META-REGULARIZATION: AN APPROACH TO ADAPTIVE CHOICE OF THE LEARNING RATE IN GRADIENT DESCENT

Guangzeng Xie, Hao Jin & Dachao Lin
Peking University
Beijing, China
{smsxgz, jin.hao, lindachao}@pku.edu.cn

Zhihua Zhang
Peking University
Beijing, China
zhzhang@pku.edu.cn

ABSTRACT

We propose Meta-Regularization, a novel approach for the adaptive choice of the learning rate in first-order gradient descent methods. Our approach modifies the objective function by adding a regularization term on the learning rate, and casts the joint updating process of parameters and learning rates into a maxmin problem. Given any regularization term, our approach facilitates the generation of practical algorithms. When Meta-Regularization takes the $\varphi$-divergence as a regularizer, the resulting algorithms exhibit comparable theoretical convergence performance with other first-order gradient-based algorithms. Furthermore, we theoretically prove that some well-designed regularizers can improve the convergence performance under the strong-convexity condition of the objective function. Numerical experiments on benchmark problems demonstrate the effectiveness of algorithms derived from some common $\varphi$-divergence in full batch as well as online learning settings.

1 Introduction

The automatic choice of the learning rate remains crucial in improving the efficiency of gradient descent algorithms. Strategies regardless of training information, such as the learning rate decay, might drive the learning rate too large or too small during the training process, which tends to negatively affect the convergence performance. In order to improve performance, adaptively updating the learning rate during the training process would be desirable.

There are two common approaches for updating the learning rate in the first-order gradient descent methods in the literature. Firstly, line search for a proper learning rate is a natural and direct approach to fully utilizing the currently received gradient information. There are practically many ways to carry out such exact or inexact line search. Specifically, Hyper-Gradient Descent [3] can be viewed as an approximate line search through using the gradient with respect to the learning rate of the update rule itself. Secondly, some methods leverage historical gradients to approximate the inverse of Hessian matrix, which is essential in Newton method [23]. Quasi-Newton Methods [20] as well as the Barzilai-Broweinin (BB) [2] method all fall in the scope of such algorithms.

In this paper we propose a novel approach to the adaptive choice for the learning rate that we call Meta-Regularization. The key idea is to impose some constraints on the updates of learning rate during the training process, which is equivalent to adding a regularization term on the learning rate to the objective function. Through introducing a regularization term on the learning rate, our approach casts the joint updating process of parameters and learning rates into a maxmin problem. In other words, our approach gives a pipeline to generate practical algorithms from any regularization term. Various regularization terms bring out various strategies of updating learning rate, which include AdaGrad [9] and WNGrad [28]. Compared with the Hyper-Gradient and BB methods, our approach is attractive due to its ability in construction and theoretical analysis of the corresponding algorithms.

Taking the regularization term derived from the $\varphi$-divergence as an instance, we theoretically analyze the overall performance of the resulting algorithms, and evaluate some representative algorithms on benchmark problems. Theoretical guarantees of these algorithms are provided in both full batch and online learning settings, which are not explicitly given in the original work of Hyper-Gradient Descent. Moreover, certain modifications of regularization terms from the $\varphi$-divergence manage to improve the theoretical convergence performance while the original objective function is strongly convex. In terms of numerical experiments, we generate several algorithms from some common
We now give the notion of the \( \varphi \)-divergence without delicate design to represent the general performance of such algorithms. Experimental results not only reveal a generally comparable performance with Hyper-Gradient Descent as well as BB method, but also demonstrate outperformance over these two algorithms in online learning and full batch settings, respectively.

The main contributions of our paper are as follows:

- To our knowledge, we are the first to formally consider the usage of regularization technique in adaptively updating the learning rate, giving rise to a pipeline to construct algorithms from any given regularization term.
- We provide theoretical analysis of the convergence performance for a family of algorithms derived from our approach when taking a generalized distance function such as the \( \varphi \)-divergence as the regularizer.
- Experimental results demonstrate that our Meta-Regularization method based on the \( \varphi \)-divergence is practically comparable with the BB method and Hyper-Gradient Descent, and even outperforms them in some cases.

## 2 Related Work

Steepest Descent uses the received gradient direction and an exact or inexact line search to obtain proper learning rates. Although Steepest Descent uses the direction that descends most and the best learning rate that gives the most reduction of objective function value, Steepest Descent may converge very slow for convex quadratic functions when the Hessian matrix is ill-conditioned [29]. In practice, some line search conditions such as Goldstein conditions or Wolfe conditions [10] can be applied to compute the learning rate. In online or stochastic settings, one observes stochastic gradients rather than exact gradients and line search methods become less effective.

The BB method [2] which was motivated by quasi-Newton methods presents a surprising result that it could lead to superlinear convergence in convex quadratic problem of two variables. Although numerical results often show that the BB method converges superlinearly in solving nonlinear optimization problems, no superlinear convergence results have been established even for an \( n \)-dimensional strictly convex quadratic problem with the order \( n > 2 \) [2, 6].

In minimizing the sum of cost functions and stochastic setting, SGD-BB proposed by [26] takes the average of the stochastic gradients in one epoch as an estimation of the full gradient. But this approach can not directly be applied to online learning settings.

In online convex optimization [31, 24, 14], AdaGrad adapts the learning rate on per parameter basis dynamically. This leads to many variants such as RMSProp [27], AdaDelta [50], Adam [17], etc.

Additionally, [5] analyzed Adaptive Stochastic Gradient Descent (ASGD) which is a generalization of Kesten’s accelerated stochastic approximation algorithm [16] for the high-dimensional case. ASGD uses a monotone decreasing function with respect to a time variable to get learning rates. Recently, [8] proposed Hyper-Gradient Descent to learn the global learning rate in SGD, SGD with Nesterov momentum and Adam. Hyper-Gradient Descent can be viewed as an approximate line search method in the online learning setting and it uses the update rule for the previous step to optimize the learning rate in the current step. However, Hyper-Gradient Descent has no theoretical guarantee.

It is worth mentioning that [12] proposed a framework named Unified Adaptive Regularization from which AdaGrad and Online Newton Step [13] can be derived. However, Unified Adaptive Regularization gives an approach for approximating the Hessian matrix in second order methods.

Our framework stems from the work of [11], who adjusted the weights of the weighted least squares problem by solving an extra objective function which adds a regularizer about the weights to origin objective function.

## 3 Problem Formulation

Before introducing our approach, we present the notation that will be used. We denote the set \( \{ x > 0 : x \in \mathbb{R} \} \) by \( \mathbb{R}_{++} \). For two vectors \( a, b \in \mathbb{R}^d \), we use \( a / b \) to denote element-wise division, \( a \odot b \) for element-wise product (the symbol \( \odot \) will be omitted in the explicit context), \( a^n = (a_1^n, a_2^n, \ldots, a_d^n) \), and \( a \geq b \) if \( a_j \geq b_j \) for all \( j \). Let \( \mathbf{1} \) be the vector of ones with an appropriate size, and \( \text{diag}(\beta) \) be a diagonal matrix with the elements of the vector \( \beta \) on the main diagonal. In addition, we define \( \| a \|_A = \sqrt{(a, A a)} \) where \( A \) is a positive semidefinite matrix.

Given a set \( X \subseteq \mathbb{R}^d \), a function \( f : X \to \mathbb{R} \) is said to satisfy \( f \in C^1_{L, 1}(X) \) if \( f \) is continuously differentiable on \( X \), and the derivative of \( f \) is Lipschitz continuous on \( X \) with constant \( L \):

\[
\| \nabla f(x) - \nabla f(y) \|_2 \leq L \| x - y \|_2.
\]

More general definition can be found in [22].

We now give the notion of the \( \varphi \)-divergence.
We usually set the auxiliary variable $\eta$ where $\alpha$ is the learning rate. To derive our meta-regularization approach, we then formulate this minimization problem:

$$x_{t+1} = \arg\min_{x \in \mathcal{X}} \left\langle g_t, x - x_t \right\rangle + \frac{1}{2\alpha} \|x - x_t\|_2^2,$$

where $\alpha$ is the learning rate. To derive our meta-regularization approach, we then formulate this minimization problem as a saddle point problem by adding a meta-regularizer about the difference between the new learning rate $\alpha$ and an auxiliary variable $\eta_t$. Accordingly, we have

$$\max_{\alpha \in \mathcal{A}_t} \min_{x \in \mathcal{X}} \Psi_t(x, \alpha) = \left\langle g_t, x - x_t \right\rangle + \frac{1}{2} \left( \frac{1}{\alpha} \|x - x_t\|_2^2 - D(\alpha, \eta_t) \right),$$

where $D(\alpha, \eta)$, a distance function, is defined as our meta-regularizer and $\mathcal{A}_t$ is a subset in $\mathbb{R}$. Our framework solves this saddle point problem for a new predictor and a new learning rate.

We usually set the auxiliary variable $\eta_t$ equal to $\alpha_t$ and consider the meta-regularizer as the penalty of the change between $\alpha_{t+1}$ and $\alpha_t$. Sometimes we also can choose the sequence $\{\eta_t\}$ in advance, before our methods start the job. In this case, our framework can be treated as a smoothing technique to stabilize the learning rate.

### 4.1 Update Rules

In this section we present two update rules of our meta-regularization framework. The first update rule is solving saddle point problem [3] exactly. That is,

$$\Psi_t(x_{t+1}, \alpha_{t+1}) = \max_{\alpha \in \mathcal{A}_t} \min_{x \in \mathcal{X}} \Psi_t(x, \alpha).$$

In the setting $\eta_t = \alpha_t$, it is more recommended to employ an alternating strategy in practice.

The second update rule is an alternatively iterative procedure between $\alpha$ and $x$. Under the assumption that the optimal value of $\alpha$ is close to $\eta_t$, we solve an approximate equation for finding $\alpha_{t+1}$:

$$\alpha_{t+1} = \arg \max_{\alpha \in \mathcal{A}_t} \Psi_t \left( \arg \min_{x \in \mathcal{X}} \Psi_t(x, \alpha_t), \alpha \right),$$
and update the new predictor $x_{t+1}$ via
\[ x_{t+1} = \arg \min_{x \in X} \Psi_t(x, \alpha_{t+1}). \]

It is worth noting that these two update rules share similar performance in some certain situations (see Theorems 8 and 4 in Section 5.3).

