DE GRUYTER Journal of Inverse and Ill-Posed Problems

On methods with preconditioning and weight decaying

Journal:	*Journal of Inverse and Ill-Posed Problems*
Manuscript ID	JIIP.2023.0080
Manuscript Type:	Original Research Article
Date Submitted by the Author:	31-Oct-2023
Complete List of Authors:	Kreinin Matvei; MIPT, DCAM
MSC-Classification -- go to www.ams.org/msc to find your classifications:	90C26
Keywords:	Adam, AdamW, OASIS, Regularization, Weight Decay, Optimization

De Gruyter Journal YYYY; aop

Research Article

Kreinin Matvei, Babkin Petr, Statkevich Ekaterina, Beznosikov Aleksandr, and Gasnokiv Alexander

On methods with preconditioning and weight decaying

https://doi.org/10.1515/sample-YYYY-XXXX
Received Month DD, YYYY; revised Month DD, YYYY; accepted Month DD, YYYY

Abstract: This paper investigates the convergence behavior of optimization methods with preconditioning that utilize weight decay regularization, specifically focusing on popular variants such as AdamW and OASIS. We explore different alternatives to these method, with the goal of investigating their convergence speed and accuracy. Also we conduct experiments on benchmark datasets and models in order to compare them on practice. Overall, our study provides insights into the design of regularization techniques methods with preconditioning.

Keywords: Adam, AdamW, OASIS, Regularization, Weight Decay, Optimization

1 Introduction

A huge part of machine learning is based on the unconstrained optimization problem

$$\min_{w \in \mathbb{R}^d} f(w). \tag{1}$$

Problems of the form (1) cover a plethora of applications, including empirical risk minimization [1], deep learning [2], and supervised learning [3] tasks such as regularized least squares [4] or logistic regression [5].

The classic base method for solving the optimization problem (1) is a gradient descent

$$w_{t+1} = w_t - \eta \nabla f(w_t), \tag{2}$$

But the minimization problem (1) can be difficult to solve, particularly when the number of training samples or problem dimension is large. In such cases, evaluating the full gradient on every iteration in the context of gradient descent becomes prohibitively expensive, especially considering that gradient descent often requires numerous iterations to converge. In modern machine learning, especially with the advent of deep learning, there is a growing interest in tackling large and more complex problems. For such cases stochastic gradient descent [6] became popular solution.

Over time, optimization methods have continuously advanced and evolved, becoming more sophisticated and intricate. One prominent aspect of optimization is the efficient selection of the step size during the iterative process. Adaptive methods with gradient scaling dynamically adjust the step size based on the gradient information. This adaptive behavior, applied to each variable, improves the optimization process by efficiently navigating complex landscapes and ensuring optimal progress for each variable [7]. In particular, these methods have gained significant popularity in the field of machine learning, where high-dimensional problems are prevalent [8, 9].

Kreinin Matvei, Babkin Petr, Statkevich Ekaterina, Moscow Institute of Physics and Technology, Phystech School of Applied Mathematics and Computer Science, Moscow, Russia, e-mail: kreinin.mv@phystech.edu, babkin.pk@phystech.edu, statkevichk@bk.ru
Beznosikov Aleksandr, Gasnokiv Alexander, Moscow Institute of Physics and Technology, Phystech School of Applied Mathematics and Computer Science, Moscow, Russia, e-mail: beznosikov.an@phystech.edu, gasnikov@yandex.ru

In more details, methods with scaled gradient refer to techniques that involve preconditioning the gradient of a problem by a specific matrix D_t, which enables the gradient to take into account the geometry of the problem. Generally, the step of preconditioned algorithms can be expressed as a following modernization of step (2):

$$w_{t+1} = w_t - \eta \cdot D_t^{-1} g_t, \tag{3}$$

where g_t is an unbiased stochastic gradient. The idea of using the scaling matrix refers to Newton's method, where $D_t = \nabla^2 f(w)$. However calculating and reversing hessian poses significant challenges, thus necessitating the utilization of certain heuristics as replacements. Such techniques are exemplified in Adagrad [10], Adam [11], RMSProp, OASIS [12] and so forth, where the computation strategies for D_t do not necessitate the evaluation of hessian. For example, Adagrad presents preconditioning in the form:

$$D_t = \operatorname{diag}\left\{ \sqrt{\sum_{t'=0}^{t} g_{t'} \odot g_{t'}} \right\},$$

where in fact we use only stochastic gradients. Here and further we use Hadamard product $\odot$. RMSProp and Adam use similar idea:

$$D_t = \operatorname{diag}\left\{ \sqrt{m_t} \right\},$$

where m_t is the second momentum [11]. OASIS uses another approach:

$$D_t = \operatorname{diag}\left\{ z_k \odot \nabla^2 f z_k \right\},$$

where the hessian matrix appears to be employed, it is actually approximated through scalar function differentiation. Here z_k is a random vector of Randamaher distribution [12].

Despite the advantages of preconditioning methods, they are prone to overfitting, thus necessitating their combined application with regularization. This approach has been widely applied to various machine learning problems, including image classification [13], speech recognition [14], and natural language processing [15], and has shown its effectiveness in improving the generalization capability of neural networks [16]. With regularization, problem (1) is reformulated as

$$\min_{w \in \mathbb{R}^d} F(w) := f(w) + r(w), \tag{4}$$

where r is a regularizer function.

In methods with preconditioning appears to be several ways to include regularization. We can include regularizer r in g_t calculation thus it will be taken into consideration while calculating D_t. This method is equal to considering optimization problem (4). Or we can include regularizer only on last step, decreasing norm of w [17]. This way of regularization is called weight decay and surprisingly turns out to be more efficient in practical problems. There is another ways of considering regularizer which will be discussed further in the paper.

Despite their practical efficiency, methods incorporating weight decay show relatively limited theoretical analysis. As a result, a number of research inquiries arise: *Do methods utilizing weight decay as a preconditioning technique converge? If so, what is the rate of convergence? To what solution do they converge?*

1.1 Related works

Stochastic methods have an extensive theory of convergence, whereas methods involving preconditioning are relatively new and have a limited history of investigations. In one of the pioneering papers on preconditioning [10], the authors conducted a meticulous analysis of Adagrad's convergence theory. However, in more recent papers such as those discussing RMSProp [18] or Adam [11], little attention is given to theoretical aspects or the existing theory may contain inaccuracies. Over time, mistakes have been rectified, leading to the development of comparatively robust convergence theories for preconditioning methods [19, 20]. In another study, Loshchilov and Hutter [17] investigated the properties of the Adam and AdamW algorithms in terms of hyperparameters and also explored restart techniques. Zhang et al. [21] delved into the lookahead mechanism of Adam. More recently, convergence theories have been established for modern methods such as OASIS [12, 22]. Additionally, a theory addressing time-varying preconditioning has also emerged [23]. Nevertheless, numerous questions in this field remain unanswered. A few of these inquiries are formulated at the end of the preceding paragraph and examined in our paper.

