[vw]$ we a vector of unit norm. It follows that

$$\begin{aligned}
\left\|\begin{bmatrix}A & C\\ C^\top & B\end{bmatrix}\begin{bmatrix}v\\ w\end{bmatrix}\right\| \;&=\; \sqrt{\|Av+Cw\|^2+\|C^\top v+Bw\|^2} \\
&\le\; \|Av+Cw\|+\left\|C^\top v+Bw\right\| \\
&\le\; \|Av\|+\|Cw\|+\left\|C^\top v\right\|+\|Bw\| \\
&\le\; \|A\|\,\|v\|+\|C\|\,\|w\|+\|C\|\,\|v\|+\|B\|\,\|w\| \\
&\le\; \|A\|+2\|C\|+\|D\|,
\end{aligned}$$

5. This follows by the definition of pseudo-inverse since $a^{+\top}a^\top a^{+\top}=a^{+\top}$, $a^\top a^{+\top}a^\top=a^\top$ and both $a^\top a^{\top+}$ and $a^{\top+}a^\top$ are symmetric.

where in the first inequality we used that, for any $a, b > 0$ we have that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, and in the last inequality we used that $\|w\|, \|v\| \leq \|[w\ v]\| = 1$. ■

Appendix D. MOTAPS detailed derivation

Here we provide the detailed derivation of the MOTAPS algorithm. In applying SGD, we partition the function (25) into $n+1$ terms, where the first n terms are given by

$$h_{i,t}(w, \alpha, \tau) = \frac{1-\lambda}{2} \frac{(f_i(w) - \alpha_i)^2}{\|\nabla f_i(w^t)\|^2 + 1} \quad \text{for } i = 1, \ldots, n.$$

The $(n+1)$th term is given by

$$\begin{aligned} h_{n+1,t}(w, \alpha, \tau) &:= \frac{(1-\lambda)n}{2}(\overline{\alpha} - \tau)^2 + \frac{\lambda}{2}\tau^2 \quad \text{thus} \\ \nabla_\tau h_{n+1,t}(w, \alpha, \tau) &= (1-\lambda)n(\tau - \overline{\alpha}) + \lambda\tau. \end{aligned} \tag{37}$$

Sampling the $(n+1)$th term and taking a gradient step to update $\tau^t \in \mathbb{R}$ gives the following update

$$\begin{aligned} \tau^{t+1} &= \tau^t - \gamma \nabla_\tau h_{n+1,t}(w, \alpha, \tau)|_{(w,\alpha,\tau)=(w^t,\alpha^t,\tau^t)} \\ &= \big(1 - \underbrace{\gamma\left(\lambda + (1-\lambda)n\right)}_{:=\gamma_\tau}\big)\tau^t + \gamma(1-\lambda)n\overline{\alpha}^t \qquad (38) \\ &= (1-\gamma_\tau)\tau^t + \gamma_\tau \frac{(1-\lambda)n}{\lambda + (1-\lambda)n}\overline{\alpha}^t \qquad (39) \end{aligned}$$

where we have introduced a separate learning rate γ_τ for updating τ. We find that a separate learning rate γ_τ is needed for updating τ, otherwise to keep τ from being negative in (38) we would need to restrict γ to be less than $\frac{1}{\lambda+(1-\lambda)n}$ which can be small when λ is close to zero. See Algorithm 2 for the resulting method. We refer to this method as the *Moving Target Stochastic Polyak Stepsize* or MOTAPS for short.

The dampening parameter λ controls how fast the stochastic gradients of $h_t(w, \alpha, \tau)$ can grow, as we show next. As a consequence, later on we will see that the λ will later control the rate of convergence of MOTAPS.

Lemma 9 *Consider $h_t(w, \alpha, \tau)$ given in (25). If*

$$\lambda \leq \frac{2n+1}{2n+3} < 1 \tag{40}$$

then

$$\frac{1}{n+1}\sum_{i=1}^{n+1} \left\|\nabla h_{i,t}(w^t, \alpha, \tau)\right\|^2 \leq 2(1-\lambda)(2n+1)h_t(w^t, \alpha, \tau). \tag{41}$$

Next we establish a general convergence theory through which we will analyse SP, TAPS and MOTAPS.

Algorithm 2 MOTAPS: Moving TArgetted Stochastic Polyak Step

Inputs: Dampening $\lambda \in [0,\ 1]$ and learning rates $\gamma, \gamma_\tau \in [0,\ 1]$
Default: $\gamma = 0.9$, $\gamma_\tau = \lambda = 0.1$, $w^0 = 0, \alpha_i^0 = \overline{\alpha}^0 = 0 = \tau$ for $i = 1, \ldots, n$
for $t = 0, \ldots, T-1$ **do**
  Sample $i \in \{1, \ldots, n+1\}$ according to some law
  if $i = n+1$ **then**
    $\alpha_j^{t+1} = \alpha_j^t + \gamma(\tau - \overline{\alpha}^t)$, for $j = 1, \ldots, n$. ▷ Updating all α's
    $\tau = (1-\gamma_\tau)\tau + \gamma_\tau \frac{(1-\lambda)n}{\lambda + (1-\lambda)n}\overline{\alpha}$ ▷ Updating target τ
    $\overline{\alpha}^{t+1} = \overline{\alpha}^t + \gamma(\tau - \overline{\alpha}^t)$
  else
    $\alpha_i^{t+1} = \alpha_i^t + \gamma \dfrac{f_i(w^t) - \alpha_i^t}{\|\nabla f_i(w^t)\|^2 + 1}$ ▷ Updating α_i
    $w^{t+1} = w^t - \gamma \dfrac{f_i(w^t) - \alpha_i^t}{\|\nabla f_i(w^t)\|^2 + 1} \nabla f_i(w^t)$. ▷ Updating w
    $\overline{\alpha}^{t+1} = \overline{\alpha}^t + \dfrac{1}{n}(\alpha_i^{t+1} - \alpha_i^t)$
Output: w^T

Appendix E. Convergence Theory

All of our methods presented thus far can be cast as a particular variant of online SGD. Indeed, SP, TAPS and MOTAPS given in (6), Algorithms 1 and 2, respectively, are equivalent to applying SGD to (7), (17) and (25), respectively. We will leverage this connection to provide a convergence theorem for these three methods. Throughout our proofs we use

$$\min_z h_t(z) := \frac{1}{n}\sum_{i=1}^n h_{i,t}(z) \tag{42}$$

as the auxiliary function in consideration. Here z represents the variables of the problem. For instance, for the SP method (6) we have that $z = w \in \mathbb{R}^d$, for TAPS in Algorithm 1 we have that $z = (w, \alpha) \in \mathbb{R}^{n+d}$ and finally for MOTAPS in Algorithm 2 we have that $z = (w, \alpha, \tau) \in \mathbb{R}^{n+d+1}$.

Consider the online SGD method applied to minimizing (42) given by

$$z^{t+1} = z^t - \gamma \nabla h_{i_t,t}(z^t), \tag{43}$$

where $i_t \in \{1, \ldots, n\}$ is sampled uniformly and i.i.d at every iteration and $\gamma > 0$ is a step size. For each method we provide a *growth condition* (see Lemmas 1, 4, 9) that we now state as an assumption.

Assumption 2 *There exists $G \geq 0$ such that*

$$\mathbb{E}\left[\left\|\nabla h_{i_t,t}(z^t)\right\|^2\right] \quad \leq \quad 2G\, h_t(z^t). \tag{44}$$

E.1. General Convergence Theory

Here we present two general convergence theorems that will then be applied to our three algorithms. The first theorem relies on a weak form of convexity known as star convexity.

Theorem 10 (Sublinear) *Suppose Assumption 2 holds with $G > 0$. Let $\gamma < 1/G$ and suppose there exists z^* such that h_t is star-convex at z^t and around z^*, that is*

$$h_t(z^*) \;\geq\; h_t(z^t) + \left\langle \nabla h_t(z^t), z^* - z^t \right\rangle, \tag{45}$$

then we have that

$$\min_{t=1,\ldots,k} \mathbb{E}\left[h_t(z^t) - h_t(z^*)\right] \leq \frac{1}{k}\frac{1}{2\gamma(1-G\gamma)}\mathbb{E}\left[\left\|z^0 - z^*\right\|^2\right] + \frac{G\gamma}{1-G\gamma}\frac{1}{k}\sum_{t=1}^{k} h_t(z^*). \tag{46}$$

Our second theorem relies on a weakened form of strong convexity.

Theorem 11 (Linear Convergence) *Suppose Assumption 2 holds with $G > 0$. Let $\gamma \leq 1/G$. If there exists $\mu > 0$ and z^* such that h_t is μ–strongly star–convex along z^t and around z^*, that is*

$$h_t(z^*) \;\geq\; h_t(z^t) + \left\langle \nabla h_t(z^t), z^* - z^t \right\rangle + \frac{\mu}{2}\left\|z^* - z^t\right\|, \quad \textit{then} \tag{47}$$

$$\mathbb{E}\left[\left\|z^{t+1} - z^*\right\|^2\right] \;\leq\; (1-\gamma\mu)^{t+1}\left\|z^0 - z^*\right\|^2 + 2G\gamma^2\sum_{i=0}^{t}(1-\gamma\mu)^i\mathbb{E}\left[h_i(z^*)\right]. \tag{48}$$

In the next three sections we specialize these theorems, and their assumptions, to the SP, TAPS and MOTAPS methods, respectively. In particular, in Section E.2 we show how two previously known convergence results for SP are special cases of Theorem 10 and 11. In Section E.3 we show that the auxiliary functions of TAPS and MOTAPS in (17) and (25) are locally strictly convex under a small technical assumption. In Section E.4 we finally prove convergence of MOTAPS.

E.2. Convergence of SPS

Before establishing the convergence of SP, we start by stating a slightly more general interpolation assumption as follows.

Assumption 3 (Interpolation) *We say that the interpolation assumption holds when*

$$\exists w^* \in \mathcal{W}^* \quad \textit{such that} \quad f_i(w^*) = \min_{w\in\mathbb{R}^d} f_i(w), \quad \textit{for } i = 1,\ldots,n. \tag{49}$$

Here we specialize Theorems 10 and 11 to the SP method (6). Both of these theorems rely on assuming that the proxy function h_t is star-convex or strongly star-convex. Thus first we establish sufficient conditions for this to hold.

Lemma 12 *Let the interpolation Assumption 3 hold. If every f_i is star convex along the iterates w^t given by (6), that is,*

$$f_i(w^*) \geq f_i(w) + \langle \nabla f_i(w), w^* - w\rangle \tag{50}$$

then $h_{i,t}(w)$ is star convex along the iterates w^t and around w^*, that is.

$$h_{i,t}(w^*) \geq h_{i,t}(w^t) + \left\langle \nabla_w h_{i,t}(w^t), w^* - w\right\rangle. \tag{51}$$

Furthermore if f_i is μ_i-strongly convex and L_i–smooth then $h_t(w)$ is $\frac{1}{2n}\sum_{i=1}\frac{\mu_i}{L_i}$–strongly star-convex, that is

$$h_t(w^*) \geq h_t(w^t) + \left\langle \nabla_w h_t(w^t), w^* - w\right\rangle + \frac{1}{4n}\sum_{i=1}^{n}\frac{\mu}{L_i}\left\|w^t - w^*\right\|^2. \tag{52}$$

Using Lemma 12, we can now prove convergence of SP as a corollaries of Theorem 10 and 11.

Corollary 13 (Convergence of SPS) *If $\gamma < 1$ and every $f_i(w)$ is star-convex along the iterates w^t given by (6) then*

$$\frac{1}{k}\sum_{t=0}^{k}\frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{f_i(w^t)-f_i(w^*)}{\|\nabla f_i(w^t)\|}\right)^2\right] \leq \frac{1}{k}\frac{1}{2\gamma(1-\gamma)}\mathbb{E}\left[\left\|w^0-w^*\right\|^2\right]. \tag{53}$$

Furthermore if the interpolation Assumption 3 holds and if each $f_i(w)$ is L_i–smooth then

$$\min_{t=0,\ldots,k}\mathbb{E}\left[f(w^t)-f^*\right] \leq \frac{1}{k}\frac{L_{\max}}{2\gamma(1-\gamma)}\mathbb{E}\left[\left\|w^0-w^*\right\|^2\right], \tag{54}$$

where $L_{\max} := \max_{i=1,\ldots,n} L_i$.

The resulting convergence in (54) has already appeared in Theorem 4.4 in [13]. There in [13] the authors use a carefully proof technique that relies on a new notion of smoothness (Lemma 4.3 in [13]). But here we have that 13 is rather a direct consequence of interpreting SP as a type of SGD method.

Corollary 14 *If $\gamma \leq 1$, the interpolation Assumption 3 holds, and every f_i is L_i–smooth and μ–strongly star-convex then the iterates w^t given by (6) converge linearly according to*

$$\mathbb{E}\left[\left\|w^{t+1}-w^*\right\|^2\right] \leq \left(1-\gamma\frac{1}{2n}\sum_{i=1}^{n}\frac{\mu_i}{L_i}\right)^{t+1}\left\|w^0-w^*\right\|^2. \tag{55}$$

This corollary shows that Theorem D.3 in [13] is a special case of Theorem 11, and again a direct result of interpreting SP as a type of SGD method. The rate of convergence in (55) is also tighter than the analysis given in Theorem 3.1 in [25] where the rate is $1-\frac{\gamma}{2}\frac{\frac{1}{n}\sum_{i=1}^{n}\mu_i}{L_{\max}}$.

E.3. Convergence of TAPS

Here we explore the consequences of Theorems 10 and 11 to the TAPS method. To this end, let $z := (w,\alpha)$ and let

$$h_t(z) = h_t(w,\alpha) := \frac{1}{n+1}\left(\sum_{i=1}^{n}\frac{1}{2}\frac{(f_i(w)-\alpha_i)^2}{\|\nabla f_i(w^t)\|^2+1}+\frac{n}{2}(\overline{\alpha}-\tau)^2\right). \tag{56}$$

As a first step, we need to determine sufficient conditions for this auxiliary function h_t of TAPS to be star-convex. Unfortunately this turned out to be very challenging. This is because the star-convexity h_t is not a consequence of f_i being star-convex, nor the converse. Instead, star-convexity of h_t translates to new nameless assumptions on the f_i functions. As an insight into this difficulty, supposing that each f_i is convex is not enough to guarantee that h_t is convex. As a simple counterexample, let $n=1$ and $f_1(w)=w^2$. Thus $\tau=0$ and from (56) we have

$$h_t(z) = \frac{1}{2}\left(\frac{1}{2}\frac{(w^2-\alpha_1)^2}{\|w^t\|^2+1}+\frac{1}{2}\alpha_1^2\right).$$

It is easy to show that for α_1 large enough, the hessian of h_t has a negative eigenvalue, and thus h_t is non-convex. Conversely, h_t can have local convexity even when the underlying loss function is arbitrarily non-convex, as we show in the next lemma and corollary.