4.2 Diagonal Meta-Regularization

Consider a generalization of the standard gradient descent [9]
\[ x_{t+1} = \Pi_{\alpha_t} g_t, \quad x_{t+1} = x_t - \alpha_t g_t, \]
\[ x_{t+1} = \arg \min_{x \in X} \|x - x_t\|^2_{\alpha_t^{-1} - \alpha_t^{-1} - D(\alpha_t, \eta_t)}. \]

Similarly, we can add our meta regularizer to the minimization problem (6) as
\[ \max_{\alpha \in A_t} \min_{x \in X} \Psi_t(x, \alpha) \triangleq \langle g_t, x - x_t \rangle + \frac{1}{2} \|x - x_t\|^2_{\alpha_t^{-1} - \alpha_t^{-1} - D(\alpha_t, \eta_t)}, \] (7)

where $A_t \subseteq \mathbb{R}^d$.

5 Algorithm Design and Analysis

In this section, we show how to design specific algorithms according to our framework, especially diagonal Meta-Regularization, and provide theoretical analysis for corresponding algorithms.

5.1 Algorithms for Two Update Rules

We choose the $\varphi$-divergence as our meta-regularizer. Accordingly, we rewrite the problem (7) as
\[ \max_{\alpha \in A_t} \min_{x \in X} \psi_t(x, \alpha) \triangleq \sum_{j=1}^d g_{t,j} (x_j - x_{t,j}) + \frac{1}{2} \left( (x_j - x_{t,j})^2 / \alpha_j - \varphi(\eta_{t,j}/\alpha_j) / \eta_{t,j} \right). \]

The form of problem (8) implies that we can solve the problem for each dimension separately, and consequently only a little extra run time is required for each step. In order to solve the problem feasibly, we always assume that $\lim_{z \to +\infty} \varphi'(z) = +\infty$.

The following lemma and Algorithm 1 give the concrete scheme of solving the saddle point problem (8) exactly.

**Lemma 2.** Considering problem (8) without constraints and solving the problem exactly, we get new predictor $x_{t+1}$ and new learning rate $\alpha_{t+1}$ such that
\[ \varphi'(\eta_{t,j}/\alpha_{t+1,j}) = \alpha_{t+1,j}^2 g_{t,j}, \]
\[ x_{t+1} = x_t - \alpha_{t+1} \circ g_t. \] (9)

**Remark 3.** Note that AdaGrad [9] and WNGrad [28] are special cases of Algorithm 1 with a particular choice of $\varphi$ (detailed derivation in Appendix 8).

- If $\varphi(z) = z + \frac{1}{z} - 2$, then we can derive AdaGrad from Algorithm 1.
- If $\varphi(z) = \frac{1}{2} - \log(\frac{1}{2}) - 1$, then we can derive WNGrad from Algorithm 1.

Applying the alternating update rule, which we described in Section 4.1 under the same assumption in Lemma 2, we obtain the following lemma and Algorithm 2.
For any $\varepsilon$ GD with Meta-regularization using alternating update rule

Algorithm 2 GD with Meta-regularization using alternating update rule

Require: $\alpha_0 = \alpha_0 > 0$, $x_0$
1: for $t = 1$ to $T$ do
2: Suffer loss $f_t(x_t)$;
3: Receive subgradient $g_t \in \partial f_t(x_t)$ of $f_t$ at $x_t$;
4: Update $\alpha_{t+1,j}$ as the solution of the equation $\varphi'(\eta_{t,j}/\alpha) = \alpha^2 g_{t,j}, j = 1, \ldots, d$;
5: Update $x_{t+1} = x_t - \alpha_{t+1} \circ g_t$;
6: end for

Lemma 3. Considering problem \cite{5} without constraint and following from the alternating update rule, we get new predictor $x_{t+1}$ and new learning rate $\alpha_{t+1}$ as

\[
\alpha_{t+1,j} = \frac{\eta_{t,j}}{\langle \varphi' \rangle^{-1}(\eta_{t,j}^2 g_{t,j}^2)}, j = 1, \ldots, d,
\]

\[
x_{t+1} = x_t - \alpha_{t+1} \circ g_t.
\]

Computing the inverse function of $\varphi'$ is usually easier than solving the equation \cite{9} in practice, especially for the widely used $\varphi$-divergences (more details can be found in Appendix ??).

5.1.1 Full Batch Setting

Instead of diagonal Meta-regularization, we consider origin Meta-regularization \cite{3} here. Recall that we set $f_t = F$ in the full batch setting, and assume that $F \in C^1_{L,1}$ without convexity. In this case, two update rules can be written as

\[
\begin{align*}
\varphi'\left(\frac{\alpha_t}{\alpha_{t+1}}\right) &= \alpha^2_{t+1} \|g_t\|^2_2, \\
x_{t+1} &= x_t - \alpha_{t+1} g_t, \\
\end{align*}
\]

\[
\begin{align*}
\alpha_{t+1} &= \alpha_t / (\varphi')^{-1}(\alpha^2 g_t^2), \\
x_{t+1} &= x_t - \alpha_{t+1} g_t.
\end{align*}
\]

Next we show that convergence of both update rules \cite{11} and \cite{12} are robust to the choice of initial learning rate.

Theorem 4. Suppose that $\varphi \in C^1_{L,1}((1, +\infty))$, $\varphi$ is $\alpha$-strongly convex, $F \in C^1_{L,1}(\mathbb{R}^d)$, and $F^* = \inf_{x} F(x) > -\infty$. For any $\varepsilon \in (0, 1)$, the sequence $\{x_t\}$ obtained from update rules \cite{11} or \cite{12} satisfies

\[
\min_{j=0:T-1} \|\nabla F(x_j)\|^2_2 \leq \varepsilon,
\]

after $T = \mathcal{O}\left(\frac{1}{\varepsilon}\right)$ steps.

More detailed results of Theorem \cite{4} for runtime can be found in Theorems \cite{23} and \cite{24} in Appendix \cite{13}. Theorem \cite{4} shows that both runtime of the two update rules can be bound as $\mathcal{O}(1/\varepsilon)$ for any constant $L$ and initial learning rate $\alpha_0$. Comparing with classical convergence result (see (1.2.13) in \cite{22} or Theorem \cite{22} in Appendix), the upper bound of runtime is $\mathcal{O}(1/\varepsilon)$ only for a certain range (related to $L$) of initial learning rates.

5.2 Logarithmic Regret Bounds

In this subsection, we show that employing some specific distance functions instead of the $\varphi$-divergence as a regularizer can improve convergence rate effectively. We make use of an example of optimization problems in which the objective function is strongly convex.

First, we define $\mu$-strong convexity.
Algorithm 3 GD with SC-Meta-regularization

**Require:** $\alpha_0 = \alpha_0, 1 > 0, x_0$

1. for $t = 1$ to $T$ do
2.  Suffer loss $f_t(x_t)$;
3.  Receive $g_t \in \partial f_t(x_t)$ of $f_t$ at $x_t$;
4.  Update $\alpha_{t+1,j}$ as the solution of the equation $\lambda(\alpha_{t,j}/\alpha^2)\varphi'(\alpha_{t,j}/\alpha) = g^2_{t,j}, j = 1, \cdots, d$;
5.  Update $x_{t+1} = x_t - \alpha_{t+1} \circ g_t$;
6. end for

**Definition 5** (Definition 2.1 in [21]). Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a convex set. We say that a function $f : \mathcal{X} \to \mathbb{R}$ is $\mu$-strongly convex if there exists $\mu \in \mathbb{R}^d$ with $\mu_j > 0$ for $j = 1, \cdots, d$ such that for all $x, y \in \mathcal{X}$,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\|y - x\|_{\text{diag}(\mu)}^2.$$ 

Let $\xi = \min_{j=1:d} \mu_j$. Then $f$ is $\xi$-strongly convex (in the usual sense), that is,

$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\xi}{2}\|y - x\|_2^2.$$

We now propose a modification of Meta-Regularization that we refer to as SC-Meta-Regularization. The modification uses a family of distance functions $D : \mathbb{R}^d_+ \times \mathbb{R}^d_+ \to \mathbb{R}$ as follows

$$D(u, v) = \sum_{j=1}^{d} \varphi(u_j/v_j),$$

(13)

where $\varphi$ is convex function with $\varphi(1) = \varphi'(1) = 0$ like we used in the $\varphi$-divergence.

**Remark 4.** Same as the $\varphi$-divergence, $D(u, v) \geq 0$ for any $u, v \in \mathbb{R}^d_+$.

---

Figure 1: Convergence performances of algorithms on CIFAR-10 in the online learning setting. **left:** training loss of the last training epoch at different initial learning rates; **right:** testing accuracy of the last training epoch at different initial learning rates.

Different from Algorithms [1] and [2] we add a hyper-parameter $\lambda > 0$ like AdaGrad to SC-AdaGrad. Rewrite problem (7) as

$$\max_{\alpha \in \mathcal{A}} \min_{x \in \mathcal{X}} \Psi_t(x, \alpha) \triangleq g^T_t (x - x_t) + \frac{1}{2}\|x - x_t\|_{\text{diag}(\alpha)}^2 - \frac{\lambda}{2} \sum_{j=1}^{d} \varphi(\alpha_{t,j}/\alpha_j),$$

(14)

and give the corresponding algorithm in Algorithm 3.

**Theorem 6.** Suppose that $f_t$ is $\mu$-strongly convex for all $t$, $\varphi \in C_1^1([1, +\infty))$, and $\varphi$ is $\gamma$-strongly convex. Assume that $\|g_t\|_{\infty} \leq G$, and $\lambda \geq G^2/((\gamma \min_{j=1:d} \mu_j))$. Then the sequence $\{x_t\}$ obtained from Algorithm 3 satisfies

$$2R(T) \leq T \left( 1 + \frac{\alpha_0 G^2}{\lambda T} \right)^2 \sum_{j=1}^{d} \ln \left( 1 + \frac{\alpha_0 \|g_{t-1,j}\|_2^2}{\lambda T} \right) + \|x_0 - x^*\|_2^2/\alpha_0.$$
5.3 Theoretical Analysis

In this subsection, we always set $\eta_t = \alpha_t$ and assume that $x$ and $\alpha$ are unconstrained, i.e., $X = \mathbb{R}^d$ and $A_t = \mathbb{R}^d_+$. We first demonstrate the monotonicity of both the two update rules from Algorithm 1 and 2 in Section 5.3.1. Afterwards, we discuss the convergence rate of the two update rules in online convex learning setting in Section 5.3.2 and establish a theorem about the regret bounds in Section 5.3.3. Furthermore, we turn to full batch setting with assumption that the objective function $F$ is $L$-smooth but not necessarily convex in Section 5.1.1. Our results for both the settings show that the convergence of our algorithms are robust to the choice of initial learning rates and do not rely on the Lipschitz constant or smoothness constant.