1.2 Contributions

In general, our paper provides insight into comparison of different consideration ways of regularization is methods with preconditioning. Here, we provide a brief summary of our main contributions:

– **Novel approach to regularization in preconditioning.** We propose a novel method of integrating regularization into preconditioning algorithms, where we utilize identical step sizes for the objective function and the regularizer. At the same time regularizer does not impact the computation of D_t.

– **Proof of preconditioned methods with weight decay convergence.** In this discourse, we elucidate the convergence properties of preconditioned optimization algorithm with regularization, particularly when supplemented with weight decay mechanisms. Our non-convex theoretical exposition revolves around the foundational assumptions encompassing the smoothness of the function (3). We achieve different convergence speed assessments based on the PL-condition (4) and conducte a stochastic analysis based on assumption (6). Through meticulous analysis, we aim to furnish robust guarantees affirming the convergent behavior of such preconditioned methods under the specified conditions.

– **Investigation of loss function behavior.** We conducted a comparative study on the convergence rate of the Adam and Adam with weight decay (AdamW). We observed that AdamW converges towards a distinct target function. Subsequently, we conducted an analysis of these target function.

2 Main part

2.1 Weight decay and how to use it with preconditioning

As it was mentioned above, in preconditioned methods, there exist several techniques for incorporating regularization into the optimization process. In this study, we consider three different approaches, which are illustrated in Algorithm 1 using different colors. In general, these methods can be characterized by the stage in which the regularization term is included into the optimization process.

Algorithm 1 Different ways of using preconditioning for regularized problem

```
Require: $\eta$ − learning rate, $f$ − objective function
  while $w$ not converged do
     $t = t + 1$
     $g_t \leftarrow$ stochastic gradient of $f$
     $g_t \leftarrow g_t + \nabla r(w_t)$                                  standart regularization
     $D_t \leftarrow$ preconditioning matrix, based on $g_t$
     $w_t \leftarrow w_{t-1} - \eta \cdot D_t^{-1} g_t$                         standart regularization,
     $w_t \leftarrow w_{t-1} - \eta \cdot D_t^{-1} (g_t + \nabla r(w_t))$             scaled weight decay,
     $w_t \leftarrow w_{t-1} - \eta \cdot D_t^{-1} g_t - \eta \cdot \nabla r(w_t)$           weight decay,
  end while
```

To be more specific, the first regularization technique illustrated in blue involves simply adding the regularization term to the objective function. This regularizer is included in the pseudo-gradient and factored into the calculation of D_t. In essence, this approach involves applying the basic preconditioning method to the problem (4) without using weight decay. The second regularization technique, shown in orange, is a novel approach. Although regularizer term does not affect its computation of preconditioner D_t, it is added before applying D_t. That means that learning rate is adopted in the same way for gradient and regularizer. The last regularization approach we consider is known as weight decay, illustrated by the color red in the Algorithm 1. Similar to the second method, matrix D_t is calculated without using regularizer and this method incorporates the regularizer during the algorithmic step, avoiding interference of regularization with the preconditioning stage.

Overall, it is important to carefully consider the impact of regularization when designing optimization algorithms, and we hope that our investigation of this techniques will prove useful to researchers in the field.

2.2 Convergence speed of preconditioning methods with weight decay

We set ourselves a goal to estimate a convergence speed of methods with preconditioning with weight decay regularization. Although step of methods with weight decay seems simple, it can be viewed in a rather unexpected way. We can put D_t^{-1} out of brackets which gives

$$w_{t+1} = w_t - \eta \cdot D_t^{-1}(\nabla f(w_t) + D_t \nabla r(w_t)). \tag{5}$$

That suggests the need to introduce a function $\tilde{r}$ such that $\nabla \tilde{r}(w) = D_t \nabla r(w)$ and new target function

$$\tilde{F}(w) := f(w) + \tilde{r}(w), \tag{6}$$

where new target function $\widetilde{F}$ changes at every time-step, because D_t changes at every time-step.

New adaptive regularizer $\widetilde{r}$ does not exist in the general case. Therefor we impose restrictions on initial regularizer and preconditioner structure. That is framed in the following assumptions 1, 2.

Assumption 1. *(Regularizer structure) Regularizer r can be viewed in the following form:*

$$r(w) = \sum_{i=1}^{d} r_i(w_i)$$

Assumption 2. *(Preconditioner structure) Preconditioner D_t can be viewed in the following form:*

$$D_t = diag\left\{D_t^1 \ldots, D_t^d\right\}$$

Although these assumptions are strict, they hold for every mentioned method with preconditioning and applicable regularizers. With this two assumptions we are able to prove existance of $\widetilde{r}$ and, consequently, $\widetilde{F}$. We frame that in the following Lemma 1. We only show existance, but not uniqueness of the function, but in our evaluations, $\widetilde{F}$ can be found up to a constant.

Lemma 1. *(Existence of $\widetilde{r}$) Suppose the Assumptions 1, 2 hold, the function $\widetilde{r}$ exists and has following form:*

$$\widetilde{r}(w) = \sum_{i=1}^{d} D_t^i r_i(w_i)$$

Using introduced Assumption 3 we can guarantee smoothness for $\widetilde{r}$ and estimate its Lipschitz constant, which is formaly framed and proved in Lemmas 2.

Lemma 2. *(L-smoothness of $\widetilde{r}$) Suppose the Assumptions 1, 2, 3 hold, The gradient of $\widetilde{r}$ is $L_{\tilde{r}}$-continuous, i.e. there exists a constant $L_{\tilde{r}} > 0$ such that $\forall x, y \in \mathbb{R}^d$,*

$$\widetilde{r}(x) \leq \widetilde{r}(y) + \langle \nabla \widetilde{r}(y), x - y \rangle + \frac{L_{\tilde{r}}}{2} ||x - y||^2,$$

where $L_{\tilde{r}} = ||D_t|| L_r$

The convergence speed is typically measured in terms of the number of iterations required to reach a certain level of error. To obtain estimates on the number of iterations required to converge to a given error, we must impose certain assumptions on the function.

Throughout theoretical analysis we assume that $f : \mathbb{R}^d \to \mathbb{R}$ is $L-$smooth and twice differentiable. Additionally we imply a PL-condition to make another evaluation concerning speed of convergence.