Lemma 15 (Locally Convex) *Consider the iterates of Algorithm 2. Let* $(w, \alpha) \in \mathbb{R}^{d+n}$ *and consider* $h_t(w, \alpha)$ *defined in* (56) *. Assume that the gradients at* w *span the entire space, that is*

$$\text{span}\left\{\nabla f_1(w), \ldots, \nabla f_n(w)\right\} \;=\; \mathbb{R}^d, \quad \forall w. \tag{57}$$

If Assumption 1 holds, every $f_i(w)$ *for* $i = 1, \ldots, n$ *is twice continuously differentiable and*

$$\frac{1}{n+1}\sum_{i=1}^{n} \nabla^2 f_i(w^t) \frac{f_i(w^t) - \alpha_i^t}{\|\nabla f_i(w^t)\|^2 + 1} \succeq 0, \tag{58}$$

then h_t *is strictly convex at* (w^t, α^t) *with*

$$\nabla^2 h_t(w^t, \alpha^t) \;\succ\; 0.$$

The condition on the span of the gradients (133) typically holds in the setting where we have more data then dimensions (features). Fortunately this occurs in precisely the setting where TAPS makes most sense since it makes sense to apply TAPS when $f^* = \tau > 0$. This can only occur in the *underparametrized* setting, where we have more data than features.

The condition in Lemma 15 that is difficult to verify is (58). A sufficient condition for (58) to hold is

$$\nabla^2 f_i(w^t)(f_i(w^t) - \alpha_i^t) \;\succeq\; 0. \tag{59}$$

Since α_i^t are essentially tracking $f_i(w^t)$ (see line 8 in Algorithm 1), we can state (59) in words as: if α_i^t is underestimating $f_i(w^t)$ then f_i should be convex at w^t, and conversely if α_i^t is overestimating $f_i(w^t)$ then f_i should be concave at w^t.

There is one point where (58) holds trivially, and that is at every point such that $\alpha_i = f_i(w)$. This includes every minimizer (w^*, α^*) since by Lemma 3 we have that $\alpha_i^* = f_i(w^*)$. Consequently, as we state in the following corollary, under minor technical assumption, we have that $h_t(w, \alpha)$ has no *degenerate* local minimas. This shows that h_t has some local convexity.

Corollary 16 (Locally Strictly Convex TAPS**)** *Consider the iterates of the* TAPS *method given in Algorithm 1. Let* w^* *be a minimizer of* (1) *and let* $\alpha_i^* = f_i(w^*)$ *for* $i = 1, \ldots, n$*. Assume that the gradients at* w^* *span the entire space, that is*

$$\text{span}\left\{\nabla f_1(w^*), \ldots, \nabla f_n(w^*)\right\} \;=\; \mathbb{R}^d. \tag{60}$$

If Assumption 1 holds and if every $f_i(w)$ *for* $i = 1, \ldots, n$ *is twice continuously differentiable then* $\nabla^2 h_t(w, \alpha) \succ 0$ *and thus* $h_t(w, \alpha)$ *is* strictly convex *at* (w^*, α^*).

Next we specialize Theorem 10 to the TAPS method in the following corollary.

Corollary 17 (Sublinear Convergence of TAPS**)**

Let $h_t(z)$ *in* (17) *be star–convex* (107) *around* $z^* = (w^*, \alpha^*)$ *and along the iterates* $z^t = (w^t, \alpha^t)$ *of Algorithm 1. If* $\gamma < 1$ *and* $f_i(w)$ *is* $L_{\max}$*–Lipschitz then*

$$\min_{t=1,\ldots,k} \frac{1}{n+1}\left(\sum_{i=1}^{n} \frac{\mathbb{E}\left[f_i(w^t) - \alpha_i^t\right]^2}{L_{\max} + 1} + \mathbb{E}\left[\overline{\alpha}^t - \tau\right]^2\right) \;\leq\; \frac{1}{k}\frac{1}{\gamma(1-\gamma)}\mathbb{E}\left[\left\|w^0 - w^*\right\|^2\right]. \tag{61}$$

Alternatively, if $h_t(z)$ *is* μ*–strongly star–convex* (111) *then*

$$\mathbb{E}\left[\left\|w^t - w^*\right\|^2 + \sum_{i=1}^{n}\left\|\alpha_i^t - f_i(w^*)\right\|^2\right] \leq (1-\gamma\mu)^t\left(\left\|w^0 - w^*\right\|^2 + \sum_{i=1}^{n}\left\|\alpha_i^0 - f_i(w^0)\right\|^2\right). \tag{62}$$

E.4. Convergence of MOTAPS

Here we explore the consequences of Theorems 10 and 11 specialized to Algorithm 2. In this case, the proxy function $h_t(z) = h_t(w, \alpha, \tau)$ is given

$$h_t(z) \;:=\; \frac{1}{n+1}\left(\sum_{i=1}^{n} \frac{1-\lambda}{2}\frac{(f_i(w)-\alpha_i)^2}{\|\nabla f_i(w^t)\|^2+1} + \frac{n(1-\lambda)}{2}(\overline{\alpha}-\tau)^2 + \frac{\lambda}{2}\tau^2\right). \tag{63}$$

Before applying Theorems 10 and 11 we should verify when $h_t(z)$ is star-convex or convex. This turns out to be much the same task as verifying that the auxiliary function for TAPS given in (56) is star-convex. This is because the only difference between the two functions is that (140) has an additional $\frac{\lambda}{2}\tau^2$ which adds strong convexity in the new τ dimension. Thus the discussion, and results around Lemma 15 and Corollary 16 remain largely true for (63). That is, we are only really able to establish when h_t is locally convex.

For the remainder of this section we impose that the dampening parameter satisfies

$$\lambda \leq \frac{2n+1}{2n+3} < 1, \tag{64}$$

so that we can apply Lemma 9. In our forth coming corollaries we will prove convergence of MOTAPS to the point $z^* = (w^*, \alpha^*, \tau^*)$ where w^* is a minimizer of (1) and

$$\alpha_i^* := f_i(w^*) \quad \text{and} \quad \tau^* = f(w^*), \quad \text{for } i = 1, \dots, n. \tag{65}$$

First we develop a corollary based on Theorem 10.

Corollary 18 *Let $\lambda \in [0,\ 1]$ satisfy (64) and let $z^t := (w^t, \alpha^t, \tau_t)$ be the iterates of Algorithm 2 when using a stepsize $\gamma = \frac{1}{2(1-\lambda)(2n+1)}$ and $\gamma_\tau = \gamma\left(\lambda + (1-\lambda)n\right)$. If $h_t(z)$ is star convex along the iterates z^t and around $z^* := (w^*, \alpha^*, \tau^*)$ then*

$$\min_{t=0,\dots,k} \mathbb{E}\left[h_t(z^t) - h_t(z^*)\right] \;\leq\; \frac{2(1-\lambda)(2n+1)}{k}\left\|z^0 - z^*\right\|^2 + \frac{\lambda f(w^*)^2}{2(n+1)}. \tag{66}$$

Furthermore, if f_i is $L_{\max}$–Lipschitz then

$$\frac{1}{n+1}\mathbb{E}\left[\sum_{i=1}^{n}\frac{1}{2}\frac{(f_i(w^t)-\alpha_i^t)^2}{L_{\max}+1} + \frac{n}{2}(\overline{\alpha}^t - \tau^t)^2 + \frac{\lambda}{2}\left((\tau^t)^2 - f(w^*)^2\right)\right]$$
$$\leq \frac{2(1-\lambda)(2n+1)}{k}\left\|z^0 - z^*\right\|^2 + \frac{\lambda f(w^*)^2}{2(n+1)}. \tag{67}$$

This Corollary 18 shows that $(f_i(w^t), \overline{\alpha}^t, \tau^t)$ converges to $(\alpha_i^t, \tau^t, f(w^*))$ sublinearly up to an additive error $\frac{\lambda f(w^*)^2}{2(n+1)}$ which is controlled by λ: When λ is very small, this additive error is very small. But λ also controls the *speed* of convergence. Indeed for λ close to 1 the method converges faster upto this additive error. Thus λ controls a trade-off between speed of convergence and radius of convergence.

The next corollary is based on Theorem 11.

Corollary 19 *Let $\lambda \in [0,\ 1]$ satisfy (64) and let $z^t := (w^t, \alpha^t, \tau_t)$ be the iterates of Algorithm 2 when using a stepsize $\gamma = \frac{1}{2(1-\lambda)(2n+1)}$ and $\gamma_\tau = \gamma\left(\lambda + (1-\lambda)n\right)$. If $h_t(z)$ is μ–strongly star–convex along the iterates z^t and around $z^* := (w^*, \alpha^*, \tau^*)$ then*

$$\mathbb{E}\left[\left\|z^{t+1} - z^*\right\|^2\right] \;\leq\; \left(1 - \frac{\mu}{(1-\lambda)(2n+1)}\right)^{t+1}\left\|z^0 - z^*\right\|^2 + \frac{\lambda f(w^*)^2}{\mu(n+1)}. \tag{68}$$

In both Corollary 18 and 19 the λ parameter controls a trade-off between speed of convergence and an additive error term. For example, for the largest value $\lambda = \frac{2n+1}{2n+3}$ (due to (64)) we have that (68), after some simplifications, gives

$$\mathbb{E}\left[\left\|z^{t+1} - z^*\right\|^2\right] \leq \left(1 - \mu/2\right)^{t+1} \left\|z^0 - z^*\right\|^2 + \frac{f(w^*)^2}{\mu(n+1)}.$$

Thus the convergence rates is now $1 - \mu/2$ and independent of n. But the additive error terms $\frac{f(w^*)^2}{\mu(n+1)}$ is now larger. On the other end, as $\lambda \to 0$ the rate of convergence tends to $(1 - \frac{\mu}{2n+1})$, which now depends on n, and the additive error term tends to zero.

By controlling this trade-off, next we use Corollary 33 to establish a total complexity of Algorithm 2.

Theorem 20 *Consider the setting of Corollary 19. For a given $\epsilon > 0$ it follows that*

$$t \geq \frac{(1-\lambda)(2n+1)}{\mu} \log\left(\frac{2\left\|z^0 - z^*\right\|^2}{\epsilon}\right) \implies \mathbb{E}\left[\left\|z^{t+1} - z^*\right\|^2\right] < \frac{\epsilon}{2} + \frac{\lambda f(w^*)^2}{\mu(n+1)}. \tag{69}$$

Consequently if we could choose

$$\lambda < \min\left\{\frac{\mu(n+1)}{f(w^*)^2}\frac{\epsilon}{2}, \frac{2n+1}{2n+3}\right\} \tag{70}$$

then

$$t \geq \frac{2n+1}{\mu} \log\left(\frac{2\left\|z^0 - z^*\right\|^2}{\epsilon}\right) \implies \mathbb{E}\left[\left\|z^{t+1} - z^*\right\|^2\right] < \epsilon. \tag{71}$$

Proof By stand arguments using the properties of logarithm we have that

$$t \geq \frac{(1-\lambda)(2n+1)}{\mu} \log\left(\frac{2}{\epsilon}\right) \implies \left(1 - \frac{\mu}{(1-\lambda)(2n+1)}\right)^{t+1} < \frac{\epsilon}{2}.$$

See for instance Lemma 11 in [12]. Furthermore, by using (70) we have that

$$\begin{aligned} t \geq \left(1 - \frac{\mu(n+1)}{f(w^*)^2}\frac{\epsilon}{2}\right)\frac{2n+1}{\mu} \log\left(\frac{2\left\|z^0 - z^*\right\|^2}{\epsilon}\right) \geq \frac{2n+1}{\mu} \log\left(\frac{2\left\|z^0 - z^*\right\|^2}{\epsilon}\right) \\ \implies \mathbb{E}\left[\left\|z^{t+1} - z^*\right\|^2\right] < \epsilon. \end{aligned} \tag{72}$$

■

Thus by choosing λ small enough, we can show that the MOTAPS method converges linearly. This is in stark contrast to SGD where, despite the presence of an additive error when using a constant step size (See Theorem 1 in [26]), this additive term only vanishes by setting the stepsize to zero. In contrast for MOTAPS we can set λ arbitrarily small without halting the method.

In practice, we would not know how to set λ using (70) since we would not know $f(w^*)$. Furthermore, we may not have a particular ϵ in mind, and instead, prefer to monitor the error and stop when resources are exhausted. To address both of these concerns, the next theorem offers another way to deal with the additive error by eventually decreasing the step size.

Theorem 21 *Consider the setting of Corollary 19. For a given $\epsilon > 0$ if we use an iteration dependent stepsize in Algorithm 2 given by*

$$\gamma_t = \begin{cases} \dfrac{1}{(1-\lambda)(2n+1)} & \text{if } t \leq 2(2n+1)\left\lceil \frac{1-\lambda}{\mu} \right\rceil \\ \dfrac{(t+1)^2 - t^2}{\mu(t+1)^2} & \text{if } t \geq 2(2n+1)\left\lceil \frac{1-\lambda}{\mu} \right\rceil \end{cases} \tag{73}$$

and if

$$\lambda \leq \min\left\{1 - \frac{2\mu}{2n+1}, \frac{2n+1}{2n+3}\right\}.$$

then

$$\mathbb{E}\left[\left\|z^t - z^*\right\|^2\right] \leq \frac{(1-\lambda)\lambda f(w^*)^2}{\mu^2}\frac{16}{t} + \frac{4(2n+1)^2}{e^2 t^2}\left\lceil\frac{1-\lambda}{\mu}\right\rceil^2 \left\|z^0 - z^*\right\|^2. \tag{74}$$

This Theorem 34 relies on knowing μ to set a *switching point* and the step size in (147). In practice it can also be difficult to estimate μ, but this theorem is still useful in that, it suggests that at some point in the execution we should decrease the stepsize

$$\gamma_t = \mathcal{O}\left(\frac{(t+1)^2 - t^2}{(t+1)^2}\right) = \mathcal{O}\left(\frac{2t+1}{(t+1)^2}\right) = \mathcal{O}\left(\frac{1}{t+1}\right),$$

much in the same way that SGD is used in practice.