5.3.1 Monotonicity

We point out the monotonicity of learning rate sequences $\{\alpha_t\}$ in our algorithms (proof can be found in Appendix 10).

**Lemma 7.** The sequences $\{\alpha_t\}$ obtained from Algorithm 1 or 2 satisfies $\alpha_{t+1} \leq \alpha_t$.

This phenomenon is common in general training setting like learning rate decay and necessary in several convergence proof including online learning [8, 11] and classical convex optimization [4].

5.3.2 Online Learning Setting

We now establish the result of regrets of Algorithms 1 and 2 in online convex learning, i.e., the $f_t$ are convex. Exactly, we try to bound regrets $\{1\}$ by $O(\sqrt{T})$ for Algorithms 1 and 2. In other words, if $f_t = f$ are the same function, we get a $O(1/\sqrt{T})$ convergent rate.

**Theorem 8.** Suppose that $\varphi \in C_1^1 \left([1, +\infty), \text{ and } \varphi \text { is } \gamma\text{-strongly convex. Assume that } \|g_t\|_\infty \leq G, \|x_t - x^*\|_\infty \leq D_\infty. \text{ Then the sequence } \{x_t\} \text{ obtained from Algorithm 1 satisfies}

$$2R(T) \leq \left(1 + \frac{D^2_\infty}{\gamma^2}\right) \sqrt{2l + 4 \alpha_0^2 G^2 \sum_{j=1}^{d} \|g_{0:T-1,j}\|_2} + \|x_0 - x^*\|_2^2 / \alpha_0,$$

and the sequence $\{x_t\}$ obtained from Algorithm 2 satisfies

$$2R(T) \leq \left(1 + \frac{D^2_\infty}{\gamma^2}\right) \max \left\{\sqrt{2l}, 2\alpha_0 G\right\} \sum_{j=1}^{d} \|g_{0:T-1,j}\|_2 + \|x_0 - x^*\|_2^2 / \alpha_0.$$

Note that under the assumption in Theorem 8, $\sum_{j=1}^{d} \|g_{0:T-1,j}\|_2 \leq dG\sqrt{T}$, hence $R(T) = O(\sqrt{T})$. Our result is comparable to the best known bound for convex online learning problem [9, 17]. We provide a proof sketch here and more detailed proof can be found in Appendix 11.
**proof sketch.** Following from $x_{t+1} = x_t - \text{diag}(\alpha_{t+1})g_t$, we can get

$$2R(T) = 2 \sum_{t=0}^{T-1} (f_t(x_t) - f_t(x_0)) \leq 2 \sum_{t=0}^{T-1} g_t^T (x_t - x^*)$$

$$= \sum_{t=0}^{T-1} \left( \|x_t - x^*\|^2_{B_{t+1}} - \|x_{t+1} - x^*\|^2_{B_t} + \|g_t\|^2_{B_{t+1}} \right)$$

$$\leq \sum_{t=0}^{T-1} \left( \|x_t - x^*\|^2_{(B_{t+1} - B_t)} + \|g_t\|^2_{B_{t+1}} \right) + \beta_0 \|x_0 - x^*\|^2_2$$

$$\leq \sum_{t=0}^{T-1} \left( \|x_t - x^*\|^2 \|\beta_{t+1} - \beta_t\| + \|g_t\|^2_{B_{t+1}} \right) + \beta_0 \|x_0 - x^*\|^2_2,$$

where $\beta_t = 1/\alpha_t$, and $B_t = \text{diag}(\beta_t)$.

For Algorithm 1,

$$\sum_{t=0}^{T-1} (\beta_{t+1,j} - \beta_{t,j}) \leq \frac{1}{\gamma} \sum_{t=0}^{T-1} \frac{g_{t,j}^2}{\beta_{t+1,j}},$$

$$\sum_{t=0}^{T-1} \frac{g_{t,j}^2}{\beta_{t+1,j}} \leq \frac{\sqrt{2l \beta_0^2 + 4G^2}}{\beta_0} \sqrt{\sum_{t=0}^{T-1} g_{t,j}^2}$$

$$= \frac{\sqrt{2l \beta_0^2 + 4G^2}}{\beta_0} \|g_{0:T-1,j}\|_2.$$

Thus

$$2R(T) \leq \left(1 + \frac{D^2}{\gamma} \right) \sqrt{2l \beta_0^2 + 4G^2} \sum_{j=1}^{d} \|g_{0:T-1,j}\|_2$$

$$+ \beta_0 \|x_0 - x^*\|^2_2$$

$$= \left(1 + \frac{D^2}{\gamma} \right) \sqrt{2l + 4\alpha_0^2G^2} \sum_{j=1}^{d} \|g_{0:T-1,j}\|_2$$

$$+ \|x_0 - x^*\|^2_2/\alpha_0.$$

Similarly, for Algorithm 2,

$$\sum_{t=1}^{T-1} (\beta_{t,j} - \beta_{t-1,j}) \leq \frac{1}{\gamma} \sum_{t=0}^{T-1} \frac{g_{t,j}^2}{\beta_{t,j}},$$

$$\sum_{j=1}^{d} \sum_{t=0}^{T-1} \frac{g_{t,j}^2}{\beta_{t+1,j}} \leq \sum_{j=1}^{d} \sum_{t=0}^{T-1} \frac{g_{t,j}^2}{\beta_{t,j}}$$

$$\leq \max \left\{ \sqrt{2l}, \frac{2G}{\beta_0} \right\} \sum_{j=1}^{d} \|g_{0:T-1,j}\|_2.$$
Therefore,
\[
2R(T) \leq \left(1 + \frac{D^2}{\gamma}\right) \max \left\{ \sqrt{2l}, \frac{2G}{\beta_0} \right\} \sum_{j=1}^{d} \|g_{0:T-1,j}\|^2 \\
+ \beta_0 \|x_0 - x^*\|^2 \\
= \left(1 + \frac{D^2}{\gamma}\right) \max \left\{ \sqrt{2l}, 2\alpha_0 G \right\} \sum_{j=1}^{d} \|g_{0:T-1,j}\|^2 \\
+ \|x_0 - x^*\|^2 / \alpha_0.
\]

6 Numerical Experiments

In this paper our principal focus has been to develop a novel approach to adaptively choosing the learning rate during the training process. It would be also interesting to empirically compare our approach with the BB method and Hyper-Gradient Descent in both full batch and online learning settings. Considering the large amount of valid regularization terms, the term is constrained to be generated from the $\varphi$-divergence in the following numerical experiments. For both simplicity and generalization, we merely utilize several common $\varphi$-divergences to derive algorithms, without any delicate design. Experimental results have revealed that these algorithms obtain comparable performance, and even outperform the BB method and Hyper-Gradient Descent in some cases.

6.1 The Set-Up

In the experiments, four common $\varphi$-divergences are used to derive the representative algorithms in Meta-Regularization framework (full implementations are displayed in the Appendix ??):

- $KL(t) = t \log t - t + 1$ leads to KL algorithm.
- $RKL(t) = - \log t + t - 1$ leads to RKL algorithm.
- $Helleinger(t) = (\sqrt{t} - 1)^2$ leads to H algorithm.
- $\chi^2(t) = (t - 1)^2$ leads to $\chi^2$ algorithm.

With any chosen $\varphi$-divergence described above, the corresponding algorithm adopts the update rule described in Algorithm ?? rather than in Algorithm [??]. This mainly comes out of the consideration on computation effectiveness (detailed explanations are displayed in Appendix ??).

To maintain stable performance, the technique of growth clipping is applied to all algorithms in our framework. Actually, growth clipping fulfills the constraints placed on the shrinking speed of the learning rate, which we fully explain in Appendix ?? Specifically, after each update, the updated learning rate can not be smaller than half of the original learning rate.

Numerical experiments involve the above four proposed algorithms as well as the BB method, and Hyper-Gradient Descent algorithms. These algorithms are evaluated on tasks of image classification with a logistic classifier on the
Figure 3: Training process at initial learning rate with least training loss on different algorithms (left: training loss at each training epoch; right: testing accuracy at each training epoch). We repeat our experiments for three times in each curve with different random seeds, and plot shadow error region with two times standard error.

databases of MNIST [19] and CIFAR-10 [18]. Experiments are run using TensorFlow [1], on a machine with Intel Xeon E5-2680 v4 CPU, 128 GB RAM, and NVIDIA Titan Xp GPU.

6.2 Full Batch Setting

We investigate our algorithms in the full batch setting on the MNIST database where algorithms receive the exact gradients of the objective loss function each iteration. The network used in the classifier merely consists of one fully connected layer. The train loss of different algorithms after 50 epochs of training is displayed in Figure 4.

Figure 4: The training of the last training epoch on MNIST at different initial learning rates in full batch setting.

All of the four algorithms derived from our framework are shown to obtain comparable performance with Hyper-Gradient Descent, regardless of the initial learning rate. Moreover, the performance of the BB method is congruously inferior to that of the algorithms from Meta-regularization. Such advantage comes more obvious while the initial learning rate goes larger.
6.3 Online Learning Setting

In the online learning setting, we train a VGG Net \cite{25} with batch normalization on the CIFAR-10 database with a batch size of 128, and an \( \ell_2 \) regularization coefficient of \( 10^{-4} \). We as well perform data augmentation as \cite{15} to improve the training. The train loss as well as test accuracy of different algorithms at different initial learning rates after 100 epochs of training are displayed in Figure\cite{1}.