Assumption 3. *(L-smoothness)*

- *The gradients of f are L_f-Lipschitz continuous $\forall w \in \mathbb{R}^d$, i.e. there exists a constant $L_f > 0$ such that $\forall x, y \in \mathbb{R}^d$,*

$$f(x) \leq f(y) + \langle \nabla f(y), x - y \rangle + \frac{L_f}{2} \|x - y\|^2$$

- *The gradient of r is L_r-Lipschitz continuous $\forall w \in \mathbb{R}^d$, i.e. there exists a constant $L_r > 0$ such that $\forall x, y \in \mathbb{R}^d$,*

$$r(x) \leq r(y) + \langle \nabla r(y), x - y \rangle + \frac{L_r}{2} ||x - y||^2$$

Assumption 4. *(PL–condition) There exists $\mu > 0$, such that $\forall w \in \mathbb{R}^d$*

$$||\nabla f(w)|| \geq 2\mu(f(w) - f^*)$$

We use popular restriction on the preconditioner, which is framed in Assumption 5.

Assumption 5. *(Preconditioner) Restrictions on preconditioner D_t*

$$\alpha I \preccurlyeq D_t \preccurlyeq \Gamma I \Leftrightarrow \frac{I}{\alpha} \preccurlyeq D_t^{-1} \preccurlyeq \frac{I}{\Gamma} \tag{7}$$

It has been proven that this assumption holds for all modern algorithms with preconditioning like Adam, Adagrad, OASIS [23].

In order to conduct a stochastic analysis we must include restrictions on the stochastic gradient of the function. This is formalized in the following assumption 6.

Assumption 6. *(Expectations) Restrictions on D_t and g_t are unbiased, i.e.*

$$\mathbb{E}\left[D_t\right] = D_t \text{ and } \mathbb{E}\left[g_t\right] = \nabla f(w_t), \mathbb{E}\left[||g_t - \nabla f||^2\right] \leq \sigma^2 \tag{8}$$

Using introduced assumptions we proved convergence of methods with preconditioning with weight decay regularization in general form. Our results are framed in Theorem 1 and Theorem 2. Proofs of the theorems can be found in Appendix 5.

Theorem 1. *Suppose the Assumptions 3, 5 hold, let $\varepsilon > 0$ and let the step-size satisfy*

$$\eta < \frac{2\alpha}{L_f + \Gamma L_{\tilde{r}} \alpha}$$

Then, the number of iterations performed by AdamW algorithm, starting from an initial point $w_0 \in \mathbb{R}^d$ with $\Delta_0 = \tilde{F}(w_0) - \tilde{F}^*$, required to obtain and ε-approximate solution of the convex problem (1) can be bounded by

$$T = \mathcal{O}\left(\frac{2\Delta_0 \Gamma \alpha}{\left(2\alpha - \left(L_f + \Gamma L_{\tilde{r}} \alpha\right)\eta\right)\eta\varepsilon} \right)$$

Theorem 2. *Suppose the Assumptions 3, 4, 5 hold, let $\varepsilon > 0$ and let the step-size satisfy*

$$\eta \leq \frac{2\alpha}{L_{\widetilde{F}}}$$

Let $\tilde{F}^*$ be a solution of the optimization function. Then, the number of iterations performed by AdamW algorithm, starting from an initial point $w_0 \in \mathbb{R}^d$ with $\Delta_0 = \tilde{F}(w_0) - \tilde{F}^*$, required to obtain and ε-approximate solution of the convex problem (1) can be bounded by

$$T = \mathcal{O}\left(\frac{\ln \frac{\Delta_0}{\epsilon}}{2\mu\eta^2 \left(\frac{1}{\eta} - \frac{L_{\tilde{F}}}{2\alpha}\right)} \right)$$

Theorem 3. *Suppose the Assumptions 3, 4, 5, 6 hold, let $\varepsilon > 0$ and let the step-size satisfy*

$$\eta \approx \sqrt{\frac{\left(\tilde{F}(w_0) - \tilde{F}(w_*)\right)\alpha}{L\sigma^2}}$$

Let $\tilde{F}^*$ be a solution of the optimization function. Then, the number of iterations performed by AdamW algorithm, starting from an initial point $w_0 \in \mathbb{R}^d$ with $\Delta_0 = \tilde{F}(w_0) - \tilde{F}^*$, required to obtain and ε-approximate solution of the convex problem (1) can be bounded by

$$T = \mathcal{O}\left(\frac{\Gamma\Delta_0}{\left(\frac{1}{\eta} - \frac{\Gamma L_{\tilde{r}}}{2} - \frac{\Gamma L_{\tilde{F}} L_f \eta^2}{2\alpha^2}\right)\varepsilon}\right)$$

These theorems establish the convergence of methods with preconditioning and weight decay under different assumptions, and further delineate their iterative rates. For our objectives, the simple act of convergence by these methods holds profound importance. However, characteristics of the solution $\widetilde{w}^*$ of a problem

$$\min_{w \in \mathbb{R}^d} \tilde{F}(w) = f(w) + \tilde{r}(w), \tag{9}$$

to which this method converge demand a deeper exploration, which will be elucidated in the subsequent section.

2.3 Solution of preconditioning methods with weight decay

In the previous subsection we have proved convergence of preconditioned methods, however we have pointed out above that methods with weight decay does not converge to the initial optimization problem (4) solution w^*, but rather to a new solution $\widetilde{w}^*$ of a problem (9). This observation is evidenced by the experiments detailed in Section 3.

Let us estimate the difference between solutions of problems (4) and (9). This restriction is based on Assumptions (3) and properties of matrix D_t.

Lemma 3. *(Lower bound) Suppose the Assumptions 1, 2 and 3 holds, then difference between a solution w^* of a problem (4) and a solution $\widetilde{w}^*$ of a problem (9) can be bounded below.*

$$\|\widetilde{w}^* - w^*\| L_F \geq \|\nabla r(\widetilde{w}^*)(I - D_t)\|$$

Consequently, it can be observed that employing techniques such as preconditioning and weight decay yields convergence towards an alternate solution. The discrepancy between these solutions is contingent upon the norm of the difference between D_t and the identity matrix ($||D_t - I||$). As a result, an analysis of the distribution of the elements of $D := \lim_{t \to \infty} D_t$ can provide insights into the convergence behavior of the weight decay method.

3 Experiments

We will consider two algorithms OASIS [12] and Adam [11], and its variations. Their main difference is in the calculation of the pseudo hessian. In Adam, the Hessian is a diagonal matrix consisting of squares of derivatives, in OASIS we have a stochastic Hessian, which is calculated through a random variable from the Randemacher distribution. We framed three methods of regularization for Adam and OASIS in Algorithm 2 and Algorithm 3 respectively.