Appendix F. Experiments

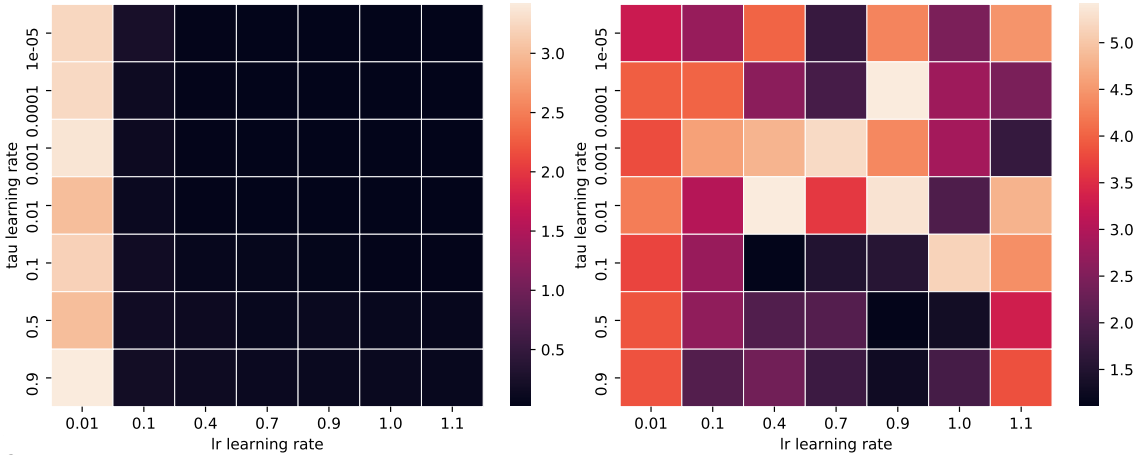
F.1. Convex Classification

We first experiment with a classification task using logistic regression. Details of these experiments and the data sets used are in Section K. For the sake of simplicity, here we test the `MOTAPS` method in Algorithm 2 with $\lambda = 0.5$. To determine a reasonable parameter setting for the `MOTAPS` methods we performed a grid search over the two parameters γ and γ_τ. See Figure 4 for the results of the grid search for an over-parametrized problem `colon-cancer` and an under-parametrized problem `mushrooms`. Through these grid searches we found that the determining factor for setting the best stepsize was the magnitude of the regularization parameter $\sigma > 0$. If σ was small or zero then $\gamma = 1$ and $\gamma_\tau = 0.001$ resulted in a good performance. On the other hand, if σ is large then $\gamma = 0.01$ and $\gamma_\tau = 0.9$ resulted in the best performance. This is most likely due to the effect that σ has on the optimal value $f(w^*)$.

F.2. Comparison to Variance Reduced Methods

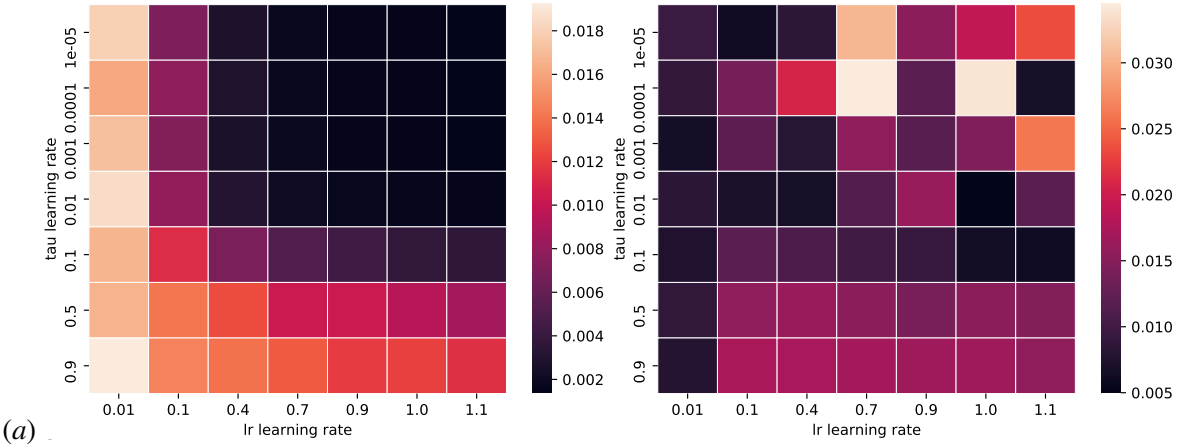
We compare our methods against SGD, and two variance reduced gradient methods SAG [8, 30] and SVRG [18] which are arguably among the state-of-the-art methods for minimizing logistic regression. For setting the parameters for SGD, based on [14] we used the learning rate schedule $\gamma_t = L_{\max}/t$ where $L_{\max}$ is the smoothness constant. For SVRG and SAG we used $\gamma = 1/2L_{\max}$. For `SP` and `TAPS` we used $\gamma = 1$ and approximated $f_i(w^*) = 0$. Because of this the `SP` is equivalent to the SPS method given in [25]. Following [25] experimental results, we also implemented `SP` with a max stepsize rule[6]. For `MOTAPS`, based

6. In [25] the authors also recommend the use of a further *smoothing* trick, but we opted for simplicity and chose not to use this smoothing.



(*a*)

Figure 2: `duke` $(n, d) = (7130, 44)$ Left: $\sigma = 0.0$. Right: $\sigma = \min_{i=1,\ldots,n} \|x_i\|^2 / n = 5.06$



(*a*)

Figure 3: `mushrooms` $(n, d) = (62, 2001)$ Left: $\sigma = 0.0$. Right: $\sigma = \min_{i=1,\ldots,n} \|x_i\|^2 / n = 2.66$

Figure 4: The resulting gradient norm of `MOTAPS` after running 50 epochs on a logistic regression

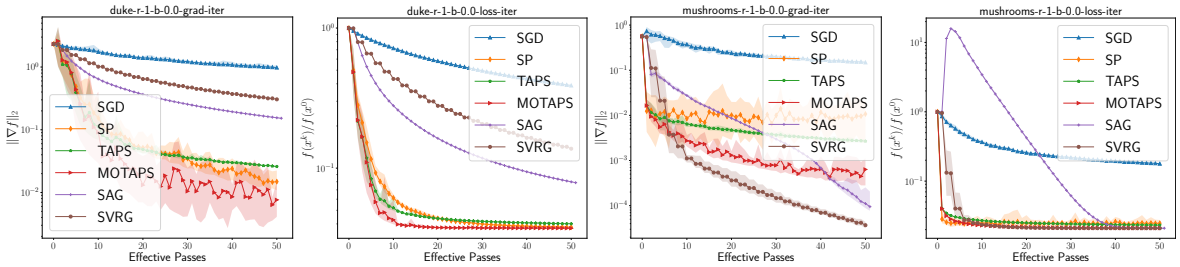


Figure 5: Logistic Regression with data set Left: duke $(n, d) = (44, 7130)$ and Right: mushrooms $(n, d) = (8124, 113)$ with regularization $\sigma = 1/n$.

on our observations in the grid search, we used the rule of thumb $\gamma = 1.0/(1 + 0.25\sigma e^{\sigma})$ and $\gamma_\tau = 1 - \gamma$. We compare all the algorithms in terms of epochs (effective passes over the data) in Figure 5. We found that in under-parametrized problem such as the mushrooms data set in Figure 5, and problems with a large regularization, SAG and SVRG were often the most efficient methods. For over-parametrized problems such as `duke`, with moderate regularization, the `MOTAPS` methods was the most efficient. Finally, for over-parametrized problems with very small regularization the `SP` method was the most efficient, see Section K.2
.

Furthermore `MOTAPS` has two additional advantages over SAG and SVRG 1) setting the stepsize does not require computing the smoothness constant and 2) does not require storing a $n \times d$ table of gradient (like SAG) or doing an occasional full pass over the date (like SVRG). We also found that adding momentum to `SP` and `MOTAPS` could speed up the methods. See Section K.3 for details on how we added momentum and additional experiments.

F.3. Deep learning tasks

We preformed a series of experiments on three benchmark problems commonly used for testing optimization methods for deep learning. CIFAR10 [22] is a computer vision classification problem and perhaps the most ubiquitous benchmark in the deep learning. We used a large and over-parameterized network for this task, the 152 layer version of the pre-activation ResNet architecture [16], which has over 58 million parameters.

For our second problem, we choose an under-parameterized computer vision task. The street-view house numbers dataset [27] is similar to the CIFAR10 dataset, consisting of the same number of classes, but with a much larger data volume of over 600k training images compared to 50k. To ensure the network can not completely interpolate the data, we used a much smaller ResNet network with 1 block per layer and 16 planes at the first layer, so that there are fewer parameters than data-points.

For our final comparison we choose one of the most popular NLP benchmarks, the IWSLT14 english-german translation task [5], consisting of approximately 170k sentence pairs. This task is relatively small scale and so overfitting is a concern on this problem. We applied a modern Transformer network with embedding size of 512, 8 heads and 3/3 encoding/decoding layers.

In each case the minimum loss is unknown so for the `TAPS` method we assume it is 0. Due to a combination of factors including the use of data-augmentation and L2 regularization, this is only an approximation. The learning rate for each method was swept on a power-of-2 grid on a single training seed, and the best value was used for the final comparison, shown over an average of 10 seeds. Error bars indicate 2 standard errors. L2 regularization was used for each task, and tuned for each problem and method separately also on a power-of-2 grid. We found that the optimal amount of regularization was not sensitive to the optimization

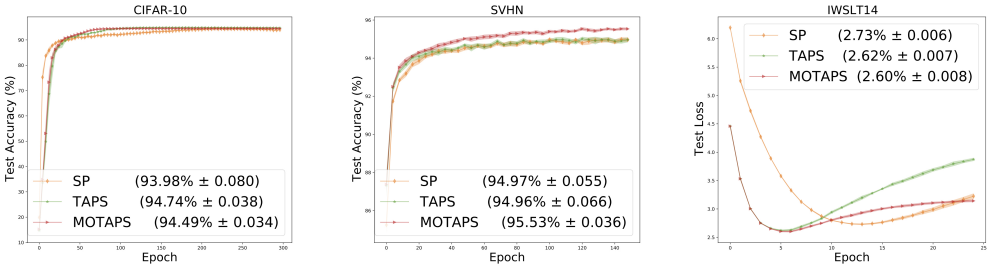


Figure 6: Deep learning experiments

method used. Results on held-out test data are shown in Figure 6; training loss plots can be found in the appendix in Figure 16.

Both TAPS and MOTAPS show favorable results compared to SP on all three problems. On the computer vision datasets, neither method quite reaches the generalization performance of SGD with a highly tuned step-wise learning rate schedule (95.2% for CIFAR10, 95.9% on SVHN). On the IWSLT14 problem, both TAPS and MOTAPS out-perform Adam [19] which achieved a 2.69 test loss and is the gold-standard for this task.

Appendix G. Missing Proofs

Here we present the missing proofs from the main text.

G.1. Proof of Lemma 3

First note that for the function in (19) we have that

$$\nabla_w h_{i,t}(w,\alpha) = \frac{f_i(w)-\alpha_i}{\|\nabla f_i(w^t)\|^2+1}\nabla f_i(w), \qquad \nabla_{\alpha_i} h_{i,t}(w,\alpha) = -\frac{f_i(w)-\alpha_i}{\|\nabla f_i(w^t)\|^2+1}, \tag{75}$$

$$\text{and} \quad \nabla_{\alpha_i} h_{n+1,t}(w,\alpha) = (\overline{\alpha}-\tau). \tag{76}$$

Proof The stationarity conditions of (17) are given by setting the gradients to zero, which from (76) we have that

$$\begin{aligned} \nabla_w h_t(w,\alpha) &= 0 \\ \nabla_{\alpha_i} h_t(w,\alpha) &= 0, \quad \text{for } i=1,\dots,n \\ &\Updownarrow \\ \frac{1}{n+1}\sum_{i=1}^n \frac{f_i(w)-\alpha_i}{\|\nabla f_i(w^t)\|^2+1}\nabla f_i(w) &= 0 \qquad (77) \\ \frac{f_i(w)-\alpha_i}{\|\nabla f_i(w^t)\|^2+1} &= (\overline{\alpha}-\tau), \quad \text{for } i=1,\dots,n. \qquad (78) \end{aligned}$$

If $\overline{\alpha}=\tau$ then from (78) we have that $f_i(w)=\alpha_i$ for all i, and thus from Assumption 1 we have that w must be a minimizer of (1), and thus a stationary point.

On the other hand, if $\overline{\alpha}\neq\tau$, then by substituting (78) into (77) gives

$$\frac{1}{n+1}\sum_{i=1}^n(\overline{\alpha}-\tau)\nabla f_i(w) = \frac{n(\overline{\alpha}-\tau)}{n+1}\left(\frac{1}{n}\sum_{i=1}^n\nabla f_i(w)\right) = 0. \tag{79}$$

Consequently since $\overline{\alpha} \neq \tau$, we have $\frac{1}{n}\sum_{i=1}^{n} \nabla f_i(w) = 0$ and thus w is a stationary point of (1).

Finally, if (w^*, α^*) is a minimizer of (17) then by Assumption 1 necessarily $h_t(w^*, \alpha^*) = 0$. Thus $f_i(w^*) = \alpha_i^*$ and $\overline{\alpha}^* = \tau$. Thus again by Assumption 1 we have that w^* must be a minimizer of (1). ■

G.2. Proof of Lemma 4

Proof First note that

$$\left\|\nabla_w h_{i,t}(w^t, \alpha)\right\|^2 \overset{(76)}{=} \left(\frac{f_i(w^t) - \alpha_i}{\left\|\nabla f_i(w^t)\right\|^2 + 1}\right)^2 \left\|\nabla f_i(w^t)\right\|^2. \tag{80}$$

Furthermore

$$\left\|\nabla_{\alpha_i} h_{i,t}(w^t, \alpha)\right\|^2 \overset{(76)}{=} \left(\frac{f_i(w^t) - \alpha_i}{\left\|\nabla f_i(w^t)\right\|^2 + 1}\right)^2. \tag{81}$$

Consequently adding (80) and (81) gives

$$\begin{aligned} \left\|\nabla h_{i,t}(w^t, \alpha)\right\|^2 &= \left\|\nabla_w h_{i,t}(w^t, \alpha)\right\|^2 + \left\|\nabla_{\alpha_i} h_{i,t}(w^t, \alpha)\right\|^2 \\ &\overset{(80)+(81)}{=} \left(\frac{f_i(w^t) - \alpha_i}{\left\|\nabla f_i(w^t)\right\|^2 + 1}\right)^2 \left(\left\|\nabla f_i(w^t)\right\|^2 + 1\right) \\ &= \frac{(f_i(w^t) - \alpha_i)^2}{\left\|\nabla f_i(w^t)\right\|^2 + 1} \overset{(19)}{=} 2h_{i,t}(w^t, \alpha). \end{aligned} \tag{82}$$

Furthermore

$$\left\|\nabla h_{n+1,t}(w, \alpha)\right\|^2 \overset{(76)}{=} \sum_{i=1}^{n} (\overline{\alpha} - \tau)^2 \overset{(19)}{=} 2h_{n+1,t}(w, \alpha). \tag{83}$$

Consequently

$$\frac{1}{n+1}\sum_{i=1}^{n+1} \left\|\nabla h_{i,t}(w^t, \alpha)\right\|^2 \overset{(83)+(82)}{=} \frac{1}{n+1}\sum_{i=1}^{n+1} 2h_{i,t}(w^t, \alpha) = 2h_t(w^t, \alpha).$$

■

G.3. Proof of Lemma 5

Lemma 22 *Let*

$$\alpha_i^* := f_i(w^*) \quad \textit{and} \quad \tau^* = f(w^*), \quad \textit{for } i = 1, \ldots, n. \tag{84}$$

It follows that

$$h_t(w^*, \alpha^*, \tau^*) = \frac{\lambda f(w^*)^2}{2(n+1)}. \tag{85}$$

Furthermore, every stationary point of (25) is a stationary point of (1). Finally if $f(w) \geq 0$ and $(w^*, \hat{\alpha}, \hat{\tau})$ is a minima of (25) then w^* is a minima of (1).