All of the four algorithms based on Meta-Regularization are shown to obtain comparable performance with the BB methods, exhibiting a relatively low training loss within a large range of initial learning rates. Besides, the advantages of these four algorithms over Hyper-Gradient Descent are obvious in the following two aspects: a generally better convergence performance and a faster convergence speed. From Figure\cite{1}, it is apparent that Hyper-Gradient fails to maintain either a low training loss or a high testing accuracy while the initial learning rate ranging from \( 10^{-2.5} \) to \( 10^{-0.5} \). Specifically, Figure\cite{2} displays the training process at several given learning rates, which conforms to the above observation. For a fair comparison of convergence speed, the initial learning rates with least training loss are respectively fixed for involved algorithms. In Figure\cite{2}, it is obviously observed that the algorithms from Meta-Regularization obtain a comparable convergence performance but a faster convergence speed than Hyper-Gradient Descent, in terms of both training loss and testing accuracy.

References

\[\begin{align*}
\text{[1]} & \quad \text{Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. Tensorflow: a system for large-scale machine learning. In OSDI, volume 16, pages 265–283, 2016.} \\
\text{[2]} & \quad \text{Jonathan Barzilai and Jonathan M Borwein. Two-point step size gradient methods. IMA journal of numerical analysis, 8(1):141–148, 1988.} \\
\text{[3]} & \quad \text{Atilim Gunes Baydin, Robert Cornish, David Martinez Rubio, Mark Schmidt, and Frank Wood. Online learning rate adaptation with hypergradient descent. In International Conference on Learning Representations, 2018.} \\
\text{[4]} & \quad \text{Sébastien Bubeck et al. Convex optimization: Algorithms and complexity. Foundations and Trends® in Machine Learning, 8(3-4):231–357, 2015.} \\
\text{[5]} & \quad \text{Pedro Cruz. Almost sure convergence and asymptotical normality of a generalization of kesten’s stochastic approximation algorithm for multidimensional case. arXiv preprint arXiv:1105.5231, 2011.} \\
\text{[6]} & \quad \text{Yu-Hong Dai. A new analysis on the barzilai-borwein gradient method. Journal of the operations Research Society of China, 1(2):187–198, 2013.} \\
\text{[7]} & \quad \text{Ingrid Daubechies, Ronald DeVore, Massimo Fornasier, and C Sinan Güntürk. Iteratively reweighted least squares minimization for sparse recovery. Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences, 63(1):1–38, 2010.} \\
\text{[8]} & \quad \text{Ofer Dekel, Ran Gilad-Bachrach, Ohad Shamir, and Lin Xiao. Optimal distributed online prediction using mini-batches. Journal of Machine Learning Research, 13(Jan):165–202, 2012.} \\
\text{[9]} & \quad \text{John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research, 12(Jul):2121–2159, 2011.} \\
\text{[10]} & \quad \text{Roger Fletcher. Practical methods of optimization. John Wiley & Sons, 2013.} \\
\text{[11]} & \quad \text{Saeed Ghadimi and Guanghui Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. SIAM Journal on Optimization, 23(4):2341–2368, 2013.} \\
\text{[12]} & \quad \text{Vineet Gupta, Tomer Koren, and Yoram Singer. A unified approach to adaptive regularization in online and stochastic optimization. arXiv preprint arXiv:1706.06569, 2017.} \\
\text{[13]} & \quad \text{Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. Machine Learning, 69(2-3):169–192, 2007.} \\
\text{[14]} & \quad \text{Elad Hazan et al. Introduction to online convex optimization. Foundations and Trends® in Optimization, 2(3-4):157–325, 2016.} \\
\text{[15]} & \quad \text{Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In Proceedings of the IEEE conference on computer vision and pattern recognition, pages 770–778, 2016.} \\
\text{[16]} & \quad \text{Harry Kesten et al. Accelerated stochastic approximation. The Annals of Mathematical Statistics, 29(1):41–59, 1958.}
\end{align*}\]
[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In International Conference on Learning Representations, 2015.

[18] Alex Krizhevsky and Geoffrey Hinton. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

[19] Yann LeCun, Corinna Cortes, and CJ Burges. Mnist handwritten digit database. AT&T Labs [Online]. Available: http://yann.lecun.com/exdb/mnist, 2, 2010.

[20] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. Mathematical programming, 45(1-3):503–528, 1989.

[21] Mahesh Chandra Mukkamala and Matthias Hein. Variants of RMSProp and Adagrad with logarithmic regret bounds. In Doina Precup and Yee Whye Teh, editors, Proceedings of the 34th International Conference on Machine Learning, Proceedings of Machine Learning Research. PMLR, 06–11 Aug 2017.

[22] Yurii Nesterov. Introductory lectures on convex optimization: A basic course, volume 87. Springer Science & Business Media, 2013.

[23] Jorge Nocedal and Stephen J Wright. Numerical optimization 2nd, 2006.

[24] Shai Shalev-Shwartz et al. Online learning and online convex optimization. Foundations and Trends® in Machine Learning, 4(2):107–194, 2012.

[25] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[26] Conghui Tan, Shiqian Ma, Yu-Hong Dai, and Yuqiu Qian. Barzilai-borwein step size for stochastic gradient descent. In Advances in Neural Information Processing Systems, pages 685–693, 2016.

[27] Tijmen Tieleman and Geoffrey Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. COURSERA: Neural networks for machine learning, 4(2):26–31, 2012.

[28] X. Wu, R. Ward, and L. Bottou. WNGrad: Learn the Learning Rate in Gradient Descent. ArXiv e-prints, March 2018.

[29] Ya-xiang Yuan. Step-sizes for the gradient method. AMS IP Studies in Advanced Mathematics, 42(2):785, 2008.

[30] M. D. Zeiler. ADADELTA: An Adaptive Learning Rate Method. ArXiv e-prints, December 2012.

[31] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In Proceedings of the 20th International Conference on Machine Learning (ICML-03), pages 928–936, 2003.
7 Solution Existence

Note that the function $h(1/\alpha) = 1/\alpha^2 \varphi'(\eta_{t,j}/\alpha)$ is an increasing continuous function and $\lim_{z \to +\infty} \varphi'(z) = +\infty$. For the same reason, the solution of (9) exists.

8 Special Cases of Algorithm 1

In this section, We will point out that Adagrad [9] and WNGrad [28] are special cases of Algorithm 1. If we set $\varphi(z) = z + \frac{1}{z} - 2$, then the new learning rate $1/\alpha_{t+1}$ can be obtained by

$$\frac{1}{\alpha_{t+1}^2} \left(1 - \frac{\alpha_{t+1,j}^2}{\alpha_{t,j}^2}\right) = g_{t,j}^2, \quad j = 1, \ldots, d,$$

that implies,

$$\frac{1}{\alpha_{t+1}^2} = \frac{1}{\alpha_t^2} + g_{t,j}^2,$$

and we drive AdaGrad from Meta-Regularization.

Similarly, we can get WNGrad by setting $\varphi(z) = \frac{1}{z} - \log \frac{1}{z} - 1$. In fact, $1/\alpha_{t+1}$ employs update

$$\frac{1}{\alpha_{t+1}^2} \left(\frac{1/\alpha_{t,j}(1/\alpha_{t+1,j} - 1/\alpha_{t,j})}{1/\alpha_{t+1,j}^2}\right) = g_{t,j}^2, \quad j = 1, \ldots, d,$$

on the other words,

$$\frac{1}{\alpha_{t+1}} = \frac{1}{\alpha_t} + \alpha_t g_{t,j}^2,$$

i.e., the update rule of WNGrad.

9 Max-min or min-max

Lemma 9. Suppose that $A_t = [b_{t,1}, B_{t,1}] \times \cdots \times [b_{t,d}, B_{t,d}]$, and $X = \mathbb{R}^d$. Let $\alpha^*$ be the solution of unconstrained problem $\max_{\alpha \in \mathcal{B}_1} (\min_{x} \Psi_t(x, \alpha))$. Then the solution of problem $\max_{\alpha \in \mathcal{B}_1} (\min_{x} \Psi_t(x, \alpha))$ is

$$\alpha_j = \min \{ \max \{ \alpha^*_j, b_{t,j} \}, B_{t,j} \}, \quad \text{for } j = 1, \ldots, d.$$

Proof. First, it is trivial to get

$$\Psi_{t,x}(\alpha) \triangleq \min_{x} \Psi_t(x, \alpha) = \Psi_t(x_t - \alpha \circ g_t, \alpha)$$

$$= -\frac{1}{2} \| g_t \|_{\text{diag}(\alpha)}^2 - \frac{1}{2} D_{\varphi}(\alpha, \eta_t)$$

$$= -\frac{1}{2} \sum_{j=1}^d \left( \alpha_j g_{t,j}^2 + \frac{1}{\eta_{t,j}} \varphi' \left( \frac{\eta_{t,j}}{\alpha_j} \right) \right).$$

The partial derivative of $\Psi_{t,x}(\alpha)$ with respect to $\alpha_j$ is

$$\frac{\partial \Psi_{t,x}(\alpha)}{\partial \alpha_j} = -\frac{1}{2} \left( g_{t,j}^2 - \frac{1}{\alpha_j^2} \varphi' \left( \frac{\eta_{t,j}}{\alpha_j} \right) \right).$$

Note that $\varphi$ is a convex function, so $\varphi'$ is a non-decreasing function, and $\frac{\partial \Psi_{t,x}(\alpha)}{\partial \alpha_j}$ is a non-increasing function. Recall that $\alpha^*$ be the solution of unconstrained problem $\max_{\alpha \in \mathcal{B}_1} (\min_{x} \Psi_t(x, \alpha))$, hence, $\alpha^*_j$ is a zero of function $\frac{\partial \Psi_{t,x}(\alpha)}{\partial \alpha_j}$.

Moreover, if $\alpha^*_j > B_{t,j}$, we have $\frac{\partial \Psi_{t,x}(\alpha)}{\partial \alpha_j} \geq 0$. Thus, $\Psi_{t,x}(\alpha)$ with respect to $\alpha_j$ is a non-increasing function, and arg $\max_{\alpha_j} \Psi_{t,x}(\alpha) = B_{t,j}$. For a similar reason, if $\alpha^*_j < b_{t,j}$, then arg $\max_{\alpha_j} \Psi_{t,x}(\alpha) = b_{t,j}$. In conclusion,

$$\arg \max_{\alpha_j \in [b_{t,j}, B_{t,j}]} \Psi_{t,x}(\alpha) = \min \{ \max \{ \alpha^*_j, b_{t,j} \}, B_{t,j} \}, \quad \text{for } j = 1, \ldots, d.$$

☐
10 Monotonicity

In this section, We provide the proof of Lemma \[ \text{Lemma 10.} \] Denote that \( \Psi_t(x, \alpha) = \min_{x \in \mathcal{X}} \Psi_t(x, \alpha) \).