Algorithm 2 Different ways of using Adam for regularized problem

```
Require: $\eta, \beta_1, \beta_2, \epsilon, f, r$
  while $\theta$ not converged do
    $t = t + 1$
    $g_t = \nabla f(w_{t-1}) + \nabla r(w_{t-1})$                                  AdamL2
    $m_t = \beta_1 \cdot m_{t-1} + (1 - \beta_1) \cdot g_t$
    $v_t = \beta_2 \cdot v_{t-1} + (1 - \beta_2) \cdot g_t^2$
    $\hat{m}_t = \frac{m_t}{1-\beta_1^t} + \nabla r(w_{t-1})$                               AdamWH
    $\hat{v}_t = \frac{v_t}{1-\beta_2^t}$
    $w_t = w_{t-1} - \eta \cdot \frac{\hat{m}_t}{\sqrt{v_t}+\epsilon} - \eta \nabla r(w_{t-1})$              AdamW
  end while
```

Algorithm 3 Different ways of using OASIS for regularized problem

```
Require: $w_0, \eta_0, D_0, \theta_0 = +\infty$
  $w_1 = w_0 - \eta \hat{D_0}^{-1} \nabla f(w_0)$
  for $k = 1, 2, ...$ do
    $g_k = \nabla f(w_k) + \nabla r(w_{t-1})$                                  OASISL2
    $D_k = \beta D_{k-1} + (1 - \beta_2) \cdot diag\left(z_k \odot \nabla^2 \left(f(w_k) + r(w_k)\right) z_k\right)$      OASISWH
    $(\hat{D_k})_{ii} = max\{|D_k|_{i,i}; \alpha\}, \forall i = \overline{1,d}$
    $\eta_k = min\{\sqrt{1 + \theta_{k-1}} \cdot \eta_{k-1}; \frac{||w_k - w_{k-1}||_{\hat{D_k}}}{2||\nabla f(w_k) - \nabla f(w_{k-1})||^*_{\hat{D_k}}}\}$
    $w_{k+1} = w_k - \eta_k g_k D_k^{-1} - \eta \nabla r(w_{t-1})$                      OASISW
    $\theta_k = \frac{\eta_k}{\eta_{k-1}}$
  end for
```

The results of our computational experiments are presented in Figures 1 and 2. It can be observed that AdamW does not converge according to the basic criterion $||\nabla F(w_t)|| = \nabla f(w_t) + D_t \nabla r(w_t)||$, but it converges only by the special criterion $||\nabla \widetilde{F}(w_t) = ||\nabla f(w_t) + D_t \nabla r(w_t)||$. Furthermore, it is noted that AdamW exhibits better convergence according to its own criterion.

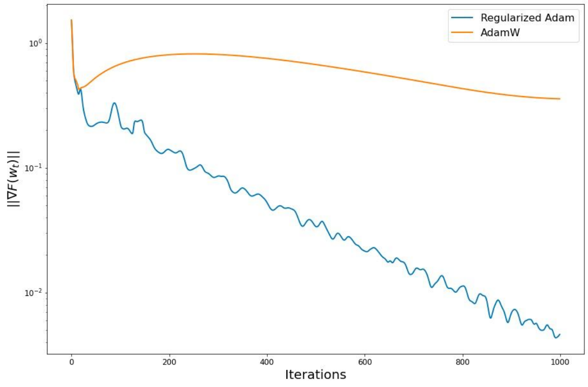


Fig. 1: Adam and AdamW with basic criterion

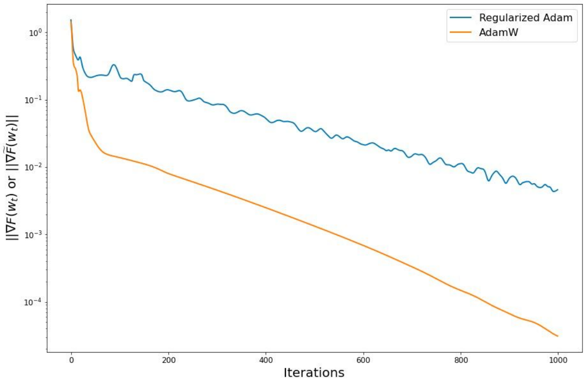


Fig. 2: Adam and AdamW with modified criterion

We think that this observation can be an explanation for the fact that AdamW performs better in the applied problems. The results of our experiments are presented below. We trained the ResNet18 [24] on four different optimizators using CosineAnnealingLr on dataset CIFAR10 [25].

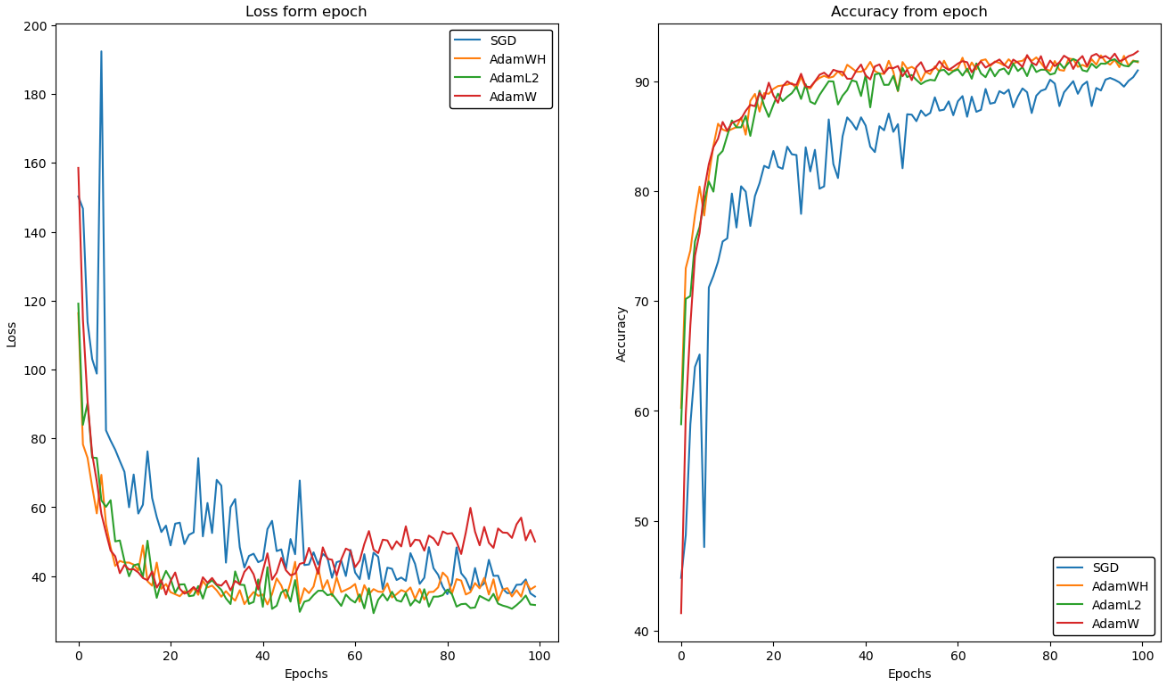
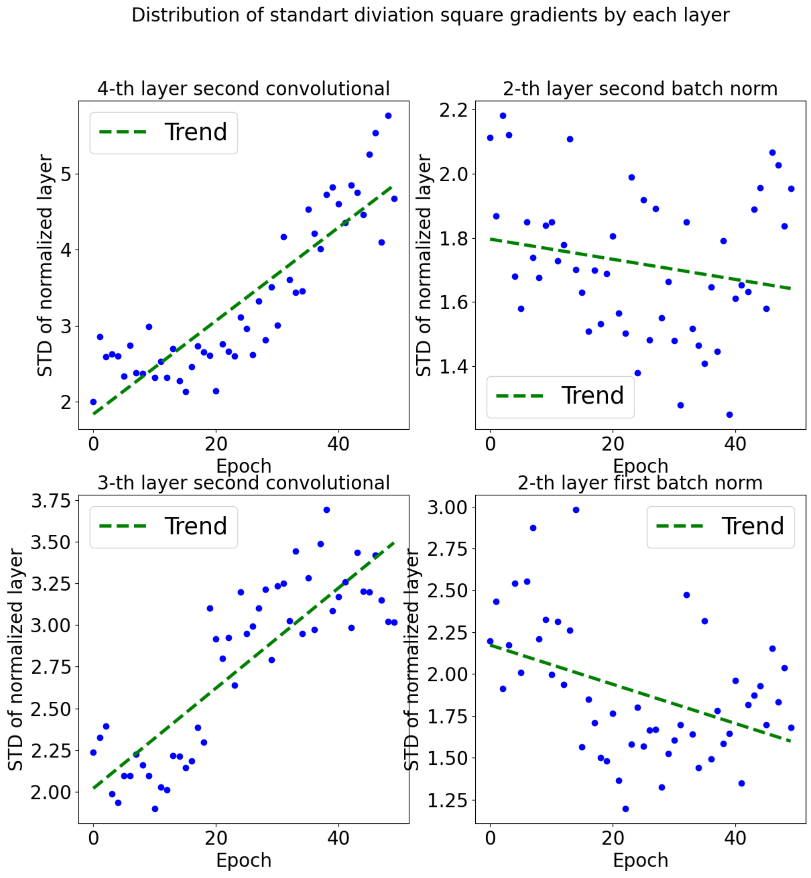