Proof Substituting (84) into (25) gives

$$h_t(w^*, \alpha^*, \tau^*) \;:=\; \frac{1}{n+1}\frac{\lambda}{2}(\tau^*)^2 \;=\; \frac{\lambda f(w^*)^2}{2(n+1)}.$$

Each stationary point of (25) satisfies

$$\nabla_w h_t(w, \alpha, \tau) \;=\; \frac{1-\lambda}{n+1}\sum_{i=1}^{n}\frac{f_i(w)-\alpha_i}{\|\nabla f_i(w^t)\|^2+1}\nabla f_i(w) = 0, \tag{86}$$

$$\nabla_{\alpha_i} h_t(w, \alpha, \tau) \;=\; \frac{1-\lambda}{n+1}\frac{\alpha_i - f_i(w)}{\|\nabla f_i(w^t)\|^2+1} + \frac{1-\lambda}{n+1}(\overline{\alpha}-\tau) = 0, \tag{87}$$

$$\nabla_{\tau} h_t(w, \alpha, \tau) \;=\; (1-\lambda)n(\tau-\overline{\alpha}) + \lambda\tau = 0. \tag{88}$$

From the last equation we have that

$$\overline{\alpha} - \tau = \frac{\lambda}{(1-\lambda)n}\tau, \tag{89}$$

and consequently substituting out $\overline{\alpha}-\tau$ in (87) by using (89) gives

$$\nabla_{\alpha_i} h_t(w, \alpha, \tau) \;=\; \frac{1-\lambda}{n+1}\frac{\alpha_i - f_i(w)}{\|\nabla f_i(w^t)\|^2+1} + \frac{1}{n+1}\frac{\lambda}{n}\tau = 0. \tag{90}$$

Passing the τ term to the other side gives

$$\frac{\lambda}{n}\tau = (1-\lambda)\frac{f_i(w)-\alpha_i}{\|\nabla f_i(w^t)\|^2+1}, \quad \text{for } i = 1, \ldots, n. \tag{91}$$

This allows us to substitute in (86) giving

$$\nabla_w h_t(w, \alpha, \tau) \;=\; \frac{\lambda\tau}{n+1}\left(\frac{1}{n}\sum_{i=1}^{n}\nabla f_i(w)\right) = 0. \tag{92}$$

From this we can conclude that if (w, α, τ) is a stationary point of (25), then w is a stationary point of our original objective function. Let (w, α, τ) be a stationary point. It follows from (89) that $\tau = \frac{(1-\lambda)n}{(1-\lambda)n+\lambda}\overline{\alpha}$, and thus after substituting into (25) gives

$$h_t(w, \alpha, \tau) \;:=\; \frac{1}{n+1}\left(\sum_{i=1}^{n}\frac{1-\lambda}{2}\frac{(f_i(w)-\alpha_i)^2}{\|\nabla f_i(w^t)\|^2+1} + \frac{n(1-\lambda)}{2}(\overline{\alpha}-\tau)^2 + \frac{\lambda}{2}\tau^2\right).$$

$$\begin{aligned} h_t(w, \alpha, \tau) &= \frac{1}{n+1}\left(\sum_{i=1}^{n}\frac{1-\lambda}{2}\frac{(f_i(w)-\alpha_i)^2}{\|\nabla f_i(w^t)\|^2+1} + \frac{n(1-\lambda)}{2}\left(\frac{\lambda}{n(1-\lambda)+\lambda}\overline{\alpha}\right)^2 + \frac{\lambda}{2}\frac{(1-\lambda)^2n^2}{(n(1-\lambda)+\lambda)^2}\overline{\alpha}^2\right) \\ &= \frac{1-\lambda}{n+1}\left(\sum_{i=1}^{n}\frac{1}{2}\frac{(f_i(w)-\alpha_i)^2}{\|\nabla f_i(w^t)\|^2+1} + \frac{1}{2}\frac{n\lambda}{n(1-\lambda)+\lambda}\overline{\alpha}^2\right) \end{aligned} \tag{93}$$

Furthermore, $\tau = \frac{(1-\lambda)n}{(1-\lambda)n+\lambda}\overline{\alpha}$ substituting into (90) and multiplying the result by $(n+1)$ gives

$$\frac{\alpha_i - f_i(w)}{\|\nabla f_i(w^t)\|^2+1} + \frac{\lambda}{n(1-\lambda)+\lambda}\overline{\alpha} = 0, \quad \text{for } i = 1, \ldots, n.$$

This can be re-arranged and written more compactly as the linear system

$$\left(\mathbf{D}^{-1} + \lambda \frac{\mathbf{1}\mathbf{1}^\top}{n(n(1-\lambda)+\lambda)}\right)\alpha = \mathbf{D}^{-1}F, \tag{94}$$

where

$$\mathbf{D} \coloneqq \mathbf{diag}\left(\left\|\nabla f_1(w^t)\right\|^2 + 1, \ldots, \left\|\nabla f_n(w^t)\right\|^2 + 1\right) \quad \text{and}$$
$$F = \left(f_1(w), \ldots, f_n(w)\right).$$

Using the Woodbury identity, the solution to the above is given by

$$\alpha = \left(\mathbf{D}^{-1} + \lambda \frac{\mathbf{1}\mathbf{1}^\top}{n(n(1-\lambda)+\lambda)}\right)^{-1} \mathbf{D}^{-1}F, \tag{95}$$

$$= \left(\mathbf{I} - \mathbf{D}\mathbf{1}\left(\frac{n(n(1-\lambda)+\lambda)}{\lambda} + \mathbf{1}^\top\mathbf{D}\mathbf{1}\right)^{-1}\mathbf{1}^\top\right)F \tag{96}$$

$$= \left(\mathbf{I} - \lambda\frac{\mathbf{D}\mathbf{1}\mathbf{1}^\top}{n(n(1-\lambda)+2\lambda) + \lambda\sum_{i=1}^n \left\|\nabla f_i(w^t)\right\|^2}\right)F. \tag{97}$$

Which reading line by line gives

$$\alpha_i = f_i(w) - \lambda\frac{\mathbf{D}e_i\sum_{j=1} f_j(w)}{n(n(1-\lambda)+2\lambda) + \lambda\sum_{j=1}^n \left\|\nabla f_j(w^t)\right\|^2}$$
$$= f_i(w) - \lambda\frac{\left(\left\|\nabla f_i(w^t)\right\|^2 + 1\right)\sum_{j=1} f_j(w)}{n(n(1-\lambda)+2\lambda) + \lambda\sum_{j=1}^n \left\|\nabla f_j(w^t)\right\|^2}. \tag{98}$$

Taking the average over i in the above gives

$$\overline{\alpha} = f(w) - \lambda f(w)\frac{n + \sum_{j=1}^n \left\|\nabla f_j(w^t)\right\|^2}{n(n(1-\lambda)+2\lambda) + \lambda\sum_{j=1}^n \left\|\nabla f_j(w^t)\right\|^2}$$
$$= f(w)\left(1 - \lambda\frac{n + \sum_{j=1}^n \left\|\nabla f_j(w^t)\right\|^2}{n(n(1-\lambda)+2\lambda) + \lambda\sum_{j=1}^n \left\|\nabla f_j(w^t)\right\|^2}\right)$$
$$= f(w)\frac{n(n(1-\lambda)+\lambda)}{n(n(1-\lambda)+2\lambda) + \lambda\sum_{j=1}^n \left\|\nabla f_j(w^t)\right\|^2} \tag{99}$$

Substituting (98) and (99) into (93) gives

$$
\begin{aligned}
h_t(w,\alpha,\tau)\frac{n+1}{1-\lambda} &= \sum_{i=1}^{n}\frac{\lambda^2}{2}\frac{\left(\frac{\left(\left\|\nabla f_i(w^t)\right\|^2+1\right)\sum_{j=1} f_j(w)}{n(n(1-\lambda)+2\lambda)+\lambda\sum_{j=1}^n\|\nabla f_j(w^t)\|^2}\right)^2}{\left\|\nabla f_i(w^t)\right\|^2+1}+\frac{1}{2}\frac{n\lambda}{n(1-\lambda)+\lambda}\overline{\alpha}^2 \\
&= \sum_{i=1}^{n}\frac{\lambda^2 n^2}{2}f(w)^2\frac{\left\|\nabla f_i(w^t)\right\|^2+1}{\left(n(n(1-\lambda)+2\lambda)+\lambda\sum_{j=1}^n\|\nabla f_j(w^t)\|^2\right)^2}+\frac{1}{2}\frac{n\lambda}{n(1-\lambda)+\lambda}\overline{\alpha}^2 \\
&= \sum_{i=1}^{n}\frac{\lambda^2 n^2}{2}f(w)^2\frac{\left\|\nabla f_i(w^t)\right\|^2+1}{\left(n(n(1-\lambda)+2\lambda)+\lambda\sum_{j=1}^n\|\nabla f_j(w^t)\|^2\right)^2} \\
&\quad +\frac{1}{2}\frac{n\lambda}{n(1-\lambda)+\lambda}\left(f(w)\frac{n(n(1-\lambda)+\lambda)}{n(n(1-\lambda)+2\lambda)+\lambda\sum_{j=1}^n\|\nabla f_j(w^t)\|^2}\right)^2 \\
&= \frac{\lambda}{2}f(w)^2\frac{n^2}{n(n(1-\lambda)+2\lambda)+\lambda\sum_{j=1}^n\|\nabla f_j(w^t)\|^2},
\end{aligned}
$$

where in first equality we used (98) and in the third equality we used (99). Since w^t is fixed, and every minima of (25) is a stationary point, we have that the minima in w of the above is given by

$$w^* \in \arg\min f(w)^2 = \arg\min f(w),$$

where we used the positivity of $f(w)$.

■

G.4. Proof of Lemma 9

Here we prove an extended version of Lemma 9 with some additional intermediary results that make the lemma easier to follow.

Lemma 23 *Consider the functions*

$$h_{i,t}(w,\alpha,\tau) := \frac{1-\lambda}{2}\frac{(f_i(w)-\alpha_i)^2}{\left\|\nabla f_i(w^t)\right\|^2+1}, \quad \text{for } i=1,\dots,n, \tag{100}$$

and $h_{n+1,t}(w,\alpha,\tau)$ given in (37). It follows that $h_t(w,\alpha,\tau)$ defined in (25) is equivalent to

$$h_t(w,\alpha,\tau) = \frac{1}{n+1}\sum_{i=1}^{n+1}h_{i,t}(w,\alpha,\tau) \tag{101}$$

Furthermore, if

$$\lambda \le \frac{2n+1}{2n+3} < 1 \tag{102}$$

then

$$\left\|\nabla h_{i,t}(w^t,\alpha,\tau)\right\|^2 = 2h_{i,t}(w^t,\alpha,\tau), \tag{103}$$

$$\|\nabla h_{n+1,t}(w,\alpha,\tau)\|^2 \le 2(1-\lambda)(2n+1)h_{n+1,t}(w,\alpha,\tau), \tag{104}$$

and consequently

$$\frac{1}{n+1}\sum_{i=1}^{n+1}\left\|\nabla h_{i,t}(w^t,\alpha,\tau)\right\|^2 \;\leq\; 2(1-\lambda)(2n+1)h_t(w^t,\alpha,\tau). \tag{105}$$

Proof Using the definitions of $h_t(w^t,\alpha,\tau)$ in (25) we have that (101) holds.

Furthermore (103) follows from Lemma 4. As for $h_{n+1,t}(w,\alpha,\tau)$ in (37) we have that

$$\begin{aligned} h_{n+1,t}(w,\alpha,\tau) &= \frac{n(1-\lambda)}{2}(\overline{\alpha}-\tau)^2+\frac{\lambda}{2}\tau^2 \\ \nabla_\tau h_{n+1,t}(w,\alpha,\tau) &= (1-\lambda)n(\tau-\overline{\alpha})+\lambda\tau. \\ \nabla_\alpha h_{n+1,t}(w,\alpha,\tau) &= (1-\lambda)\mathbf{1}(\overline{\alpha}-\tau) \end{aligned} \tag{106}$$

Consequently

$$\begin{aligned} \|\nabla h_{n+1,t}(w,\alpha,\tau)\|^2 &= ((1-\lambda)n(\tau-\overline{\alpha})+\lambda\tau)^2+(1-\lambda)^2\|\mathbf{1}\|^2(\overline{\alpha}-\tau)^2 \\ &\leq 2(1-\lambda)^2n^2(\tau-\overline{\alpha})^2+2\lambda^2\tau^2+(1-\lambda)^2n(\overline{\alpha}-\tau)^2 \\ &= 2(1-\lambda)(2n+1)\frac{(1-\lambda)n(\tau-\overline{\alpha})^2}{2}+4\lambda\frac{\lambda\tau^2}{2} \\ &\leq 2\max\{(1-\lambda)(2n+1),\;2\lambda\}h_{n+1,t}(w,\alpha,\tau). \end{aligned}$$

Due to (102) we have that

$$\max\{(1-\lambda)(2n+1),\;2\lambda\} \;=\; (1-\lambda)(2n+1).$$

This proves (103). As a consequence from (103) and (104) we have that

$$\begin{aligned} \frac{1}{n+1}\sum_{i=1}^{n+1}\left\|\nabla h_{i,t}(w^t,\alpha,\tau)\right\|^2 &\leq \frac{2\max\{1,(1-\lambda)(2n+1)\}}{n+1}\sum_{i=1}^{n+1}h_{i,t}(w^t,\alpha,\tau) \quad \text{(Using (103) and (104))} \\ &\leq 2(1-\lambda)(2n+1)h_t(w^t,\alpha,\tau). \quad \text{(Using (102) and (25))} \end{aligned}$$

■

G.5. Proof of Theorem 10

Here we give the proof of Theorem 10. We prove a slightly more general version of Theorem 10 by not requiring that the auxiliary function is zero at the optimal. That is $h_t(z^*)$ may be non-zero. The exact result in Theorem 10 follows from applying the following Theorem 24 with $h_t(z^*)=0$.

Theorem 24 (Star-convexity) *Suppose Assumption 2 holds with $G>0$. Let $\gamma<1/G$ and suppose there exists z^* such that h_t is star-convex at z^t and around z^*, that is*

$$h_t(z^*) \;\geq\; h_t(z^t)+\left\langle\nabla h_t(z^t),z^*-z^t\right\rangle, \tag{107}$$

then we have that

$$\begin{aligned} \min_{t=1,\ldots,k}\mathbb{E}\left[h_t(z^t)-h_t(z^*)\right] &\leq \frac{1}{k}\sum_{t=0}^{k}\mathbb{E}\left[h_t(z^t)-h_t(z^*)\right] \\ &\leq \frac{1}{k}\frac{1}{2\gamma(1-G\gamma)}\mathbb{E}\left[\left\|z^0-z^*\right\|^2\right]+\frac{G\gamma}{1-G\gamma}\frac{1}{k}\sum_{t=1}^{k}h_t(z^*). \end{aligned} \tag{108}$$

Proof []

This proof is partially based on Theorems 4.3 [35]. Let $\mathbb{E}_t[\cdot] := \mathbb{E}\left[\cdot \mid z^t\right]$ denote the expectation conditioned on z^t.