\text{Lemma 10.} \( \alpha_{t+1} \) obtained from \text{equation (9)} satisfies \( \alpha_{t+1} \leq \eta_t \).

\text{Proof.} Recall that \( \varphi(1) = \varphi'(1) = 0 \), so \( \varphi(x) \geq 0 \) for all \( x \) and \( D_\varphi(\alpha, \eta_t) = \varphi(\eta_t/\alpha)/\eta_t \geq 0 \). If \( \alpha > \eta_t \), then for all \( x \in \mathcal{X} \)

\[ \Psi_t(x, \alpha) = g_t^T(x - x_t) + \frac{1}{2\alpha} \|x - x_t\|^2 - \frac{1}{2} D_\varphi(\alpha, \eta_t) \]
\[ < g_t^T(x - x_t) + \frac{1}{2\eta_t} \|x - x_t\|^2 \]
\[ = \Psi_t(x, \eta_t). \]

Hence, \( \min_{x \in \mathcal{X}} \Psi_t(x, \alpha) < \min_{x \in \mathcal{X}} \Psi_t(x, \eta_t) \), i.e., \( \Psi_t(x, \alpha) < \Psi_t(x, \eta_t) \).

It means \( \alpha_{t+1} = \arg \max_{\alpha \in \mathcal{A}} \Psi_t(x, \alpha) \leq \eta_t \).

\text{Lemma 11.} \( \alpha_{t+1} \) obtained from \text{equation (10)} satisfies \( \alpha_{t+1} \leq \eta_t \).

\text{Proof.} Let \( y = \arg \min_x \Psi_t(x, \eta_t) \). If \( \alpha > \eta_t \), then

\[ \Psi_t(y, \alpha) = g_t^T(y - x_t) + \frac{1}{2\alpha} \|y - x_t\|^2 - \frac{1}{2} D_\varphi(\alpha, \eta_t) \]
\[ < g_t^T(y - x_t) + \frac{1}{2\eta_t} \|y - x_t\|^2 \]
\[ = \Psi_t(y, \eta_t). \]

Hence \( \alpha_{t+1} = \arg \max_{\alpha \in \mathcal{A}} \Psi_t(y, \alpha) \leq \eta_t \).

11 Regrets in online learning setting

Recall the definition of regret

\[ R(T) = \sum_{t=0}^{T-1} (f_t(x_t) - f_t(x^*)) \]

where \( x^* = \arg \min_{x \in \mathcal{X}} \sum_{t=0}^{T-1} f_t(x) \). We show our Algorithms derived from meta-regularization have \( O(\sqrt{T}) \) regret bounds.

\text{Lemma 12.} Consider an arbitrary real-valued sequence \( \{a_i\} \) and its vector representation \( a_{1:t} = (a_1, \cdots, a_t)^T \). Then

\[ \sum_{i=1}^{T} \frac{a_i^2}{\|a_{1:t}\|^2} \leq 2\|a_{1:T}\|^2 \]

(16)

holds.

\text{Proof.} Let us use induction on \( T \) to prove inequality (12). For \( T = 1 \), the inequality trivially holds. Assume the bound (16) holds true for \( T - 1 \), in which case

\[ \sum_{i=1}^{T} \frac{a_i^2}{\|a_{1:t}\|^2} \leq 2\|a_{1:T-1}\|^2 + \frac{a_T^2}{\|a_{1:T}\|^2}. \]

We denote \( b_T = \sum_{t=1}^{T} a_t^2 \) and have

\[ 2\|a_{1:T-1}\|^2 + \|a_{1:T}\|^2 \]
\[ = \frac{a_T^2}{\|a_{1:T}\|^2} + \frac{a_T^2}{\sqrt{b_T}} \]
\[ \leq \frac{a_T^2}{\|a_{1:T-1}\|^2} + \frac{a_T^2}{\|a_{1:T}\|^2} \]
\[ = 2\sqrt{b_T}. \]

\[ \square \]
Lemma 13. Suppose the sequence \( \{x_t\} \) and sequence \( \{\alpha_t\} \) satisfy \( x_{t+1} = x_t - \alpha_{t+1} \circ g_t \). Then the regret satisfies

\[
2R(T) \leq \sum_{t=0}^{T-1} \| g_t \|^2_{\text{diag}(\alpha_{t+1})} + \sum_{t=0}^{T-1} \| x_t - x^* \|^2_{\text{diag}(\alpha_{t+1})^{-1}} + \| x_0 - x^* \|^2_{\text{diag}(\alpha_0)^{-1}}.
\]

**Proof.** Note that

\[
x_{t+1} = x_t - \text{diag}(\alpha_{t+1}) g_t,
\]

and

\[
\| x_{t+1} - x^* \|^2_{\text{diag}(\alpha_{t+1})^{-1}} = \| x_t - x^* - \text{diag}(\alpha_{t+1}) g_t \|^2_{\text{diag}(\alpha_{t+1})^{-1}} = \| x_t - x^* \|^2_{\text{diag}(\alpha_{t+1})^{-1}} + \| g_t \|^2_{\text{diag}(\alpha_{t+1})} - 2g_t^T (x_t - x^*),
\]

i.e.,

\[
2g_t^T (x_t - x^*) = \| g_t \|^2_{\text{diag}(\alpha_{t+1})} + \left( \| x_t - x^* \|^2_{\text{diag}(\alpha_{t+1})^{-1}} - \| x_{t+1} - x^* \|^2_{\text{diag}(\alpha_{t+1})^{-1}} \right).
\]

Hence

\[
2R(T) = 2 \sum_{t=0}^{T-1} (f_t(x_t) - f_t(x^*))
\]

\[
\leq 2 \sum_{t=0}^{T-1} g_t^T (x_t - x^*)
\]

\[
= \sum_{t=0}^{T-1} \| g_t \|^2_{\text{diag}(\alpha_{t+1})} + \sum_{t=0}^{T-1} \left( \| x_t - x^* \|^2_{\text{diag}(\alpha_{t+1})^{-1}} - \| x_{t+1} - x^* \|^2_{\text{diag}(\alpha_{t+1})^{-1}} \right)
\]

\[
\leq \sum_{t=0}^{T-1} \| g_t \|^2_{\text{diag}(\alpha_{t+1})} + \sum_{t=0}^{T-1} \| x_t - x^* \|^2_{\text{diag}(\alpha_{t+1})^{-1}} + \| x_0 - x^* \|^2_{\text{diag}(\alpha_0)^{-1}}.
\]

\[
\square
\]

Lemma 14. Suppose an increasing function \( \psi \) satisfies \( \psi(1) = 0 \) and \( \psi(x) \leq l(x-1) \). Consider a real valued sequence \( \{g_t\}_{t=0}^{T-1} \) and a positive sequence \( \{\beta_t\}_{t=0}^{T} \) which satisfies \( \| g_t \| \leq G, \beta^2_{t+1} \psi \left( \frac{\beta_{t+1}}{\beta_t} \right) = g_t^2, t = 0, \cdots, T-1, \beta_0 \geq 0 \). We can bound \( \beta_T \) as

\[
\beta_t \geq c \sqrt{\beta_0^2 + \frac{2}{T} \sum_{i=0}^{t-1} g_i^2}, t = 1, \cdots, T
\]

where \( c = \sqrt{\frac{\beta_0^2}{2G^2}} \). Moreover, we have

\[
\sum_{t=0}^{T-1} g_t^2 \beta_{t+1}^2 \leq \frac{2l\beta_0^2 + 4G^2}{\beta_0} \sum_{t=0}^{T-1} g_t^2.
\]

**Remark 5.** We point out that

- \( \beta_{t+1} \geq \beta_t \) (If \( \beta_{t+1} < \beta_t \), then \( \beta^2_{t+1} \psi (\beta_{t+1}/\beta_t) < 0 \leq g_t^2 \)).

- \( \beta_{t+1} \) is unique with respect to \( \beta_t \) due to the fact that the function \( \hat{\psi}(\beta) = \beta^2 \psi(\beta/\beta_0) \) is strictly increasing.

**Proof.** Assume that \( \beta_t \geq c \sqrt{\beta_0^2 + \frac{2}{T} \sum_{i=0}^{t-1} g_i^2} \), where \( c > 0 \) is a variable coefficient.

Let us find out a specific \( c \) such that \( \beta_{t+1} \geq c \sqrt{\beta_0^2 + \frac{2}{T} \sum_{i=0}^{t-1} g_i^2} \).

Note that

\[
\beta^2_{t+1} \psi \left( \frac{\beta_{t+1}}{\beta_t} \right) \leq l \beta^2_{t+1} \left( \frac{\beta_{t+1}}{\beta_t} - 1 \right).
\]

15
Define a cubic polynomial

\[ h(\beta) = \frac{l}{\beta_t} \beta^3 - l\beta^2 - g_t^2, \]

and \( h \) is an increasing function when \( \beta \geq \beta_t \).

If \( h\left(c \sqrt{\beta_0^2 + \frac{2}{l} \sum_{i=0}^{t-1} g_t^2}\right) \leq 0 \), according to \( h(\beta_{t+1}) \geq 0 \), then \( \beta_{t+1} \geq c \sqrt{\beta_0^2 + \frac{2}{l} \sum_{i=0}^{t} g_t^2} \).