Fig. 3: Compare different optimization algorithms on dataset: CIFAR10



AdamW shows the worst results for the loss function, which repeats the results obtained with linear regression above. However, AdamW outperforms all other methods in terms of accuracy on the test dataset, demonstrating its generalization ability, as mentioned earlier.

We previously mentioned that the difference between the solutions of problems (4) and (9) can be bounded below by $||D_t - I||$. This is why we were interested in investigating the deviation of the elements of $||D_t||$ during training. Below, we examine the standard deviation of the normalized weights of the model across its layers, and we can observe certain trends in the convolutional and batch normalization layers that are framed in Figure 3.

Deviation of the normalized weights in the convolutional layers has rising trend. Hence, difference between solutions of different problems is bounded below and methods converge to a different optimums.

4 Conclusion

This study investigates the application of preconditioning methods with weight decay regularization. We propose a novel approach that combines preconditioning methods with weight decaying and analyze the convergence of these methods under various conditions. Theoretical and experimental analyses are conducted to investigate the solution to problem (9), demonstrating that methods incorporating preconditioning and weight decaying do not converge to the optimum of the initial problem (4). We anticipate that this paper will be valuable for researchers in this field. Further investigation is needed to understand the generalization ability of the introduced method and preconditioned methods with weight decaying.

References

[1] Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[3] Pádraig Cunningham, Matthieu Cord, and Sarah Jane Delany. Supervised learning. *Machine learning techniques for multimedia: case studies on organization and retrieval*, pages 21–49, 2008.

[4] Ryan M Rifkin and Ross A Lippert. Notes on regularized least squares. 2007.

[5] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[6] Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[7] Elad Hazan, Alexander Rakhlin, and Peter Bartlett. Adaptive online gradient descent. *Advances in Neural Information Processing Systems*, 20, 2007.

[8] Guodong Zhang, Chaoqi Wang, Bowen Xu, and Roger Grosse. Three mechanisms of weight decay regularization. *arXiv preprint arXiv:1810.12281*, 2018.

[9] Zhewei Yao, Amir Gholami, Sheng Shen, Mustafa Mustafa, Kurt Keutzer, and Michael W. Mahoney. Adahessian: An adaptive second order optimizer for machine learning, 2021.

[10] John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

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

[12] Andrew Goldberg, Xiaojin Zhu, Alex Furger, and Jun-Ming Xu. Oasis: Online active semi-supervised learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 25, pages 362–367, 2011.

[13] Feng Zhu, Hongsheng Li, Wanli Ouyang, Nenghai Yu, and Xiaogang Wang. Learning spatial regularization with image-level supervisions for multi-label image classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5513–5522, 2017.

[14] Yingbo Zhou, Caiming Xiong, and Richard Socher. Improved regularization techniques for end-to-end speech recognition. *arXiv preprint arXiv:1712.07108*, 2017.

[15] Tingting Wu, Xiao Ding, Minji Tang, Hao Zhang, Bing Qin, and Ting Liu. Stgn: an implicit regularization method for learning with noisy labels in natural language processing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7587–7598, 2022.

[16] Federico Girosi, Michael Jones, and Tomaso Poggio. Regularization theory and neural networks architectures. *Neural computation*, 7(2):219–269, 1995.

[17] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[18] T. Tieleman and G. Hinton. Lecture 6.5 - rmsprop: Divide the gradient by a running average of its recent magnitude. *Lecture 6.5 - rmsprop: Divide the gradient by a running average of its recent magnitude*, 2012.

[19] Sashank J. Reddi, Satyen Kale, and Sanjiv Kumar. On the convergence of adam and beyond, 2019.

[20] Alexandre Défossez, Léon Bottou, Francis Bach, and Nicolas Usunier. A simple convergence proof of adam and adagrad. *arXiv preprint arXiv:2003.02395*, 2020.

[21] Michael R. Zhang, James Lucas, Geoffrey Hinton, and Jimmy Ba. Lookahead optimizer: k steps forward, 1 step back, 2019.

[22] Abdurakhmon Sadiev, Aleksandr Beznosikov, Abdulla Jasem Almansoori, Dmitry Kamzolov, Rachael Tappenden, and Martin Takáč. Stochastic gradient methods with preconditioned updates. *arXiv preprint arXiv:2206.00285*, 2022.

[23] Aleksandr Beznosikov, Aibek Alanov, Dmitry Kovalev, Martin Takáč, and Alexander Gasnikov. On scaled methods for saddle point problems. *arXiv preprint arXiv:2206.08303*, 2022.

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition, 2015.

[25] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi. Cifar10-dvs: an event-stream dataset for object classification. *Frontiers in neuroscience*, 11:309, 2017.