Expanding the squares we have

$$\begin{aligned}
\mathbb{E}_t\left[\left\|z^{t+1}-z^*\right\|^2\right] &\leq \left\|z^t-z^*\right\|^2 - 2\gamma\left\langle \nabla_w h_t(z^t), z^t-z^*\right\rangle + \gamma^2\mathbb{E}_t\left[\left\|\nabla_w h_{t,i_t}(z^t)\right\|^2\right] \\
&\overset{(44)}{\leq} \left\|z^t-z^*\right\|^2 - 2\gamma\left\langle \nabla_w h_t(z^t), z^t-z^*\right\rangle + 2G\gamma^2 h_t(z^t) \\
&\overset{(107)}{\leq} \left\|z^t-z^*\right\|^2 - 2\gamma(h_t(z^t)-h_t(z^*)) + 2G\gamma^2 h_t(z^t) \\
&= \left\|z^t-z^*\right\|^2 - 2\gamma(1-G\gamma)(h_t(z^t)-h_t(z^*)) + 2G\gamma^2 h_t(z^*) \qquad (109)
\end{aligned}$$

Taking expectation, re-arranging and summing both sides from $t=0,\ldots,k$ we have that

$$\begin{aligned}
\sum_{t=0}^{k}\mathbb{E}\left[h_t(z^t)-h_t(z^*)\right] &\leq \frac{1}{2\gamma(1-G\gamma)}\sum_{t=0}^{k}\left(\mathbb{E}\left[\left\|z^t-z^*\right\|^2\right]-\mathbb{E}\left[\left\|z^{t+1}-z^*\right\|^2\right]\right)+\frac{G\gamma}{1-G\gamma}\sum_{t=0}^{k}h_t(z^*) \\
&\leq \frac{1}{2\gamma(1-G\gamma)}\mathbb{E}\left[\left\|z^0-z^*\right\|^2\right]+\frac{G\gamma}{1-G\gamma}\sum_{t=0}^{k}h_t(z^*). \qquad (110)
\end{aligned}$$

Now dividing through by k gives (108). ■

G.6. Proof of Theorem 11

Theorem 25 *Suppose Assumption 2 holds with $G>0$. Let $\gamma \leq 1/G$. If there exists $\mu>0$ and z^* such that h_t is μ–strongly star–convex along z^t and around z^*, that is*

$$h_t(z^*) \geq h_t(z^t) + \left\langle \nabla h_t(z^t), z^*-z^t\right\rangle + \frac{\mu}{2}\left\|z^*-z^t\right\|, \qquad (111)$$

then

$$\mathbb{E}\left[\left\|z^{t+1}-z^*\right\|^2\right] \leq (1-\gamma\mu)^{t+1}\left\|z^0-z^*\right\|^2 + 2G\gamma^2\sum_{i=0}^{t}(1-\gamma\mu)^i\mathbb{E}\left[h_i(z^*)\right]. \qquad (112)$$

Finally, if $h_t(z^*)=0$ for all t then we have that (112) and (44) together imply that $\mu \leq G$ and thus (112) gives linear convergence.

Proof This proof is partially based on 4.10 in [35], which in turn is based on Theorem 6 in [32], Thereom 4.1 in [13] and Theorem 3.1 in [14].

Expanding the squares we have that

$$\begin{aligned}
\mathbb{E}_t\left[\left\|z^{t+1}-z^*\right\|^2\right] &\leq \left\|z^t-z^*\right\|^2 - 2\gamma\left\langle \nabla_w h_t(z^t), z^t-z^*\right\rangle + \gamma^2\mathbb{E}_t\left[\left\|\nabla_w h_{t,i_t}(z^t)\right\|^2\right] \\
&\overset{(44)}{\leq} \left\|z^t-z^*\right\|^2 - 2\gamma\left\langle \nabla_w h_t(z^t), z^t-z^*\right\rangle + 2G\gamma^2 h_t(z^t) \\
&\overset{(111)}{\leq} (1-\gamma\mu)\left\|z^t-z^*\right\|^2 - 2\underbrace{\gamma(1-G\gamma)(h_t(z^t)-h_t(z^*))}_{\geq 0} + 2G\gamma^2 h_t(z^*) \\
&\leq (1-\gamma\mu)\left\|z^t-z^*\right\|^2 + 2G\gamma^2 h_t(z^*), \qquad (113)
\end{aligned}$$

where to get to the last line we used that $(1-G\gamma)(h_t(z^t)-h_t(z^*)) \geq 0$ which holds because $\gamma \leq \frac{1}{G}$. Taking the expectation and applying the above recursively gives

$$\mathbb{E}_t\left[\left\|z^{t+1}-z^*\right\|^2\right] \;\leq\; (1-\gamma\mu)^{t+1}\left\|z^0-z^*\right\|^2 + 2G\gamma^2\sum_{i=0}^{t}(1-\gamma\mu)^i h_i(z^*) \tag{114}$$

which is the result (112).

Furthermore, if $h_t(z^*)=0$ we have that $\mu \leq G$ follows from a small modification of Theorem 4.10 in [35]. Indeed taking expectation over (111) and using (44) we have that

$$\begin{aligned} h_t(z^*) &\geq \frac{1}{2G}\mathbb{E}\left[\left\|\nabla h_{t,i_t}(z^t)\right\|^2\right] + \left\langle \nabla h_t(z^t), z^*-z^t\right\rangle + \frac{\mu}{2}\left\|z^*-z^t\right\| \\ &= \frac{G}{2}\mathbb{E}\left[\left\|z^*-z^t-\frac{1}{L}\nabla h_{t,i_t}(z^t)\right\|^2\right] - \frac{G-\mu}{2}\left\|z^*-z^t\right\|. \end{aligned} \tag{115}$$

Rearranging and using that $h_t(z^*)=0$ gives

$$\frac{G-\mu}{2}\left\|z^*-z^t\right\| \geq \frac{L}{2}\mathbb{E}\left[\left\|z^*-z^t-\frac{1}{G}\nabla h_{t,i_t}(z^t)\right\|^2\right] \geq 0.$$

Thus $\mu \leq G$. ■

Appendix H. Convergence of The Stochastic Polyak Method

Here we explore sufficient conditions for the assumptions in Theorems 10 and 11 to hold for the SP method (6). To this end, let

$$h_t(w) \;:=\; \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\frac{(f_i(w)-f_i(w^*))^2}{\|\nabla f_i(w^t)\|^2}, \tag{116}$$

$$h_{i,t}(w) \;:=\; \frac{1}{2}\frac{(f_i(w)-f_i(w^*))^2}{\|\nabla f_i(w^t)\|^2}. \tag{117}$$

We will also explore the consequences of these theorems. In these section we say that f_i is L_i smooth if

$$f_i(z) \;\leq\; f_i(w) + \langle \nabla f_i(w), z-w\rangle + \frac{L_i}{2}\|z-w\|^2, \quad \forall z,w\in\mathbb{R}^d. \tag{118}$$

We will also use the interpolation Assumption 3 throughout this section. Thus

$$f_i(w^*) \;=\; \min_{w\in\mathbb{R}^d} f_i(w) \;\leq\; f(z), \quad \text{for all } i\in\{1,\dots,n\},\ z\in\mathbb{R}^d.$$

Using smoothness and interpolation, we first establish the following descent lemma.

Lemma 26 *If the interpolation Assumption 3 holds and each $f_i(w)$ is L_i–smooth (118) then*

$$f_i(w)-f_i(w^*) \;\geq\; \frac{1}{2L_i}\|\nabla f_i(w)\|^2, \quad \forall w\in\mathbb{R}^d,\ i=1,\dots,n. \tag{119}$$

Proof Let w^* be a minimizer of $f(w)$. Consequently by the interpolation assumption for every $z \in \mathbb{R}^d$ we have that $f_i(w^*) - f_i(z) \leq 0$ and for every $w \in \mathbb{R}^d$ we have that

$$\begin{aligned} f_i(w^*) - f_i(w) &\leq f_i(w^*) - f_i(z) + f_i(z) - f_i(w) \\ &\leq f_i(z) - f_i(w) \\ &\overset{(118)}{\leq} \langle \nabla f_i(w), z - w \rangle + \frac{L_i}{2} \|z - w\|^2 \end{aligned}$$

Minimizing the right hand side in z gives $z = w - \frac{1}{L_i}\nabla f_i(w)$ which when plugged in the above gives

$$f_i(w^*) - f_i(w) \leq -\frac{1}{2L_i} \|\nabla f_i(w)\|^2 .$$

Re-arranging gives (119).

■

H.1. Proof of Lemma 12

First we show that, under interpolation, if f_i is star-convex, then the auxiliary functions in (116) and (117) are also star convex....

Lemma 27 *Let the interpolation Assumption 3 hold. If every f_i is star convex along the iterates (w^t) given by (6), that is,*

$$f_i(w^*) \geq f_i(w) + \langle \nabla f_i(w), w^* - w \rangle \tag{120}$$

then $h_{i,t}(w)$ is star convex along the iterates (w^t) with

$$h_{i,t}(w^*) \geq h_{i,t}(w^t) + \left\langle \nabla_w h_{i,t}(w^t), w^* - w \right\rangle , \tag{121}$$

so long as $w^t \neq w^*$. Consequently we have that h_t is star convex around w^*.

Furthermore if f_i is μ_i-strongly convex and L_i–smooth then $h_{i,t}$ is $\frac{1}{2}\frac{\mu_i}{L_i}$–strongly star-convex. Consequently $h_t(w)$ is $\frac{1}{2n}\sum_{i=1}\frac{\mu_i}{L_i}$–strongly star-convex

$$h_t(w^*) \geq h_t(w^t) + \left\langle \nabla_w h_t(w^t), w^* - w \right\rangle + \frac{1}{4n}\sum_{i=1}^{n} \frac{\mu}{L_i} \left\|w^t - w^*\right\|^2 . \tag{122}$$

Proof Using that $h_{i,t}(w^*) = 0$ and that $h_{i,t}(w^t) > 0$ since $w^t \neq w^*$ we have that

$$h_{i,t}(w^*) \geq h_{i,t}(w^t) + \left\langle \nabla_w h_{i,t}(w^t), w^* - w \right\rangle$$

$$\Updownarrow \text{ (By definition (117))}$$

$$0 \geq \frac{1}{2}\left(\frac{f_i(w^t) - f_i(w^*)}{\|\nabla f_i(w^t)\|}\right)^2 + \left\langle \frac{f_i(w^t) - f_i(w^*)}{\|\nabla f_i(w^t)\|^2}\nabla f_i(w^t), w^* - w^t \right\rangle$$

$$\Updownarrow \left(\text{Multipling by } \left\|\nabla f_i(w^t)\right\|^2 / (f_i(w^t) - f_i(w^*)) \geq 0.\right)$$

$$0 \geq \frac{1}{2}(f_i(w^t) - f_i(w^*)) + \left\langle \nabla f_i(w^t), w^* - w^t \right\rangle$$

$$\Uparrow \text{ (Using that } f_i(w^t) - f_i(w^*) \geq 0 \text{)}$$
$$f_i(w^*) \geq f_i(w^t) + \left\langle \nabla f_i(w^t), w^* - w^t \right\rangle,$$

where we used $f_i(w^t) - f_i(w^*) \geq 0$ which is a consequence of interpolation. This proves (121)

Now if we assume that f_i is μ-strongly star-convex and L_i–smooth then we have that by

$$h_{i,t}(w^*) \geq h_{i,t}(w^t) + \left\langle \nabla_w h_{i,t}(w^t), w^* - w \right\rangle + \frac{1}{4}\frac{\mu}{L_i}\left\|w^t - w^*\right\|^2 \tag{123}$$
$$\Updownarrow \text{ (By definition (117))}$$
$$0 \geq \frac{1}{2}\left(\frac{f_i(w^t) - f_i(w^*)}{\|\nabla f_i(w^t)\|}\right)^2 + \left\langle \frac{f_i(w^t) - f_i(w^*)}{\|\nabla f_i(w^t)\|^2}\nabla f_i(w^t), w^* - w^t \right\rangle + \frac{1}{4}\frac{\mu}{L_i}\left\|w^t - w^*\right\|^2$$
$$\Updownarrow \left(\text{Multipling by } \left\|\nabla f_i(w^t)\right\|^2 / (f_i(w^t) - f_i(w^*)) \geq 0.\right)$$
$$0 \geq \frac{1}{2}(f_i(w^t) - f_i(w^*)) + \left\langle \nabla f_i(w^t), w^* - w^t \right\rangle + \frac{\left\|\nabla f_i(w^t)\right\|^2}{f_i(w^t) - f_i(w^*)}\frac{1}{4}\frac{\mu}{L_i}\left\|w^t - w^*\right\|^2$$
$$\Uparrow \text{ (Using that } f_i(w^t) - f_i(w^*) \geq 0 \text{)}$$
$$f_i(w^*) \geq f_i(w^t) + \left\langle \nabla f_i(w^t), w^* - w^t \right\rangle + \frac{\left\|\nabla f_i(w^t)\right\|^2}{f_i(w^t) - f_i(w^*)}\frac{1}{4}\frac{\mu}{L_i}\left\|w^t - w^*\right\|^2.$$

Finally, from smoothness and Lemma 26 we have that $1 \geq \frac{1}{2L_i}\frac{\left\|\nabla f_i(w^t)\right\|^2}{f_i(w^t) - f_i(w^*)}$ consequently

$$\begin{aligned} f_i(w^*) &\geq f_i(w^t) + \left\langle \nabla f_i(w^t), w^* - w^t \right\rangle + \frac{\mu}{2}\left\|w^t - w^*\right\|^2 \\ &\geq f_i(w^t) + \left\langle \nabla f_i(w^t), w^* - w^t \right\rangle + \frac{1}{4}\frac{\mu}{L_i}\frac{\left\|\nabla f_i(w^t)\right\|^2}{f_i(w^t) - f_i(w^*)}\left\|w^t - w^*\right\|^2. \end{aligned} \tag{124}$$

Consequently the above implications hold, and thus $h_{i,t}$ is $\frac{1}{4}\frac{\mu}{L_i}$-strongly star convex. Taking the average of (123) over i gives (122), which concludes the proof. ■

H.2. Proof of Corollary 13 and 14

Having established when h_t is star convex and strongly star convex, we can now apply Theorems 10 and Theorem 11, which when specialized to SP gives the following corollaries. This result has already been established in Theorem 4.4 and Theorem D.3 in [13]. Thus here we have showed that the results in [13] follow as a direct consequence of the interpretation of SP as a variant of the online SGD method.