Denote \( b = \beta_0^2 + \frac{2}{l} \sum_{i=0}^{t-1} g_t^2 \). So we just need to choose \( c \) such that

\[ h\left(c \sqrt{\beta_0^2 + \frac{2}{l} \sum_{i=0}^{t} g_t^2}\right) \leq c^2 (b + 2g_t^2/l) \left( \frac{\sqrt{b + 2g_t^2/l}}{\sqrt{b}} - 1 \right) - g_t^2 \leq 0, \]

where the first inequality holds for the assumption \( \beta_t \geq c \sqrt{\beta_0^2 + \frac{2}{l} \sum_{i=0}^{t-1} g_t^2} \), or

\[ \frac{c^2}{\sqrt{b}} (b + 2g_t^2/l) \frac{2g_t^2/l}{\sqrt{b + 2g_t^2/l} + \sqrt{b}} \leq g_t^2/l, \]

or

\[ \frac{2c^2}{\sqrt{b}} (b + 2g_t^2/l) \leq \sqrt{b + 2g_t^2/l} + \sqrt{b}. \]

Thus, \( c \) just need to satisfy

\[ c^2 \leq \frac{b}{b + 2g_t^2/l}. \]

According to \( b \geq \beta_0^2, g_t^2 \leq G^2 \), hence

\[ \frac{b}{b + 2g_t^2/l} \geq \frac{\beta_0^2}{\beta_0^2 + 2G^2/l}. \]

So if we choose \( c = \sqrt{\frac{\beta_0^2}{\beta_0^2 + 2G^2/l}} \), then \( \beta_t > \beta_0 > c\beta_0 \), hence

\[ \beta_t \geq c \sqrt{\beta_0^2 + \frac{2}{l} \sum_{i=0}^{t-1} g_t^2}, \quad t = 1, \ldots, T. \]

Moreover, following from Lemma 12, we have

\[ \sum_{t=0}^{T-1} \frac{g_t^2}{\beta_{t+1}} \leq \sum_{t=0}^{T-1} \frac{g_t^2}{c \sqrt{2/l} \sqrt{\sum_{i=0}^{t} g_t^2}} \leq \frac{\sqrt{2l}}{c} \sqrt{\sum_{t=0}^{T-1} g_t^2}. \]

**Lemma 15.** Suppose an increasing function \( \psi \) satisfies \( \psi(1) = 0 \) and \( \psi(x) \leq l(x-1) \). Consider a real valued sequence \( \{g_t\}_{t=0:T-1} \) and a positive sequence \( \{\beta_t\}_{t=0:T} \) which satisfies \( |g_t| \leq G, \beta_t^2 \psi \left( \frac{\beta_{t+1}}{\beta_t} \right) = g_t^2, \quad t = 0, \ldots, T-1, \beta_0 \geq 0 \).

We can bound \( \beta_T \) as

\[ \beta_t \geq \sqrt{\beta_0^2 + \frac{2}{l} \sum_{i=0}^{t-1} g_t^2}, \quad t = 1, \ldots, T. \quad (21) \]

Moreover, we have

\[ \sum_{t=0}^{T-1} \frac{g_t^2}{\beta_t} \leq \max \left\{ \sqrt{2l}, \left( \frac{2G}{\beta_0} \right)^{\frac{1}{2}} \right\} \sqrt{\sum_{t=0}^{T-1} g_t^2}. \quad (22) \]
Theorem 16. Suppose that $\varphi \in C_i^{1,1}([1, +\infty))$, and $\varphi$ is $\gamma$-strongly convex. Assume that $\|g_t\|_\infty \leq G$, and $\|x_t - x^*\|_\infty \leq D_\infty$. Then the sequence $\{x_t\}$ obtained from Algorithm 1 satisfies

$$2R(T) \leq \left(1 + \frac{D_\infty^2}{\gamma}\right)\sqrt{2l + 4\alpha_0^2G^2} \sum_{j=1}^{d} \|g_{0:T-1,j}\|_2 + \|x_0 - x^*\|_2^2/\alpha_0.$$  

Proof. Let $\beta_t = 1/\alpha_t$. Following from Lemma 13

$$2R(T) \leq \sum_{t=0}^{T-1} \|g_t\|^2_{\text{diag}(\alpha_{t+1}^{-1})} + \sum_{t=0}^{T-1} \|x_t - x^*\|^2_{\text{diag}(\alpha_{t+1}^{-1} - \alpha_t)} + \|x_0 - x^*\|^2_{\text{diag}(\alpha_0)}$$

$$\leq \sum_{t=0}^{T-1} \|g_t\|^2_{\text{diag}(\beta_{t+1}^{-1})} + \sum_{t=0}^{T-1} \|x_t - x^*\|^2_{\text{diag}(\beta_{t+1}^{-1} - \beta_t)} + \|x_0 - x^*\|^2_{\text{diag}(\beta_0)}$$

$$\leq \sum_{t=0}^{T-1} \sum_{j=1}^{d} g_{t,j}^2 \beta_{t+1,j} + \max_{0 \leq t < T} \|x_t - x^*\|_\infty^2 \sum_{t=0}^{T-1} \sum_{j=1}^{d} (\beta_{t+1,j} - \beta_{t,j}) + \|x_0 - x^*\|_{\text{diag}(\beta_0)}.$$  

Recall $\varphi$ is a $\gamma$-strongly convex function, and $\varphi'(\alpha_{t+1,j}/\alpha_{t,j}) = \alpha_{t+1,j}^2 g_{t,j}^2$.

so,

$$g_{t,j}^2 = \beta_{t+1,j}^2 \varphi' \left( \frac{\beta_{t+1,j}}{\beta_{t,j}} \right) \geq \gamma \beta_{t+1,j} \beta_{t,j} \left( \frac{\beta_{t+1,j}}{\beta_{t,j}} - 1 \right),$$

and

$$\sum_{t=0}^{T-1} (\beta_{t+1,j} - \beta_{t,j}) \leq \frac{1}{\gamma} \sum_{t=0}^{T-1} g_{t,j}^2 \beta_{t+1,j}.$$  

(23)

The function $\psi = \varphi'$ satisfies $\psi(1) = 0$ and $\psi(x) \leq l(x-1)$ according to the smoothness of $\varphi$. Following from Lemma 14 we have

$$\sum_{t=0}^{T-1} g_{t,j}^2 \beta_{t+1,j} \leq \sqrt{2l/\beta_{0,j}^2 + 4G^2} \sum_{t=0}^{T-1} g_{t,j}^2 \beta_{0,j} \leq \sqrt{2l/\beta_{0,j}^2 + 4G^2} \sum_{t=0}^{T-1} g_{0:T-1,j}.$$  

(24)
Combining inequality 23 and 24, we have

\[ 2R(T) \leq \left( 1 + \frac{\max_{0 \leq t < T} \| x_t - x^* \|_2^2}{\gamma} \right) \sum_{j=1}^{d} \sum_{t=0}^{T-1} \frac{g_{t,j}^2}{\beta_{t+1,j}} \| x_0 - x^* \|_{\text{diag}(\beta_0)}^2 \]

\[ \leq \left( 1 + \frac{D^2}{\gamma} \right) \sum_{j=1}^{d} \sum_{t=0}^{T-1} \frac{\sqrt{2l \beta^2_{0,j} + 4G^2}}{\beta_{0,j}} \| g_{0:T-1,j} \|_2 + \| x_0 - x^* \|_{\text{diag}(\beta_0)}^2 \]

\[ = \left( 1 + \frac{D^2}{\gamma} \right) \sum_{j=1}^{d} \sum_{t=0}^{T-1} \frac{\sqrt{2l \beta^2_{0,j} + 4G^2}}{\beta_{0,j}} \| g_{0:T-1,j} \|_2 + \| x_0 - x^* \|_2^2 / \alpha_0. \]

\[ \square \]

**Theorem 17.** Suppose that \( \varphi \in C^{1,1}_{\gamma}([1, +\infty]), \) and \( \varphi \) is \( \alpha \)-strongly convex. Assume that \( \| g_t \|_{\infty} \leq G, \) and \( \| x_t - x^* \|_{\infty} \leq D_\infty. \) Then the sequence \( \{ x_t \} \) obtained from Algorithm 2 satisfies

\[ 2R(T) \leq \left( 1 + \frac{D^2}{\gamma} \right) \max \left\{ \sqrt{2l}, 2\alpha G \right\} \sum_{j=1}^{d} \| g_{0:T-1,j} \|_2 + \| x_0 - x^* \|_2^2 / \alpha_0. \]

**Proof.** Let \( \beta_t = 1 / \alpha_t. \) Similar to the proof of Theorem 16 for Algorithm 2, we have

\[ 2R(T) \leq \sum_{t=0}^{T-1} \sum_{j=1}^{d} \frac{g_{t,j}^2}{\beta_{t+1,j}} + \max_{0 \leq t < T} \| x_t - x^* \|_2^2 \sum_{t=0}^{T-1} \sum_{j=1}^{d} \left( \beta_{t+1,j} - \beta_{t,j} \right) + \| x_0 - x^* \|_{\text{diag}(\beta_0)}^2 \]

\[ \leq \sum_{t=0}^{T-1} \sum_{j=1}^{d} \frac{g_{t,j}^2}{\beta_{t,j}} + \max_{0 \leq t < T} \| x_t - x^* \|_2^2 \sum_{t=0}^{T-1} \sum_{j=1}^{d} \left( \beta_{t+1,j} - \beta_{t,j} \right) + \| x_0 - x^* \|_{\text{diag}(\beta_0)}^2. \]

Note that in Algorithm 2, \( \alpha^2_{t,j} g_{t,j} = \varphi'(\alpha_{t,j} / \alpha_{t+1,j}), \) thus

\[ g_{t,j}^2 = \beta_{t,j}^2 \varphi' \left( \frac{\beta_{t+1,j}}{\beta_{t,j}} \right) \geq \gamma \beta_{t,j}^2 \left( \frac{\beta_{t+1,j}}{\beta_{t,j}} - 1 \right), \]

and

\[ \sum_{t=0}^{T-1} \left( \beta_{t,j} - \beta_{t-1,j} \right) \leq \sum_{t=0}^{T-1} \left( \beta_{t+1,j} - \beta_{t,j} \right) \leq \frac{1}{\gamma} \sum_{t=0}^{T-1} \frac{g_{t,j}^2}{\beta_{t,j}}. \]

Thus, following from Lemma 15 and similar reason in our proof of Theorem 16, we have

\[ 2R(T) \leq \left( 1 + \frac{D^2}{\gamma} \right) \sum_{j=1}^{d} \sum_{t=0}^{T-1} \frac{g_{t,j}^2}{\beta_{t,j}} + \beta_0 \| x_0 - x^* \|_2^2 \]

\[ \leq \left( 1 + \frac{D^2}{\gamma} \right) \max \left\{ \sqrt{2l}, \frac{2G}{\beta_0} \right\} \sum_{j=1}^{d} \| g_{0:T-1,j} \|_2 + \beta_0 \| x_0 - x^* \|_2^2 \]

\[ = \left( 1 + \frac{D^2}{\gamma} \right) \max \left\{ \sqrt{2l}, 2\alpha_0 G \right\} \sum_{j=1}^{d} \| g_{0:T-1,j} \|_2 + \| x_0 - x^* \|_2^2 / \alpha_0. \]

\[ \square \]
12 Logarithmic Bounds

In this section, we will use a different class of ‘distance’ function for problem (3), and establish logarithmic regret bounds under assumption \(f_t\) is strongly convex. Our analysis and proof follow from \([13, 21]\).