5 Appendix

5.1 Experiments

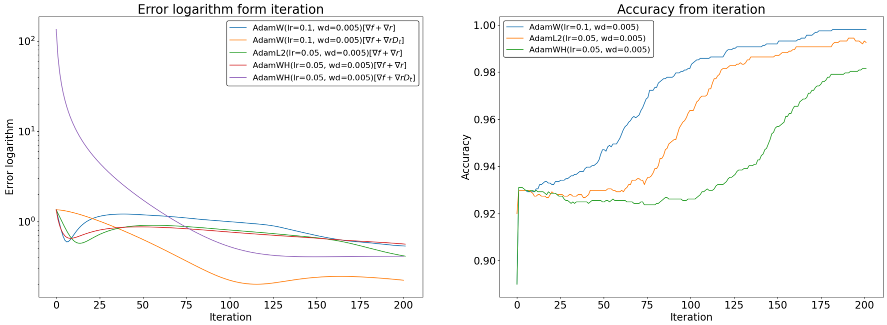


Fig. 4: Different versions Adam algrorithms on dataset: mushrooms

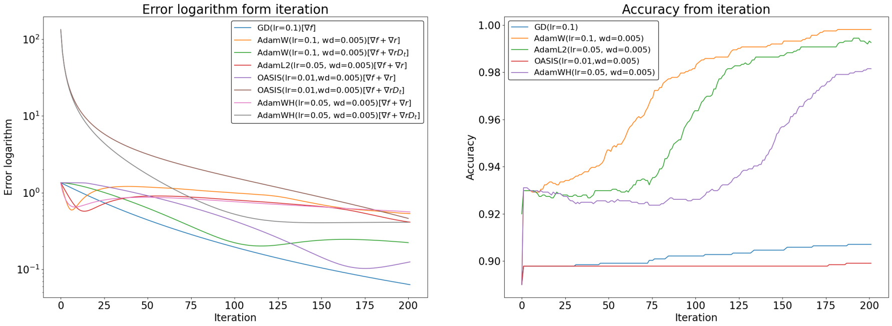


Fig. 5: Compare different optimization algorithms on dataset: mushrooms

5.2 Proofs of lemmas

Proof. (Proof of Lemma 1)

Using assumptions 1, 2 we can write the gradient of $\widetilde{r}$

$$\nabla\widetilde{r} = \nabla\left(\sum_{i=1}^{d} D_t^i r_i(w_i)\right) = D_t \begin{pmatrix} r'_1(w_1) \\ \vdots \\ r'_d(w_d) \end{pmatrix} = D_t \nabla r$$

□

Proof. (Proof of Lemma 2)

We can write definition of smoothness, using Lemma 1 and then apply 3

$$||\nabla\widetilde{r}(x) - \nabla\widetilde{r}(y)|| = ||\nabla\left(\sum_{i=1}^{d} D_t^i r_i(x_i)\right) - \nabla\left(\sum_{i=1}^{d} D_t^i r_i(y_i)\right)|| = ||D_t\left(\nabla r(x) - \nabla r(y)\right)|| \leq ||D_t||L_r$$

□

Proof. (Proof of lemma 3)

Lets write definitions of solutions w^*, $\widetilde{w}^*$:

$$\begin{cases} \nabla f(\widetilde{w}^*) + D_t \nabla r(\widetilde{w}^*) = 0 \\ \nabla f(w^*) + \nabla r(w^*) = 0 \end{cases},$$

Then we are able to get lower bound from the definition of L_F-contentiousness of the function F.

$$\|\widetilde{w}^* - w^*\|L_F \geq \|\nabla f(\widetilde{w}^*) + \nabla r(\widetilde{w}^*) - \nabla f(w^*) - \nabla r(w^*)\| = \| - D_t \nabla r(\widetilde{w}^*) + \nabla r(\widetilde{w}^*)\| = \|\nabla r(\widetilde{w}^*)(I - D_t)\|$$

□

5.3 Proofs of theorems

Proof. (Proof of theorem 1)

Let us use Assumption (3) for step t and $t+1$:

$$f(w_{t+1}) \leq f(w_t) + \langle \nabla f(w_t), w_{t+1} - w_t \rangle + \frac{L_f}{2}||w_{t+1} - w_t||^2, \tag{10}$$

By definition for our algorithm we have:

$$w_{t+1} - w_t = -\eta D_t^{-1} \nabla f(w_t) - \eta \nabla r(w_t).$$

From the previous expression , we select the gradient of the function

$$\nabla f(w_t) = \frac{1}{\eta} D^t (w_t - w_{t+1}) - D^t \nabla r(w_t),$$

replace $\nabla f(w_t)$ in 10 and by definition of matrix D_t, $I \preccurlyeq \frac{D_t}{\alpha}$

$$f(w_{t+1}) \leq f(w_t) + \langle \frac{1}{\eta} D_t (w_t - w_{t+1}) - D_t \nabla r(w_t), w_{t+1} - w_t \rangle + \frac{L_f}{2\alpha}||w_{t+1} - w_t||^2_{D_t} =$$

$$= f(w_t) + \left(\frac{L_f}{2\alpha} - \frac{1}{\eta}\right)||w_{t+1} - w_t||^2_{D_t} - \langle D_t \nabla r(w_t), w_{t+1} - w_t \rangle,$$

using the notation of $\tilde{r} : \nabla\tilde{r} = D_t \nabla r(w_t)$, we can rewrite step using the variable and Assumption (3)

$$\tilde{r}(w_{t+1}) \leq \tilde{r}(w_t) + \langle \nabla\tilde{r}(w_t), w_{t+1} - w_t \rangle + \frac{L_{\tilde{r}}}{2}||w_{t+1} - w_t||_2^2$$

Let us replace the old regularization function with a new one

$$f(w_{t+1}) \leq f(w_t) + \left(\frac{L_f}{2\alpha} - \frac{1}{\eta}\right)||w_{t+1} - w_t||^2_{D_t} + \tilde{r}(w_t) - \tilde{r}(w_{t+1}) + \frac{\Gamma L_{\tilde{r}}}{2}||w_{t+1} - w_t||^2_{D_t}.$$

Now let us define a new loss function $\tilde{F}(w) = f(w) + \tilde{r}(w)$, $F(w) = f(w) + r(w)$, $(\tilde{L} = L_f + \Gamma L_{\tilde{r}\alpha})$, we get:

$$\tilde{F}(w_{t+1}) \leq \tilde{F}(w_t) + \left(\frac{\tilde{L}}{2\alpha} - \frac{1}{\eta}\right)||w_{t+1} - w_t||^2_{D_t},$$

we select the step in such a way that $\frac{\tilde{L}}{2\alpha} - \frac{1}{\eta} < 0$, $\eta < \frac{2\alpha}{\tilde{L}}$

$$\left(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha}\right)||w_{t+1} - w_t||^2_{D_t} \leq \tilde{F}(w_t) - \tilde{F}(w_{t+1}).$$