Corollary 28 *If $\gamma < 1$ and every $f_i(w)$ is star-convex along the iterates (w^t) given by (6) then*

$$\frac{1}{k}\sum_{t=0}^{k}\frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}\left[\left(\frac{f_i(w^t) - f_i(w^*)}{\|\nabla f_i(w^t)\|}\right)^2\right] \leq \frac{1}{k}\frac{1}{2\gamma(1-\gamma)}\mathbb{E}\left[\left\|w^0 - w^*\right\|^2\right]. \tag{125}$$

Furthermore if the interpolation Assumption 3 holds and if each $f_i(w)$ is L_i–smooth then

$$\min_{t=0,\ldots,k}\mathbb{E}\left[f(w^t) - f^*\right] \leq \frac{1}{k}\frac{L_{\max}}{2\gamma(1-\gamma)}\mathbb{E}\left[\left\|w^0 - w^*\right\|^2\right], \tag{126}$$

where $L_{\max} := \max_{i=1,\ldots,n} L_i$.

Proof The proof of (125) follows as a special case of Theorem 10 by identifying h_t with (116) and $h_{i_t,t}$ with (117). Indeed, according to (10) we have that h_t satisfies the growth condition (44) with $G = 1$ and according to (121) h_t is star-convex (107) around w^*. Finally since $h_t(w^*) = 0$ the result (125) follows by Theorem 10.

The result (126) would follow from (125) if

$$L_{\max} \frac{1}{n} \sum_{i=1}^{n} \frac{(f_i(w) - f_i(w^*))^2}{\|\nabla f_i(w)\|^2} \geq 2(f(w) - f^*). \tag{127}$$

This Assumption has appeared recently in [13] where it was proven that (127) is a consequence of each $f_i(w)$ being L_i–smooth. We give a simpler proof next for completeness. That is, assuming that there exists w such that $f_i(w) \neq f_i(w^*)$ and thus $\nabla f_i(w) \neq 0$ (otherwise (127) holds trivially) we have from (119) that

$$\frac{1}{\|\nabla f_i(w)\|^2} \geq \frac{1}{2L_i(f_i(w) - f_i(w^*))}.$$

Multiplying both sides by $(f_i(w) - f_i(w^*))^2$ and averaging over $i = 1, \ldots, n$ gives

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} \frac{(f_i(w) - f_i(w^*))^2}{\|\nabla f_i(w)\|^2} \geq 2 \frac{1}{n} \sum_{i=1}^{n} \frac{f_i(w) - f_i(w^*)}{L_i} &\geq \frac{1}{n} \sum_{i=1}^{n} \frac{2f_i(w) - f_i(w^*)}{\max_{i=1,\ldots,n} L_i} \\ &= \frac{2(f(w) - f^*)}{L_{\max}}. \end{aligned}$$

Using (127) and (125) we have

$$\begin{aligned} \min_{t=0,\ldots,k} \mathbb{E}\left[f(w^t) - f^*\right] &\leq \frac{1}{k} \sum_{t=0}^{k} \mathbb{E}\left[f(w^t) - f^*\right] \\ &\leq \frac{1}{k} \sum_{t=0}^{k} \frac{L_{\max}}{2n} \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{f_i(w^t) - f_i(w^*)}{\|\nabla f_i(w^t)\|}\right)^2\right] \\ &\leq \frac{1}{k} \frac{L_{\max}}{2\gamma(1-\gamma)} \mathbb{E}\left[\left\|w^0 - w^*\right\|^2\right] \end{aligned}$$

which concludes the proof of ■

Corollary 29 *If $\gamma \leq 1$, the interpolation Assumption 3 holds, and every f_i is L_i–smooth and μ–strongly star-convex then the iterates w^t given by (6) converge linearly according to*

$$\mathbb{E}\left[\left\|w^{t+1} - w^*\right\|^2\right] \leq \left(1 - \gamma \frac{1}{2n} \sum_{i=1}^{n} \frac{\mu_i}{L_i}\right)^{t+1} \left\|w^0 - w^*\right\|^2 \tag{128}$$

Proof The proof of (128) follows as a special case of Theorem 11 by identifying h_t with (116) and $h_{i_t,t}$ with (117). Indeed, according to (10) we have that h_t satisfies the growth condition (44) with $G = 1$. Furthermore f_i is μ_i–strongly star convex and L_i–smooth, then from Lemma 27 we have that h_t is $\frac{1}{2n} \sum_{i=1}^{n} \frac{\mu_i}{L_i}$–strongly star convex. Finally since $h_t(w^*) = 0$ the result (125) follows by Theorem 11. ■

Appendix I. Convergence of the Targeted Stochastic Polyak Stepsize

Here we explore the consequences and conditions of Theorem 10 for the TAPS method given in Algorithm 1.

I.1. Proof of Corollary 30 and more

First we re-state Theorem 10 specialized to Algorithm 1.

Corollary 30 *Let $h_t(z)$ be defined in (17) and suppose that $h_t(z)$ is star convex (107) around $z^* = (w^*, \alpha^*)$ and along the iterates $z^t = (w^t, \alpha^t)$ of Algorithm 1.*

If $\gamma < 1$ and in addition $f_i(w)$ is $L_{\max}$–Lipschitz then

$$\min_{t=1,\ldots,k} \frac{1}{n+1}\left(\sum_{i=1}^{n} \frac{\mathbb{E}\left[f_i(w^t)-\alpha_i^t\right]^2}{L_{\max}+1} + \mathbb{E}\left[\overline{\alpha}^t - \tau\right]^2\right) \;\leq\; \frac{1}{k}\frac{1}{\gamma(1-\gamma)}\mathbb{E}\left[\left\|w^0 - w^*\right\|^2\right]. \tag{129}$$

Alternatively, if $h_t(z)$ is μ–strongly star–convex (111) then

$$\mathbb{E}\left[\left\|w^t - w^*\right\|^2 + \sum_{i=1}^{n}\left\|\alpha_i^t - f_i(w^*)\right\|^2\right] \leq (1-\gamma\mu)^t\left(\left\|w^0 - w^*\right\|^2 + \sum_{i=1}^{n}\left\|\alpha_i^0 - f_i(w^0)\right\|^2\right). \tag{130}$$

Theorem 30 provides us with a $\mathcal{O}(1/k)$ convergence in expectation when $h_t(z)$ is star convex. Indeed, the bound in (129) shows that $\overline{\alpha}$ converges to τ at a rate of $\mathcal{O}(1/k)$. Finally from the target assumption (12) we have that $h_t(z^*) = 0$, thus $f_i(w^t)$ and α_i^t converge to $f_i(w^*)$ at a rate of $\mathcal{O}(1/k)$.

Proof The proof follows by applying Theorem 10. Indeed, by letting $h_{i,t}(z) = \frac{1}{2}\frac{(f_i(w)-\alpha_i)^2}{\|\nabla f_i(w^t)\|^2+1}$ for $i = 1,\ldots,n$ and $h_{n+1,t}(z) = \frac{n}{2}(\overline{\alpha}-\tau)^2$. Thus $h_t(z) = \frac{1}{n+1}\sum_{i=1}^{n+1} h_{i,t}(z)$. By Lemma 4 we have that

$$\begin{aligned}
\mathbb{E}_{i\sim\frac{1}{n+1}}\left[\|\nabla h_{i,t}(z^t)\|^2\right] &= \frac{1}{n+1}\sum_{i=1}^{n+1}\|\nabla h_{i,t}(z^t)\|^2 \\
&= \frac{1}{n+1}\left(\sum_{i=1}^{n}\left\|\nabla f_{i,w^t}(w^t,\alpha^t)\right\|^2 + \left\|\nabla h_{n+1}(\alpha^t)\right\|^2\right) \\
&\overset{(23)}{=} \frac{1}{n+1}\left(\sum_{i=1}^{n} 2f_{i,w^t}(w^t,\alpha^t) + 2f_{n_1}(\alpha^t)\right) \\
&\overset{(56)+(19)}{=} 2h_t(z^t).
\end{aligned}$$

Consequently h_t satisfies the growth condition (44) with $G = 1$. By assumption h_t is star convex along the iterates z^t, thus the two condition required for Theorem 10 to hold are satisfied, and as a consequence, we have that (108) holds. Substituting out $h_t(z^t)$ we have that

$$\frac{1}{k}\sum_{t=0}^{k}\frac{1}{n+1}\left(\sum_{i=1}^{n}\frac{1}{2}\frac{(f_i(w^t)-\alpha_i^t)^2}{\|\nabla f_i(w^t)\|^2+1} + \frac{n}{2}(\overline{\alpha}^t-\tau)^2\right) \;\leq\; \frac{1}{k}\frac{1}{2\gamma(1-\gamma)}\mathbb{E}\left[\left\|w^0-w^*\right\|^2\right]. \tag{131}$$

Furthermore, if f_i is $L_{\max}$–Lipschitz, that is if $\|\nabla f_i(w^t)\| \leq L_{\max}$ then from (131) we have that

$$\frac{1}{k}\sum_{t=0}^{k}\frac{1}{n+1}\left(\sum_{i=1}^{n}\frac{1}{2}\frac{(f_i(w^t)-\alpha_i^t)^2}{L_{\max}+1} + \frac{n}{2}(\overline{\alpha}^t-\tau)^2\right) \;\leq\; \frac{1}{k}\frac{1}{2\gamma(1-\gamma)}\mathbb{E}\left[\left\|w^0-w^*\right\|^2\right], \tag{132}$$

from which (129) follows by lower bounding the average over k by the minimum.

Finally, if there exists $\mu > 0$ such that $h_t(z)$ is strongly star-convex (111), then by noting that

$$\|z - z^*\|^2 = \left\|w^t - w^*\right\|^2 + \left\|\alpha^t - \alpha^*\right\|^2 = \left\|w^t - w^*\right\|^2 + \sum_{i=1}^n \left\|\alpha_i^t - f_i(w^*)\right\|^2$$

we have that (112) gives (130).

■

I.2. Proof of Lemmas 15 and Corollary 16

For ease of reference, we first re-state the lemmas.

Lemma 31 (Locally Convex) *Consider the iterates of Algorithm 2. Let $(w, \alpha) \in \mathbb{R}^{d+n}$ and consider $h_t(w, \alpha)$ defined in (56) . Assume that the gradients at w spans the entire space, that is*

$$\text{span}\left\{\nabla f_1(w), \ldots, \nabla f_n(w)\right\} \;=\; \mathbb{R}^d, \quad \forall w. \tag{133}$$

If Assumption 1 holds, every $f_i(w)$ for $i = 1, \ldots, n$ is twice continuously differentiable and

$$\frac{1}{n+1} \sum_{i=1}^n \nabla^2 f_i(w^t) \frac{f_i(w^t) - \alpha_i^t}{\|\nabla f_i(w^t)\|^2 + 1} \succeq 0, \quad \forall t, \tag{134}$$

then h_t is strictly convex with at (w^t, α^t) that is

$$\nabla^2 h_t(w^t, \alpha^t) \;\succ\; 0, \quad \forall t.$$

Proof We have $(f_i(w) - \alpha_i)^2$ is locally convex, and thus star convex, iff its Hessian is positive definite around (w^*, α^*). Computing the gradient of $(f_i(w) - \alpha_i)^2$ we have that

$$\nabla (f_i(w) - \alpha_i)^2 = 2 \begin{bmatrix} \nabla f_i(w) \\ -1 \end{bmatrix} (f_i(w) - \alpha_i)$$

Computing the Hessian gives

$$\begin{aligned} \nabla^2 (f_i(w) - \alpha_i)^2 &= 2 \begin{bmatrix} \nabla f_i(w) \\ -1 \end{bmatrix} \begin{bmatrix} \nabla f_i(w)^\top & -1 \end{bmatrix} + 2 \begin{bmatrix} \nabla^2 f_i(w) & 0 \\ 0 & 0 \end{bmatrix} (f_i(w) - \alpha_i) \\ &= 2 \begin{bmatrix} \nabla f_i(w) \nabla f_i(w)^\top & -\nabla f_i(w) \\ -\nabla f_i(w)^\top & 1 \end{bmatrix} + 2 \begin{bmatrix} \nabla^2 f_i(w) & 0 \\ 0 & 0 \end{bmatrix} (f_i(w) - \alpha_i) \end{aligned} \tag{135}$$

Now let $\mathbf{I}_n \in \mathbb{R}^{n \times n}$ be the identity matrix in $\mathbb{R}^{n \times n}$, let

$$\begin{aligned} \mathbf{D}_t &\coloneqq \mathbf{diag}\left(\frac{1}{\|\nabla f_1(w^t)\|^2 + 1}, \ldots, \frac{1}{\|\nabla f_n(w^t)\|^2 + 1} \right) \in \mathbb{R}^{n \times n} \\ \mathbf{H}_t(w, \alpha) &\coloneqq \sum_{i=1}^{n+1} \nabla^2 f_i(w) \frac{f_i(w) - \alpha_i}{\|\nabla f_i(w^t)\|^2 + 1} \end{aligned} \tag{136}$$

and let

$$DF(w) \;\coloneqq\; \big[\nabla f_1(w), \ldots, \nabla f_n(w)\big] \in \mathbb{R}^{d\times n}.$$

Using (135) and by the definition of h_t in (56) we have that

$$\nabla^2 h_t(w,\alpha) \;=\; \frac{1}{n+1}\underbrace{\begin{bmatrix} DF(w)\mathbf{D}_t DF(w)^\top & -DF(w)\mathbf{D}_t \\ -(DF(w)\mathbf{D}_t)^\top & \mathbf{I}_n(1+\frac{1}{n}), \end{bmatrix}}_{\coloneqq \mathbf{M}_t(w)} + \begin{bmatrix} \mathbf{H}_t(w,\alpha) & 0 \\ 0 & 0 \end{bmatrix} \tag{137}$$

where we used the $\nabla^2 \frac{n}{2}(\overline{\alpha}-\tau)^2 = \frac{1}{n}\mathbf{I}_n$. Thus the matrix (137) is a sum of two terms. By the assumption (134) we have that the second part that contains $\mathbf{H}_t(w,\alpha)$ is positive semi-definite. Next we will show that the first matrix $\mathbf{M}_t(w)$ is symmetric positive definite. Indeed, left and right multiplying the above by $[x,\, a] \in \mathbb{R}^{d+n}$ gives

$$\begin{aligned}
\begin{bmatrix} x & a \end{bmatrix}^\top \mathbf{M}_t(w) \begin{bmatrix} x \\ a \end{bmatrix} &\overset{(137)}{=} \begin{bmatrix} x & a \end{bmatrix}^\top \begin{bmatrix} DF(w)\mathbf{D}_t DF(w)^\top x - DF(w)\mathbf{D}_t a \\ -(DF(w)\mathbf{D}_t)^\top x + a(1+\frac{1}{n}). \end{bmatrix} \\
&= \left\|\mathbf{D}_t^{1/2} DF(w)^\top x\right\|^2 - 2a(DF(w)\mathbf{D}_t)^\top x + (1+\frac{1}{n})\left\|a\right\|^2 \\
&= \left\|\mathbf{D}_t^{1/2}(DF(w)^\top x - a)\right\|^2 - \left\|\mathbf{D}_t^{1/2} a\right\|^2 + (1+\frac{1}{n})\left\|a\right\|^2,
\end{aligned}$$

or in short

$$\begin{bmatrix} x & a \end{bmatrix}^\top \mathbf{M}_t(w) \begin{bmatrix} x \\ a \end{bmatrix} \;=\; \left\|DF(w)^\top x - a\right\|^2_{\mathbf{D}_t} + \left\|a\right\|^2_{(1+\frac{1}{n})\mathbf{I}_n - \mathbf{D}_t} \tag{138}$$

Next we show that (138) is strictly positive for every $(x,a) \neq 0$. To this end, first note that the matrix $(1+\frac{1}{n})\mathbf{I}_n - \mathbf{D}_t$ is positive definite, which follows since the ith diagonal element is positive with

$$\big[(1+\tfrac{1}{n})\mathbf{I}_n - \mathbf{D}_t\big]_{ii} \;=\; 1 + \frac{1}{n} - \frac{1}{\left\|\nabla f_i(w^t)\right\|^2 + 1} \;>\; 0.$$

Consequently if $a \neq 0$ we have that (138) is strictly positive. On the other hand, if $a = 0$ let us prove by contradiction that (138) is still positive for $x \neq 0$. Indeed suppose that $x \neq 0$ and

$$\left\|DF(w)^\top x\right\|^2_{\mathbf{D}_t} = 0 \;\overset{\mathbf{D}_t \succ 0}{\Longrightarrow}\; \sum_{i=1}^n \nabla f_i(w)^\top x = 0.$$

But due to our assumption (60), we have that $DF(w)^\top$ has full column rank, and thus $x = 0$, which is a contradiction. Thus (138) is positive for every $(x,a) \neq 0$ from which we conclude that the Hessian $\nabla^2 h_t(w,\alpha)$ in (137) is positive definite.