First, we define \(\mu\)-strongly convexity.

**Definition 18** (Definition 2.1 in \([21]\)). Let \(\mathcal{X} \subseteq \mathbb{R}^d\) be a convex set. We say that a function \(f : \mathcal{X} \to \mathbb{R}\) is \(\mu\)-strongly convex, if there exists \(\mu \in \mathbb{R}_+^d\) with \(\mu_j > 0\) for \(j = 1, \cdots, d\) such that for all \(x, y \in \mathcal{X}\),

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{1}{2}\|y - x\|_{\text{diag}(\mu)}^2.
\]

Let \(\xi = \min_{j=1:d} \mu_j\), then this function is \(\xi\)-strongly convex (in the usual sense), that is

\[
f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\xi}{2}\|y - x\|_2^2.
\]

The modification SC-Meta-Regularization of Meta-Regularization which we propose in the following uses a family of distance function \(D: \mathbb{R}^d_+ \times \mathbb{R}^d_+ \to \mathbb{R}\) formulated as

\[
D(u, v) = \sum_{j=1}^d \varphi(v_j / u_j),
\]

where \(\varphi\) is convex function with \(\varphi(1) = \varphi'(1) = 0\) like we used in \(\varphi\)-divergence.

**Remark 6.** Same as \(\varphi\)-divergence, \(D(u, v) \geq 0\) for any \(u, v \in \mathbb{R}^d_+\).

Different from Algorithm \([1, 2]\) we add a hyper-parameter \(\lambda > 0\) like AdaGrad to SC-AdaGrad. Rewrite problem (3) as

\[
\max_{\alpha \in \mathcal{A}} \min_{x \in \mathcal{X}} \Psi_t(x, \alpha) \triangleq g_t^T (x - x_t) + \frac{1}{2}\|x - x_t\|_{\text{diag}(\alpha)}^2 - \frac{\lambda}{2} \sum_{j=1}^d \varphi(\alpha_{t,j}/\alpha_j).
\]

Similarly, we can also derive two algorithms according to two update rules respectively.

**Algorithm 4** GD with SC-Meta-Regularization (Algorithm \([3]\) in Section 5.2)

**Require:** \(\alpha_0 > 0, x_0\)

1: for \(t = 1\) to \(T\) do
2: Suffer loss \(f_t(x_t)\);
3: Receive \(g_t \in \partial f_t(x_t)\) of \(f_t\) at \(x_t\);
4: Update \(\alpha_{t+1, j}\) as the solution of the equation \(\lambda(\alpha_{t,j}/\alpha^2)\varphi'(\alpha_{t,j}/\alpha) = g_{t,j}, \forall j = 1, \cdots, d\);
5: Update \(x_{t+1} = x_t - \alpha_{t+1} g_t\);
6: end for

**Algorithm 5** GD with SC-Meta-Regularization using alternating update rule

**Require:** \(\alpha_0 > 0, x_0\)

1: for \(t = 1\) to \(T\) do
2: Suffer loss \(f_t(x_t)\);
3: Receive \(g_t \in \partial f_t(x_t)\) of \(f_t\) at \(x_t\);
4: Update \(\alpha_{t+1} = \alpha_{t,j} / (\varphi')^{-1}(\alpha_{t,j} g_{t,j}^2 / \lambda), j = 1, \cdots, d\);
5: Update \(x_{t+1} = x_t - \alpha_{t+1} g_t\);
6: end for

**Remark 7.** Same as Lemma \([4]\) the monotonicity of Algorithm \([2]\) and \([3]\) also holds.

**Theorem 19.** Suppose that \(f_t\) is \(\mu\)-strongly convex for all \(t\), \(\varphi \in C^{1,1}_\text{Lips}([1, +\infty))\), and \(\varphi\) is \(\gamma\)-strongly convex. Assume that \(\|g_t\|_\infty \leq G, \lambda \geq G^2 / (\gamma \min_{j=1:d} \mu_j)\). Then the sequence \(\{x_t\}\) obtained from Algorithm \([4]\) satisfies

\[
2R(T) \leq l \left(1 + \frac{\alpha_0 G^2}{\lambda}\right)^2 \sum_{j=1}^d \ln \left(1 + \frac{\alpha_0 \|g_{0:T-1,j}\|_2^2}{\lambda}\right) + \|x_0 - x^*\|_2^2 / \alpha_0,
\]

19
and the sequence \( \{x_t\} \) obtained from Algorithm \([2\, 5]\) satisfies
\[
2R(T) \leq l \sum_{j=1}^{d} \ln \left( 1 + \frac{\alpha_0 \|g_{0:T-1,j}\|_2^2}{\lambda} \right) + \|x_0 - x^*\|_2^2 / \alpha_0.
\]

**Remark 8.** Under assumption in Theorem \([19]\) we have \( \|g_{0:T-1,j}\|_2^2 \leq G^2T \), so \( R(T) = O(\ln(T)) \).

To prove Theorem \([19]\) we first prove following lemma.

**Lemma 20.** For an arbitrary real-valued sequence \( \{a_i\} \) and a positive real number \( b \),
\[
\sum_{t=1}^{T} \frac{a_t^2}{b + \sum_{i=1}^{t} a_i^2} \leq \ln \left( 1 + \frac{\sum_{t=1}^{T} a_t^2}{b} \right).
\]

**Proof.** Let \( b_0 = b, b_i = b + \sum_{i=1}^{t} a_i^2, t \geq 1 \), then
\[
\sum_{t=1}^{T} \frac{a_t^2}{b + \sum_{i=1}^{t} a_i^2} = \sum_{t=1}^{T} \frac{b_t - b_{t-1}}{b_t} = \sum_{t=1}^{T} \int_{b_{t-1}}^{b_t} \frac{1}{x} \, dx \\
\leq \sum_{t=1}^{T} \int_{b_{t-1}}^{b_t} \frac{1}{x} \, dx = \int_{b}^{b_T} \frac{1}{x} \, dx = \ln \left( 1 + \frac{\sum_{t=1}^{T} a_t^2}{b} \right).
\]

Like Lemma \([14\, 15]\) similar lemma holds for Algorithm \([4\, 5]\).

**Lemma 21.** Suppose an increasing function \( \psi \) satisfies \( \psi(1) = 0 \) and \( \psi(x) \leq l(x-1) \). Consider a real valued sequence \( \{g_t\}_{t=T-1}^{0} \) and a positive sequence \( \{\beta_t\}_{t=0}^{T} \) which satisfies \( |g_t| \leq G, \beta_0 > 0 \).

If \( (\beta_{t+1}/\beta_t)\psi(\beta_{t+1}/\beta_t) = g_t^2, t = 0, \cdots, T-1, \) then we have
\[
\beta_t \geq \left( \frac{\beta_0}{\beta_0 + G^2/l} \right)^2 \left( \beta_0 + \frac{1}{l} \sum_{i=0}^{t-1} g_i^2 \right), \; t = 1, \cdots, T \tag{28}
\]
and
\[
\sum_{t=0}^{T-1} \frac{g_t^2}{\beta_{t+1}} \leq \ln \left( \frac{\beta_0 + G^2/l}{\beta_0} \right)^2 \left( 1 + \frac{\sum_{t=0}^{T-1} g_t^2}{l\beta_0} \right) \tag{29}
\]

Meanwhile, if \( \beta_t\psi(\beta_{t+1}/\beta_t) = g_t^2, t = 0, \cdots, T-1, \) then we have
\[
\beta_t \geq \beta_0 + \frac{1}{l} \sum_{i=0}^{t-1} g_i^2, \; t = 1, \cdots, T \tag{30}
\]
and
\[
\sum_{t=0}^{T-1} \frac{g_t^2}{\beta_{t+1}} \leq \ln \left( 1 + \frac{\sum_{t=0}^{T-1} g_t^2}{l\beta_0} \right) \tag{31}
\]