Let us sum up our inequalities and evaluate the left part from below

$$\frac{\eta^2(T+1)}{\Gamma}\left(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha}\right)\cdot \min_{k=0,T}||\nabla f(w_t) + \nabla\tilde{r}(w_t)||^2 \leq \frac{\eta^2}{\Gamma}\left(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha}\right)\cdot\sum_{t=0}^{T}||\nabla f(w_t) + \nabla\tilde{r}(w_t)||^2 \leq \tilde{F}(w_0) - \tilde{F}(w_*).$$

Moving everything to the right we get the following estimate

$$\min_{t=0,T}||\nabla f(w_t) + \nabla\tilde{r}(w_t)||^2 \leq \frac{(\tilde{F}(w_0) - \tilde{F}(w_*))\Gamma}{(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha})\eta^2(T+1)} = \varepsilon$$

$$T + 1 \geq \frac{\Delta_0\Gamma}{(\frac{1}{\eta} - \frac{\tilde{L}}{2\alpha})\eta^2\varepsilon}.$$

We get an estimate for the number of steps required for a given accuracy

$$T = \mathcal{O}\left(\frac{2\Delta_0\Gamma\alpha}{(2\alpha - \tilde{L}\eta)\eta\varepsilon}\right).$$

□

Proof. (Proof of theorem 2)

The proof of this theorem will be similar to the previous one, the main difference is that we impose another Assumption 3 on the original function Assume

$$\nabla\tilde{F} = \nabla f + \nabla\tilde{r}$$

$$L_F + ||D_t||L_r = L_{\tilde{F}}$$

rewrite step in terms of new function

$$w_{t+1} - w_t = -\eta D_t^{-1}\nabla r(w_t) - \eta\nabla r(w_t) = -\eta D_t^{-1}(\nabla f + \nabla\tilde{r})(w_t) = -\eta D_t^{-1}\nabla\tilde{F}(w_t),$$

Then we write $\tilde{L}$-smoothness for $\tilde{F}$

$$\tilde{F}(w_{t+1}) - \tilde{F}(w_t) \leq \langle \nabla\tilde{F}(w_t), w_{t+1} - w_t\rangle + \frac{L_{\tilde{F}}}{2}||w_{t+1} - w_t||^2,$$

then combine it together and use constraints on the matrix $\alpha \cdot I \preccurlyeq D_t \preccurlyeq \Gamma \cdot I$

$$\tilde F(w_{t+1}) - \tilde F(w_t) \le -\langle \frac{1}{\eta} D_t(w_{t+1} - w_t), w_{t+1} - w_t\rangle + \frac{L_{\tilde F}}{2}||w_{t+1} - w_t||^2 = \left(\frac{L_{\tilde F}}{2\alpha} - \frac{1}{\eta}\right)||w_{t+1} - w_t||^2_{D_t} =$$

$$= \left(\frac{L_{\tilde F}}{2\alpha} - \frac{1}{\eta}\right)||w_{t+1} - w_t||^2_{D_t} = \left(\frac{L_{\tilde F}}{2\alpha} - \frac{1}{\eta}\right)|| - \eta D_t^{-1}\nabla \tilde F(w_t)||^2_{D_t} \le \left(\frac{L_{\tilde F}}{2\alpha} - \frac{1}{\eta}\right)\eta^2||\nabla \tilde F(w_t)||^2_{D_t^{-1}},$$

We use PL-condition 4 for the function $\tilde F$:

$$||\nabla \tilde F(w_t)||^2_{D_t^{-1}} \ge 2\mu(\tilde F(w_t) - \tilde F^*),$$

subtract the exact solution from both parts and apply PL-condition

$$\tilde F(w_t) - F^* \ge \tilde F(w_{t+1}) - \tilde F^* + \left(\frac{1}{\eta} - \frac{L_{\tilde F}}{2\alpha}\right)\eta^2 2\mu(\tilde F(w_t) - \tilde F^*) = \left(1 + 2\mu\eta^2\left(\frac{1}{\eta} - \frac{L_{\tilde F}}{2\alpha}\right)\right)(\tilde F(w_{t+1}) - \tilde F^*).$$

Let us apply the expression for each step and use expression $\Delta_0 = \tilde F(w_0) - \tilde F(w_*)$

$$\varepsilon \ge \Delta_0\left(1 + 2\mu\eta^2\left(\frac{1}{\eta} - \frac{L_{\tilde F}}{2\alpha}\right)\right)^{-T} \ge (\tilde F(w_T) - \tilde F^*),$$

from this expression we get the necessary number of steps to get together with the error ε

$$T = \frac{\ln\frac{\Delta_0}{\varepsilon}}{\ln\left(1 + 2\mu\eta^2\left(\frac{1}{\eta} - \frac{L_{\tilde F}}{2\alpha}\right)\right)} \approx \frac{\ln\frac{\Delta_0}{\varepsilon}}{2\mu\eta^2\left(\frac{1}{\eta} - \frac{L_{\tilde F}}{2\alpha}\right)}.$$

We get an estimate on the number of steps of the algorithm for a given error on the difference of the loss function

$$T = \mathcal{O}\left(\frac{\ln\frac{\Delta_0}{\varepsilon}}{2\mu\eta^2(\frac{1}{\eta} - \frac{L_{\tilde F}}{2\alpha})}\right)$$

□

Proof. (Proof of theorem 3)

We start from L-smoothness of f (Assumption (3)).

$$f(w_{t+1}) \le f(w_t) + \langle\nabla f(w_t), w_{t+1} - w_t\rangle + \frac{L_f}{2}||w_{t+1} - w_t||_2^2$$

Next we substitute an update of w:

$$f(w_{t+1}) \le f(w_t) + \langle\nabla f(w_t), w_{t+1} - w_t\rangle + \frac{L_f\eta^2}{2}||D_t^{-1}g_t + \nabla r(w_t)||_2^2 =$$