■

The proof of Corollary 16 then follows from Lemma 31 by plugging in $\alpha_i^* = f_i(w^*)$ into (136).

Appendix J. Convergence of the Moving Target Stochastic Polyak Stepsize

Here we explore the consequences of Theorems 10 and 11 specialized to Algorithm 2. Throughout this section let

$$\lambda \leq \frac{2n+1}{2n+3} < 1 \tag{139}$$

and let $z^t := (w^t, \alpha^t, \tau_t)$ be the iterates of Algorithm 2 when using a stepsize $\gamma = \gamma_\tau$. Let

$$h_t(z) := \frac{1}{n+1}\left(\sum_{i=1}^{n} \frac{1-\lambda}{2}\frac{(f_i(w)-\alpha_i)^2}{\|\nabla f_i(w^t)\|^2+1} + \frac{n(1-\lambda)}{2}(\overline{\alpha}-\tau)^2 + \frac{\lambda}{2}\tau^2\right). \tag{140}$$

and let w^* be a minimizer of (1) and let

$$\alpha_i^* := f_i(w^*) \quad \text{and} \quad \tau^* = f(w^*), \quad \text{for } i = 1, \dots, n. \tag{141}$$

J.1. Proof of Corollary 18

Corollary 32 *If $\gamma = \gamma_\tau = \frac{1}{2(1-\lambda)(2n+1)}$ and if $h_t(z)$ is star convex along the iterates z^t and around $z^* := (w^*, \alpha^*, \tau^*)$ then*

$$\min_{t=0,\dots,k} \mathbb{E}\left[h_t(z^t) - h_t(z^*)\right] \leq \frac{2(1-\lambda)(2n+1)}{k}\left\|z^0 - z^*\right\|^2 + \frac{\lambda f(w^*)^2}{2(n+1)}. \tag{142}$$

Furthermore, if f_i is $L_{\max}$–Lipschitz then

$$\frac{1}{n+1}\mathbb{E}\left[\sum_{i=1}^{n}\frac{1}{2}\frac{(f_i(w^t)-\alpha_i^t)^2}{L_{\max}+1} + \frac{n}{2}(\overline{\alpha}^t - \tau^t)^2 + \frac{\lambda}{2}\left((\tau^t)^2 - f(w^*)^2\right)\right]$$
$$\leq \frac{2(1-\lambda)(2n+1)}{k}\left\|z^0 - z^*\right\|^2 + \frac{\lambda f(w^*)^2}{2(n+1)}. \tag{143}$$

Proof The proof follows by applying Theorem 10 and Lemmas 9 and 5. Indeed h_t satisfies the growth condition (44) with $G = (1-\lambda)(2n+1)$. By assuming that h_t is star convex along the iterates z^t we have satisfied the two condition required for Theorem 10 to hold, which when substituting in G and $\gamma = \frac{1}{2(1-\lambda)(2n+1)}$ gives

$$\min_{t=1,\dots,k} \mathbb{E}\left[h_t(z^t) - h_t(z^*)\right] \leq \frac{2(1-\lambda)(2n+1)}{k}\mathbb{E}\left[\left\|z^0 - z^*\right\|^2\right] + \frac{1}{k}\sum_{t=1}^{k} h_t(z^*). \tag{144}$$

Furthermore using the bound in Lemma 22 we have that

$$\frac{1}{k}\sum_{t=1}^{k} h_t(z^*) = \frac{\lambda f(w^*)^2}{2(n+1)}$$

and thus (142) holds. Finally, if f_i is $L_{\max}$–Lipschitz, that is if $\|\nabla f_i(w^t)\| \leq L_{\max}$, then using the definition of $h_t(z)$ in (140) we can lower bound $h_t(z^t) - h_t(z^*)$ by the left-hand side of (143). ■

J.2. Proof of Corollary 33

Corollary 33 *If $\gamma = \gamma_\tau = \frac{1}{(1-\lambda)(2n+1)}$ and if $h_t(z)$ is μ–strongly star–convex along the iterates z^t and around $z^* := (w^*, \alpha^*, \tau^*)$ then*

$$\mathbb{E}\left[\left\|z^{t+1} - z^*\right\|^2\right] \leq \left(1 - \frac{\mu}{(1-\lambda)(2n+1)}\right)^{t+1}\left\|z^0 - z^*\right\|^2 + \frac{\lambda f(w^*)^2}{\mu(n+1)}. \tag{145}$$

Proof The proof follows by applying Theorem 11 and Lemmas 9 and 5. Indeed by Lemma 9 h_t satisfies the growth condition (44) with $(1-\lambda)(2n+1)$. By assuming that h_t is μ–strongly star convex along the iterates z^t we have satisfied the two condition required for Theorem 11 to hold. Finally using Lemma 22 we have that

$$h_t(z^*) = \frac{\lambda f(w^*)^2}{2(n+1)}$$

and as a consequence Theorem 11 gives

$$\begin{aligned}
\mathbb{E}\left[\left\|z^{t+1}-z^*\right\|^2\right] &\leq (1-\frac{\mu}{(1-\lambda)(2n+1)})^{t+1}\left\|z^0-z^*\right\|^2 \\
&\quad +\frac{2}{(1-\lambda)(2n+1)}\sum_{i=0}^{t}(1-\gamma\mu)^i\frac{\lambda f(w^*)^2}{2(n+1)}. \\
&\leq (1-\frac{\mu}{(1-\lambda)(2n+1)})^{t+1}\left\|z^0-z^*\right\|^2 \\
&\quad +\frac{2}{(1-\lambda)(2n+1)}\frac{1}{\gamma\mu}\frac{\lambda f(w^*)^2}{2(n+1)}. \\
&= (1-\frac{\mu}{(1-\lambda)(2n+1)})^{t+1}\left\|z^0-z^*\right\|^2+\frac{\lambda f(w^*)^2}{\mu(n+1)}, \qquad (146)
\end{aligned}$$

where in the last equality we used that $\gamma = \frac{1}{(1-\lambda)(2n+1)}$. ■

J.3. Proof of Theorem 34

Theorem 34 *Let $h_t(z)$ be μ-strongly star–convex along the iterates z^t and around $z^* := (w^*, \alpha^*, \tau^*)$. Let $\epsilon > 0$. If we use an iteration dependent stepsize in Algorithm 2 given by*

$$\gamma_t = \begin{cases} \dfrac{1}{(1-\lambda)(2n+1)} & \text{if } t \leq 2(2n+1)\left\lceil\frac{1-\lambda}{\mu}\right\rceil \\ \dfrac{(t+1)^2-t^2}{\mu(t+1)^2} & \text{if } t \geq 2(2n+1)\left\lceil\frac{1-\lambda}{\mu}\right\rceil \end{cases} \qquad (147)$$

and if

$$\lambda \leq \min\left\{1-\frac{2\mu}{2n+1}, \frac{2n+1}{2n+3}\right\}.$$

then

$$\mathbb{E}\left[\left\|z^t-z^*\right\|^2\right] \leq \frac{(1-\lambda)\lambda f(w^*)^2}{\mu^2}\frac{16}{t}+\frac{4(2n+1)^2}{e^2t^2}\left\lceil\frac{1-\lambda}{\mu}\right\rceil^2\left\|z^0-z^*\right\|^2. \qquad (148)$$

Proof

Following the proof of Theorem 11 upto (113), we have that for $\gamma \leq \frac{1}{G} = \frac{1}{(1-\lambda)(2n+1)}$ and $h_t(z^*) = \frac{\lambda f(w^*)^2}{2(n+1)}$ that

$$\begin{aligned}
\mathbb{E}_t\left[\left\|z^{t+1}-z^*\right\|^2\right] &\leq (1-\gamma\mu)\left\|z^t-z^*\right\|^2+2\gamma^2(1-\lambda)(2n+1)\frac{\lambda f(w^*)^2}{2(n+1)} \\
&\leq (1-\gamma\mu)\left\|z^t-z^*\right\|^2+4\gamma^2(1-\lambda)\lambda f(w^*)^2. \qquad (149)
\end{aligned}$$

Taking expectation and using the abbreviations

$$r_t \coloneqq \mathbb{E}\left[\left\|z^t - z^*\right\|^2\right] \quad \text{and} \quad \sigma^2 \coloneqq 2(1-\lambda)\lambda f(w^*)^2, \tag{150}$$

gives that

$$r^{t+1} \leq (1-\gamma\mu)r^t + 2\gamma^2\sigma^2. \tag{151}$$

With this notation, this is now identical to the setting of Theorem 3.2 in [14]. Using the notation of Theorem 3.2 in [14] we have that $2\mathcal{L} = (1-\lambda)(2n+1)$ and consequently $\mathcal{K} = \frac{\mathcal{L}}{\mu} = \frac{1}{2}(2n+1)\left\lceil\frac{1-\lambda}{\mu}\right\rceil$. As a result of Theorem 3.2 in [14] we have that

$$r^t \leq \frac{\sigma^2}{\mu^2}\frac{8}{t} + \frac{16\lceil\mathcal{K}\rceil^2}{e^2t^2}r^0. \tag{152}$$

Substituting back the definitions given in (150) gives (148). Though one detail in the proof of Theorem 3.2 in [14] is that $\mathcal{K} \geq 1$, which in our case holds it

$$\lambda \leq 1 - \frac{2\mu}{2n+1}.$$

■

Appendix K. Convex Classification: Additional Experiments

For our experiments on convex classification tasks, we focused on logistic regression. That is

$$f(w) = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i^\top w) + \frac{\sigma}{2}\|w\|_2^2 \tag{153}$$

where $\phi_i(t) = \ln\left(1+e^{-y_i t}\right)$, $(x_i, y_i) \in \mathbb{R}^{d+1}$ are the features and labels for $i = 1, \ldots, n$, and $\sigma > 0$ is the regularization parameter. We experimented with the five diverse data sets: leu [11] , duke [34], colon-cancer [1], mushrooms [9] and phishing [9]. Details of these datasets and their properties can be found in Table 1.

				$\sigma = 0$			$\sigma = \min_{i=1,\ldots,n}\|x_i\|^2/n$			
dataset	d	n	$L_{\max}$	γ^*	γ_τ^*	f^*	γ^*	γ_τ^*	f^*	σ
leu	7130	38	824.6	1.1	10^{-5}	0.0	0.01	0.4	0.449	11.74
duke	7130	44	683.2	1.1	10^{-3}	0.0	0.1	0.4	0.4495	5.06
colon-cancer	2001	62	137.8	1.1	10^{-5}	0.0	0.1	0.9	0.453	2.66
mushrooms	112	8124	5.5	1.1	10^{-4}	0.0	1.0	0.01	0.083	0.0027
phishing	68	11055	7.75	0.01	0.5	0.142	0.01	0.9	0.188	0.0028

Table 1: Binary datasets used in the logistic regression experiments together with the best parameters settings for γ and γ_τ for two different regularization settings.

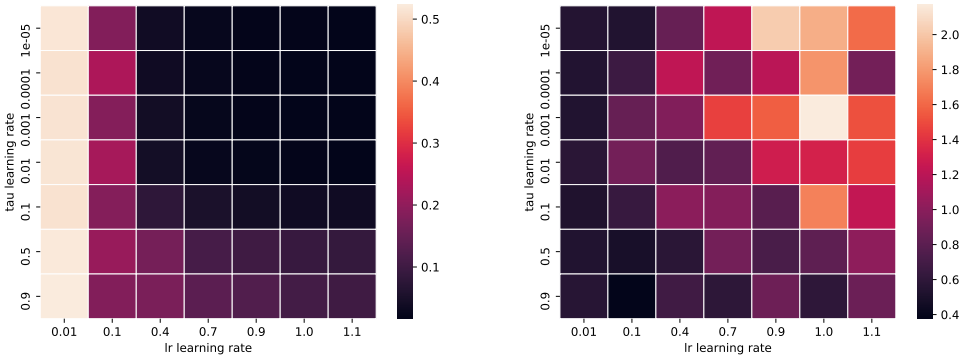


Figure 7: `colon-cancer` $(n, d) = (62, 2001)$ Left: $\sigma = 0.0$. Right: $\sigma = \min_{i=1,\dots,n} \|x_i\|^2 / n = 2.66$.

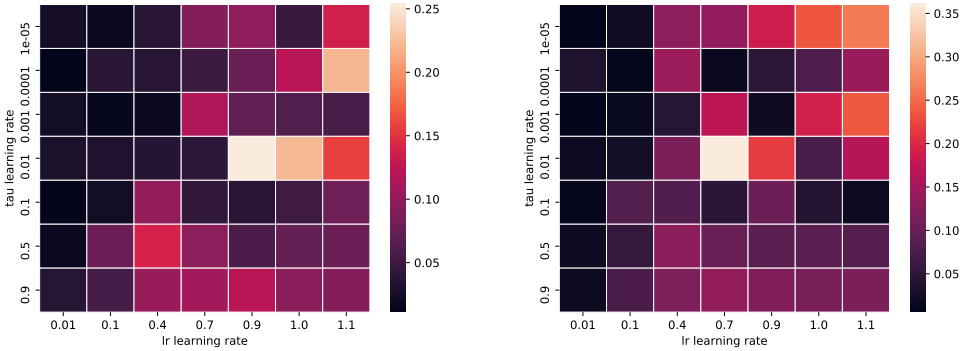


Figure 8: Logistic Regression with data set phishing $(n, d) = (11055, 68)$ and regularization. Left: $\sigma = 0.0$ and Right: $\sigma = \min_{i=1,\dots,n} \|x\|_i^2 / n$.