Then we added and subtracted $D_t^{-1}\nabla f(w_t)$ under of the norm:

$$= f(w_t) + \langle\nabla f(w_t), w_{t+1} - w_t\rangle + \frac{L_f\eta^2}{2}||D_t^{-1}g_t - D_t^{-1}\nabla f(w_t) + D_t^{-1}\nabla f(w_t) + \nabla r(w_t)||_2^2 =$$

$$= f(w_t) + \langle\nabla f(w_t), w_{t+1} - w_t\rangle + \frac{L_f\eta^2}{2}||D_t^{-1}g_t - D_t^{-1}\nabla f(w_t)||_2^2 + \frac{L_f\eta^2}{2}\langle D_t^{-1}g_t - D_t^{-1}\nabla f(w_t), D_t^{-1}\nabla f(w_t) + \nabla r(w_t)\rangle +$$

$$+ \frac{L_f\eta^2}{2}||D_t^{-1}\nabla f(w_t) + \nabla r(w_t)||_2^2.$$

Then we take full expectation:

$$\mathbb{E}f(w_{t+1}) \le \mathbb{E}f(w_t) + \frac{L_f\eta^2}{2}\mathbb{E}||D_t^{-1}g_t - D_t^{-1}\nabla f(w_t)||_2^2 + \frac{L_f\eta^2}{2}\mathbb{E}\langle D_t^{-1}g_t - D_t^{-1}\nabla f(w_t), D_t^{-1}\nabla f(w_t) + \nabla r(w_t)\rangle$$

$$+\frac{L_f\eta^2}{2}\mathbb{E}||D_t^{-1}\nabla f(w_t)+\nabla r(w_t)||_2^2.$$

Using Assumption 4, $\mathbb{E}\left[D_t^{-1}(g_t-\nabla f(w_t)\right]=0$, $||g_t-\nabla f(w_t)||_2^2\le\sigma^2$

$$\mathbb{E}f(w_{t+1})\le\mathbb{E}f(w_t)+\mathbb{E}\langle\nabla f(w_t),w_{t+1}-w_t\rangle+\frac{L_f\eta^2}{2\alpha^2}\mathbb{E}||g_t-\nabla f(w_t)||_2^2+\frac{L_f\eta^2}{2\alpha^2}\mathbb{E}||\nabla f(w_t)+D_t\nabla r(w_t)||_2^2,$$

then apply second part of Assumption 4:

$$\mathbb{E}f(w_{t+1})\le\mathbb{E}f(w_t)+\mathbb{E}\langle\nabla f(w_t),w_{t+1}-w_t\rangle+\frac{L_f\eta^2\sigma^2}{2\alpha^2}+\frac{L_f\eta^2}{2\alpha^2}\mathbb{E}||\nabla f(w_t)+D_t\nabla r(w_t)||_2^2.$$

Then again use 4

$$\nabla f(w_t)=\mathbb{E}g_t=\mathbb{E}\frac{1}{\eta}D_t(w_t-w_{t+1})+D_t\nabla r(w_t)$$

and put it in the inequality:

$$\mathbb{E}f(w_{t+1})\le\mathbb{E}f(w_t)+\mathbb{E}\langle\mathbb{E}\frac{1}{\eta}D_t(w_t-w_{t+1})+D_t\nabla r(w_t),w_{t+1}-w_t\rangle+\frac{L_f\eta^2}{2\alpha^2}\mathbb{E}||\nabla f(w_t)+D_t\nabla r(w_t)||_2^2+\frac{L_f\eta^2\sigma^2}{2\alpha^2}.$$

Let's rewrite it in a convenient way:

$$\mathbb{E}f(w_{t+1})\le\mathbb{E}f(w_t)-\frac{1}{\eta}\mathbb{E}||w_{t+1}-w_t||_{D_t}^2+\mathbb{E}\langle D_t\nabla r(w_t),w_{t+1}-w_t\rangle+\frac{L_f\eta^2}{2\alpha^2}\mathbb{E}||\nabla f(w_t)+D_t\nabla r(w_t)||_2^2+\frac{L_f\eta^2\sigma^2}{2\alpha^2}.$$

Using Lemma 2 about $L_{\tilde{r}}-smoothness$ of $\tilde{r}(w_t)=D_t\nabla r(w_t)$

$$\mathbb{E}f(w_{t+1})\le\mathbb{E}f(w_t)-\frac{1}{\eta}\mathbb{E}||w_{t+1}-w_t||_{D_t}^2+\mathbb{E}\tilde{r}(w_t)-\mathbb{E}\tilde{r}(w_{t+1})+\frac{L_{\tilde{r}}}{2}\mathbb{E}||w_{t+1}-w_t||_2^2+\frac{L_f\eta^2}{2\alpha^2}\mathbb{E}||\nabla f(w_t)+D_t\nabla r(w_t)||_2^2+$$

$$+\frac{L_f\eta^2\sigma^2}{2\alpha^2}.$$

Then we apply $L_{\tilde{F}}-smoothness$ and get:

$$\mathbb{E}\left(f(w_t)+\tilde{r}(w_t)\right)\le\mathbb{E}\left(f(w_{t+1})+\tilde{r}(w_{t+1})\right)+\mathbb{E}||w_{t+1}-w_t||_{D_t}^2\left(-\frac{1}{\eta}+\frac{\Gamma L_{\tilde{r}}}{2}+\frac{\Gamma L_{\tilde{F}}L_f\eta^2}{2\alpha^2}\right)+\frac{L_f\eta^2\sigma^2}{2\alpha^2}.$$

And with restrictions on η such that: $\left(-\frac{1}{\eta}+\frac{\Gamma L_{\tilde{r}}}{2}+\frac{\Gamma L_{\tilde{F}}L_f\eta^2}{2\alpha^2}\right)\le 0$:

$$\left(-\frac{1}{\eta}+\frac{\Gamma L_{\tilde{r}}}{2}+\frac{\Gamma L_{\tilde{F}}L_f\eta^2}{2\alpha^2}\right)\mathbb{E}||w_{t+1}-w_t||_{D_t}^2\le\mathbb{E}\tilde{F}(w_t)-\mathbb{E}\tilde{F}(w_{t+1})+\frac{L_f\eta^2\sigma^2}{2\alpha^2}.$$

Then using the expectation and $L_{\tilde{F}}-smoothness$:

$$\frac{T}{\Gamma}\left(\frac{1}{\eta}-\frac{\Gamma L_{\tilde{r}}}{2}-\frac{\Gamma L_{\tilde{F}}L_f\eta^2}{2\alpha^2}\right)\min_{k=0,T-1}||\nabla f(w_t)+\nabla\tilde{r}(w_t)||_2^2\le\frac{1}{\Gamma}\cdot\left(\frac{1}{\eta}-\frac{\Gamma L_{\tilde{r}}}{2}-\frac{\Gamma L_{\tilde{F}}L_f\eta^2}{2\alpha^2}\right)\sum_{i=0}^{T-1}||\nabla f(w_t)+\nabla\tilde{r}(w_t)||_2^2$$

$$\le\tilde{F}(w_0)-\tilde{F}(w_*)+T\cdot\frac{L_f\eta^2\sigma^2}{2\alpha^2}\le\varepsilon$$

Choose $\eta\sim\frac{(\tilde{F}(w_0)-\tilde{F}(w_*))\alpha^2}{L_f\sigma^2}$, we get an estimate the number of steps of the algorithm for a given error ε:

$$T=\mathcal{O}\left(\frac{\Gamma\Delta_0}{\left(\frac{1}{\eta}-\frac{\Gamma L_{\tilde{r}}}{2}-\frac{\Gamma L_{\tilde{F}}L_f\eta^2}{2\alpha^2}\right)\varepsilon}\right)$$

□