K.1. Grid search and Parameter Sensitivity

To investigate how sensitve `MOTAPS` is to setting its two parameters $\gamma \in [0, 1]$ and $\gamma_\tau \in [0, 1]$ we did a parameter sweep. We searched over the grid given by

$$\gamma \in \{0.01, 0.1, 0.4, 0.7, 0.9, 1.0, 1.1\}$$

and

$$\gamma_\tau \in \{0.00001, 0.0001, 0.001, 0.01, 0.1, 0.5, 0.9\}$$

and ran `MOTAPS` for 50 epochs over the data, and recorded the resulting norm of the gradient. See Figures 3, 8, 7 and 4 2 for the results of the grid search on the datasets `mushrooms`, `phishing`, `colon-cancer` and `duke` respectively. In Table 1 we resume the results of the parameter search, together with the details of each data set.

Ultimately the determining factor for finding the best parameter was the magnitude of the optimal value $f(w^*)$. Since this quantity is unknown to us a priori, we used the size of the regularization parameter as an proxy. Based on these parameter results we devised the following rule-of-thumb for setting γ and γ_τ with

$$\gamma = 1.0/(1 + 0.25\sigma e^\sigma) \quad \text{and} \quad \gamma_\tau = 1 - \gamma \tag{154}$$

where σ is regularization parameter.

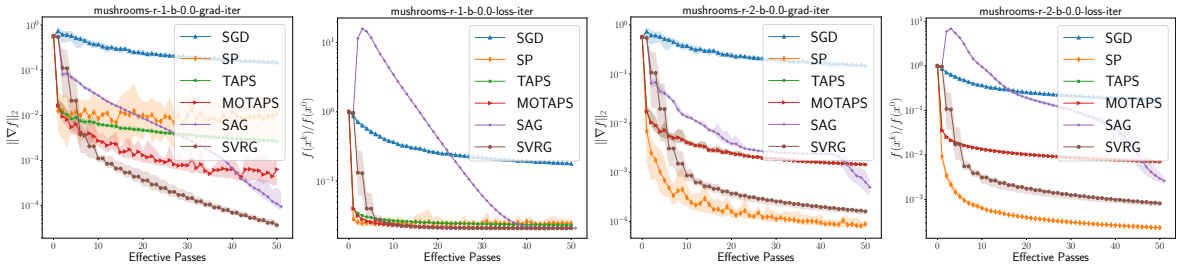


Figure 9: Logistic Regression with data set mushrooms $(n, d) = (8124, 112)$. Left: $\sigma = 1/n$ and Right: $\sigma = 1/n^2$. See [6]

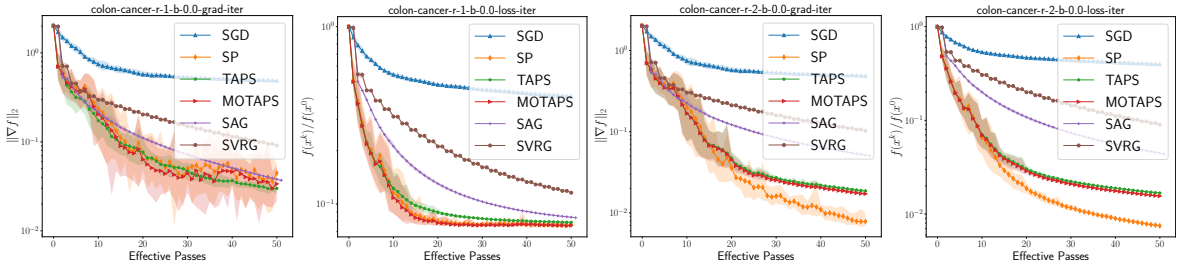


Figure 10: Logistic Regression with data set colon-cancer $(n, d) = (62, 2001)$ and regularization. Left: $\sigma = 1/n$ and Right: $\sigma = 1/n^2$. See [6]

K.2. Comparing to Variance Reduced Gradient Methods

In Figures 9, 10, 11 and 12 we present further comparisons between SP, TAPS and MOTAPS against SGD, SAG and SVRG. We found that the variance reduced gradients methods were able to better exploit strong convexity, in particular for problems with a large regularization, and problems that were under-parameterized, with the phishing problem in Figure 12 being the most striking example. For problems with moderate regularization, and that were over-parametrized, the MOTAPS performed the best. See for example the left of Figure 10 and Figure 11. When the regularization is very small, and the problem is over-parameterized, thus making interpolation much more likely, the SP converged the fastest. See for example the right of Figure 10 and 9.

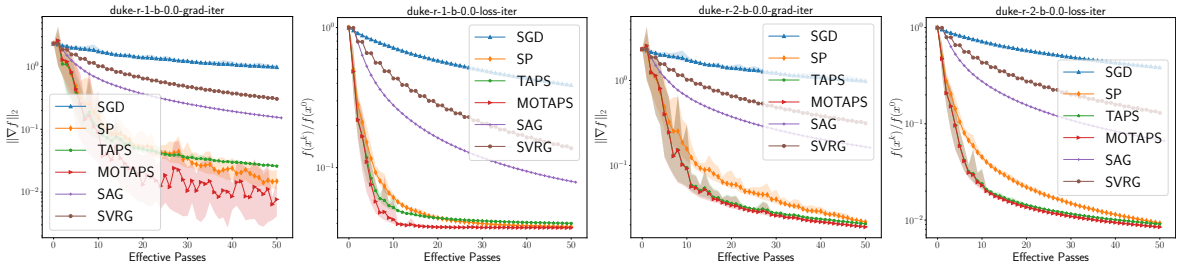


Figure 11: Logistic Regression with data set duke $(n, d) = (44, 7130)$. Left: $\sigma = 1/n$ and Right: $\sigma = 1/n^2$.

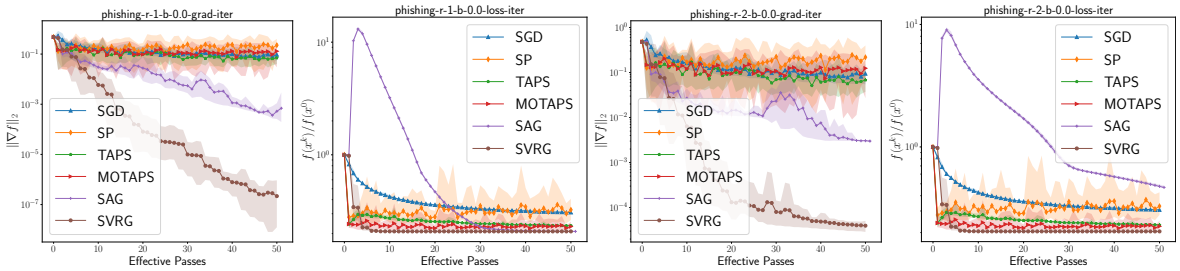


Figure 12: Logistic Regression with data set phishing $(n, d) = (11055, 68)$. Left: $\sigma = 1/n$ and Right: $\sigma = 1/n^2$.

K.3. Momentum variants

We also found that adding momentum to SP and MOTAPS could speed up the methods. To add momentum we used the iterate averaging viewpoint of momentum [31] where by we replace the updates in w^t by a weighted average over past iterates. For TAPS and MOTAPS this is equivalent to introducing a sequences of z^t variables and updating according to

$$z^{t+1} = z^t - \eta \frac{f_i(w^t) - \alpha_i^t}{\|\nabla f_i(w^t)\|^2 + 1} \nabla f_i(w^t)$$
$$w^{t+1} = \beta w^t + (1-\beta) z^{t+1}$$

where $\eta = \gamma \left(1 + \frac{\beta}{1-\beta}\right)$ is the adjusted stepsize [7]. See Figures 13, 14 and 15 for the results of our experiments with momentum as compared to ADAM [20]. We found that in regimes of moderate regularization ($\sigma = 1/n$) the MOTAPS method was the fastest among all method, even faster than TAPS despite not having access to f^*, see the left side of Figures 13, 14 and 15. Yep when using moderate regularization, adding on momentum gave no benefit to SP, TAPS, and MOTAPS. Quite the opposite, for momentum $\beta = 0.5$, we see that MOTAPSM-0.5, which is the MOTAPS method with momentum and $\beta = 0.5$, hurt the convergence rate of the method.

In the regime of small regularization $\sigma = \frac{1}{n^2}$, we found that momentum sped up the convergence of our methods, see the right of Figures 13, 14 and 15. On the under-parameterized problem mushrooms, the gains from momentum were marginal, and the ADAM method was the fastest overall, see the right of Figure 13. On the over-parametrized problem colon-cancer, adding momentum to SP gave a significant boost in convergence speed, see the right of Figure 14. Finally on the most over-parametrized problem duke, adding momentum offered a significant speed-up for MOTAPS, but still the ADAM method was the fastest, see the right of Figure 15.

Appendix L. Deep learning experimental setup details

In this section we detail the specific implementation choices for each environment. Across all environments, minibatching was accomplished by treating each minibatch as a single data-point. Since per-datapoint values are tracked across epochs, our training setup used minibatches which contain the same set of points each epoch.

7. See Proposition 1.6 in [31] for the details of form of momentum and parameter settings.

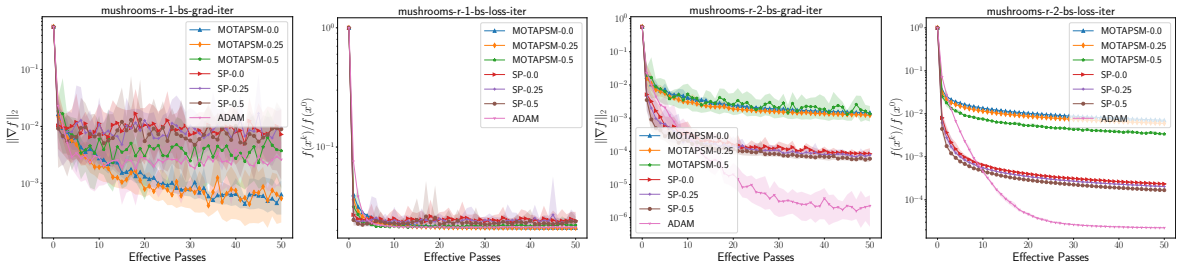


Figure 13: Experiments on momentum with mushrooms $(n, d) = (8124, 112)$. Left: $\sigma = 1/n$ and Right: $\sigma = 1/n^2$.

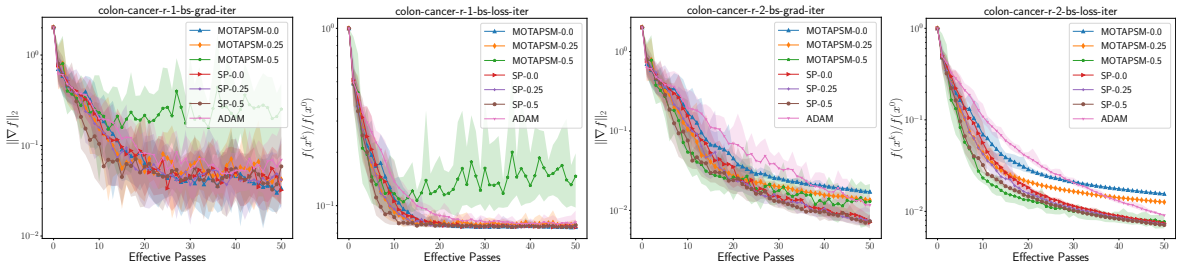


Figure 14: Experiments on momentum with colon-cancer $(n, d) = (62, 2001)$ and regularization. Left: $\sigma = 1/n$ and Right: $\sigma = 1/n^2$.

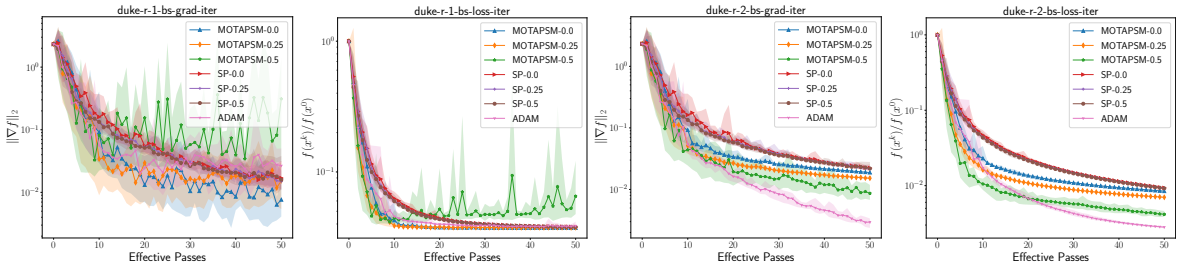


Figure 15: Experiments on momentum with duke $(n, d) = (44, 7130)$. Left: $\sigma = 1/n$ and Right: $\sigma = 1/n^2$.

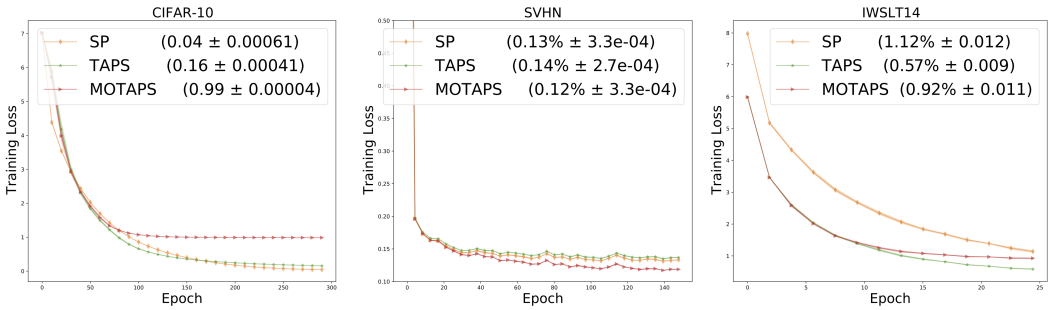


Figure 16: Deep learning experiments training loss

L.1. CIFAR10

We trained for 300 epochs using batch size 256 on 1 GPU. Momentum 0.9 was used for all methods. The pre-activation ResNet used has 58,144,842 parameters. Following standard practice we apply data augmentation of the training data; horizontal flipping, 4 pixel padding followed by random cropping to 32x32 square images.

L.2. SVHN

We trained for 150 epochs on a single GPU, using a batch size of 128. Momentum 0.9 was used for each method. Data augmentations were the same as for our CIFAR10 experiments. The ResNet-1-16 network has a total of 78,042 parameters, and uses the classical, non-preactivation structure.

L.3. IWSLT14

We used a very simple preprocessing pipeline, consisting of the Spacy de_core_news_sm/en_core_web_sm tokenizers and filtering out of sentences longer than 100 tokens to fit without our GPU memory constraints. Training used batch-size 32, across 1 GPU for 25 epochs. Other hyper-parameters include momentum of 0.9, weight decay of 5e-6, and a linear learning rate warmup over the first 5 epochs