**Proof.** Using same methods in proof of Lemma \([14\, 15]\) the conclusion can be deduced from Lemma \([20]\) easily.

**proof of Theorem \([19]\)** Like Lemma \([13]\) in strongly convex case, we have
\[
2R(T) = 2 \sum_{t=0}^{T-1} f_t(x_t) - f_t(x^*) \\
\leq 2 \sum_{t=0}^{T-1} \langle g_t, x_t - x^* \rangle - \sum_{t=0}^{T-1} \|x_t - x^*\|_{\text{diag}(\mu)}^2 \\
= \sum_{t=0}^{T-1} \|g_t\|_{\text{diag}(\alpha_{t+1})}^2 + \sum_{t=0}^{T-1} \left( \|x_t - x^*\|_{\text{diag}(\alpha_{t+1})}^2 - \|x_{t+1} - x^*\|_{\text{diag}(\alpha_{t+1})}^2 \right) + \sum_{t=0}^{T-1} \|x_t - x^*\|_{\text{diag}(\mu)}^2 \\
\leq \sum_{t=0}^{T-1} \|g_t\|_{\text{diag}(\alpha_{t+1})}^2 + \sum_{t=0}^{T-1} \|x_t - x^*\|_{\text{diag}(1/\alpha_{t+1}-1/\alpha_t-\mu)}^2 + \|x_0 - x^*\|_{\text{diag}(\alpha_{0})}^2.
\]
Note that in Algorithm\(^4\) \(\lambda (\alpha_{t,j}/\alpha_{t+1,j}^2) \varphi'(\alpha_{t,j}/\alpha_{t+1,j}) = g_{t,j}^2\), so
\[
\frac{1}{\alpha_{t+1,j}} - \frac{1}{\alpha_{t,j}} = \frac{1}{\alpha_{t,j}} \left( \frac{\alpha_{t,j}}{\alpha_{t+1,j}} - 1 \right)
\leq \frac{1}{\gamma \alpha_{t,j}} \varphi' \left( \frac{\alpha_{t,j}}{\alpha_{t+1,j}} \right) = \frac{\alpha_{t+1,j}^2 g_{t,j}^2}{\alpha_{t,j}^2} \frac{\gamma^2}{\lambda^2} \leq \frac{G^2}{\lambda^2}.
\]
And in Algorithm\(^5\) \(\alpha_{t+1,j} = \alpha_{t,j}^2 / (\varphi')^{-1}(\alpha_{t,j} g_{t,j}^2 / \lambda)\), thus same conclusion holds:
\[
\frac{1}{\alpha_{t+1,j}} - \frac{1}{\alpha_{t,j}} = \frac{1}{\alpha_{t,j}} \left( \frac{\alpha_{t,j}}{\alpha_{t+1,j}} - 1 \right)
\leq \frac{1}{\alpha_{t,j}^2 \gamma} \varphi' \left( \frac{\alpha_{t,j}}{\alpha_{t+1,j}} \right) = \frac{g_{t,j}^2}{\lambda^2} \leq \frac{G^2}{\lambda^2}.
\]
Hence, if \(\lambda \geq \max_{j=1..d} \frac{g_{t,j}^2}{\gamma \mu_{j}}\), then \(1/\alpha_{t+1} - 1/\alpha_t \leq \mu\), and
\[
\sum_{t=0}^{T-1} \|x_t - x^*\|^2_{\text{diag}(1/\alpha_{t+1} - 1/\alpha_t - \mu)} \leq 0.
\]
On the other hand, let \(\beta_t = 1/\alpha_t\),
\[
\sum_{t=0}^{T-1} \|g_t\|^2_{\text{diag}(\alpha_{t+1})} = \sum_{j=1}^d \sum_{t=0}^{T-1} \frac{g_{t,j}^2}{\beta_{t+1,j}},
\]
following from Lemma\(^2\)\(^1\) we have
\[
\sum_{t=0}^{T-1} \|g_t\|^2_{\text{diag}(\alpha_{t+1})} \leq l \left(1 + \frac{G^2}{\lambda \beta_0} \right) \sum_{j=1}^d \ln \left(1 + \frac{\|g_{0:T-1,j}\|^2_{\text{diag}}}{\lambda \beta_0} \right) \text{ in Algorithm}\(^4\)
\]
\[
\sum_{t=0}^{T-1} \|g_t\|^2_{\text{diag}(\alpha_{t+1})} \leq l \sum_{j=1}^d \ln \left(1 + \frac{\|g_{0:T-1,j}\|^2_{\text{diag}}}{\lambda \beta_0} \right) \text{ in Algorithm}\(^5\)
\]

13 Run-time in Full batch Setting

In this section, we will discuss the convergence of our methods in full batch setting.

We first review a classical result on the convergence rate for gradient descent with fixed learning rate.

**Theorem 22.** Suppose that \(F \in C_{L,1}^1(\mathbb{R}^d)\) and \(F^* = \inf_x F(x) > -\infty\). Consider gradient descent with constant step size, \(x_{t+1} = x_t - \nabla F(x_t) / b\). If \(b > \frac{L}{2}\), then
\[
\min_{0 \leq t \leq T-1} \|\nabla F(x_t)\|^2_{\text{2}} \leq \varepsilon
\]
after at most a number of steps
\[
T = \frac{2b^2 (F(x_0) - F^*)}{\varepsilon (2b - L)} = \mathcal{O} \left( \frac{1}{\varepsilon} \right)
\]

**Proof.** Following from the fact that \(F\) is \(L\)-smooth, we have
\[
F(x_{t+1}) \leq F(x_t) - \frac{1}{b} \|\nabla F(x_t)\|^2_{\text{2}} + \frac{L}{2} \|x_{t+1} - x_t\|^2_{\text{2}}
\]
\[
= F(x_t) - \frac{1}{b} \left(1 - \frac{L}{2b} \|\nabla F(x_t)\|^2_{\text{2}} \right)
\]
\[
= F(x_t) - \frac{1}{b} \left(1 - \frac{L}{2b} \|\nabla F(x_t)\|^2_{\text{2}} \right).
\]

(32)
We denote the reciprocal of learning rate $\alpha$.
while in update rule (12),
(Run-time of update rule (12))

Theorem 24

Following Theorem 23 and 24 are detailed version of Theorem 4.

Hence,

$$\frac{\beta_t}{\gamma} \text{ for } t \in \mathbb{R}^+$$

Our proof is followed from the proof of Theorem 2.3 in WNGrad [28].

Next we will show that convergence of both update rules (11) and (12) are robust to the choice of initial learning rate.

Remark 9.
If we choose $b \leq \frac{L}{2}$, then convergence of gradient descent with constant learning rate is not guaranteed at all.

We denote the reciprocal of learning rate $\alpha_t$, i.e., $\beta_t = 1/\alpha_t$. Note that in update rule (11), $\beta_{t+1}$ satisfies

$$\beta_{t+1} \phi'(\beta_{t+1}/\beta_t) = \|g_t\|^2,$$

while in update rule (12), $\beta_{t+1}$ satisfies

$$\beta_{t+1} \phi'(\beta_{t+1}/\beta_t) = \|g_t\|^2.$$ 

Following Theorem 23 and 24 are detailed version of Theorem 4.

Theorem 23 (Run-time of update rule (11)). Suppose that $\phi \in C^1_1([1, +\infty))$, $\phi$ is $\gamma$-strongly convex, and $F \in C^{1,1}_L(\mathbb{R}^d)$, $F^* = \inf_x F(x) > -\infty$. For any $\varepsilon \in (0, 1)$, the sequence $\{x_t\}$ obtained from update rule (11) satisfies

$$\min_{j=0:T}\|\nabla F(x_j)\|_2^2 \leq \varepsilon,$$

after $T$ steps, where

$$T = \left\{ \begin{array}{ll}
1 + \left[ \frac{2(\beta_0 + 2(\gamma \beta_0) + 2(\gamma \beta_0) + \gamma F(x_0) - F^*)}{\gamma} \right] & \text{if } \beta_0 \geq L \text{ or } \beta_1 \geq L, \\
1 + \left[ \frac{\log(\frac{L}{\gamma})}{\log(1 + \frac{1}{\gamma})} \right] + \left[ \frac{L(1 + \frac{1}{\gamma}) (F(x_0) - F^* + \frac{L}{2(\gamma \beta_0)})^2}{\varepsilon} \right] & \text{otherwise.}
\end{array} \right.$$ 

Theorem 24 (Run-time of update rule (12)). Suppose that $\phi \in C^1_1([1, +\infty))$, $\phi$ is $\gamma$-strongly convex, and $F \in C^{1,1}_L(\mathbb{R}^d)$, $F^* = \inf_x F(x) > -\infty$. For any $\varepsilon \in (0, 1)$, the sequence $\{x_t\}$ obtained from update rule (12) satisfies

$$\min_{j=0:T}\|\nabla F(x_j)\|_2^2 \leq \varepsilon,$$

after $T$ steps, where

$$T = \left\{ \begin{array}{ll}
1 + \left[ \frac{2(\beta_0 + \|g_0\|^2_2/(\gamma \beta_0) + 2(\gamma \beta_0) + \gamma F(x_0) - F^*)}{\gamma} \right] & \text{if } \beta_0 \geq L \text{ or } \beta_1 \geq L, \\
1 + \left[ \frac{\log(\frac{L}{\gamma})}{\log(1 + \frac{1}{\gamma})} \right] + \left[ \frac{L + \frac{2L}{\gamma \beta_0} + \frac{2L}{\gamma} (F(x_0) - F^* + \frac{L}{2(\gamma \beta_0)})^2}{\varepsilon} \right] & \text{otherwise.}
\end{array} \right.$$ 

We begin our proof by following lemma.

Lemma 25. Suppose $\phi \in C^1_1(\mathbb{R}^+)$. Fix $\varepsilon \in (0, 1)$. In both update rules (11) and (12), after $T = \left[ \frac{\log(\frac{L}{\gamma})}{\log(1 + \frac{1}{\gamma})} \right] + 1$ steps, either $\min_{t=0:T-1}\|g_t\|^2_2 \leq \varepsilon$, or $\beta_T \geq L$ holds.

Proof. Assume that $\beta_T < L$ and $\min_{t=0:T-1}\|g_t\|^2_2 > \varepsilon$. Recall that the sequence $\{\beta_t\}$ is an increasing sequence. Hence, $\beta_t < L$ for $0 \leq t \leq T$. So, for all $0 \leq t \leq T - 1$,

$$\phi' \left( \frac{\beta_{t+1}}{\beta_t} \right) = \frac{\|g_t\|^2_2}{\beta_t} > \frac{\varepsilon}{L^2} \text{ (for update rule (11)),}$$

$$\phi' \left( \frac{\beta_{t+1}}{\beta_t} \right) = \frac{\|g_t\|^2_2}{\beta_t} > \frac{\varepsilon}{L^2} \text{ (for update rule (12)).}$$
Note that \( \varphi \) is a \( l \)-smooth convex function, and \( \beta_{t+1}/\beta_t \geq 1 \). So

\[
\varphi'(\frac{\beta_{t+1}}{\beta_t}) \leq l \left( \frac{\beta_{t+1}}{\beta_t} - 1 \right),
\]

then

\[
\frac{\beta_{t+1}}{\beta_t} > \frac{\varepsilon}{L^2} + 1.
\]

In this case,

\[
L > \beta_T = \beta_0 \left( \frac{\varepsilon}{L^2} + 1 \right)^T,
\]

however, it is impossible according to the setting of \( T \) in the lemma.

We first prove Theorem 23 using following lemma.

**Lemma 26.** In update rule (11), suppose \( F \in C^{1,1}_L(\mathbb{R}^d), \varphi \in C^{1,1}_1(\mathbb{R}^{++}), \) and \( \varphi \) is \( \gamma \)-strongly convex function. Denote \( F^* = \inf_{x} F(x) > -\infty \). Let \( t_0 \geq 1 \) be the first index such that \( \beta_{t_0} \geq L \).

\[
\beta_t \leq \beta_{t_0} - 1 + \frac{2}{\gamma} (F(x_{t_0-1}) - F^*). \tag{34}
\]

and moreover,

\[
F(x_{t_0-1}) - F^* \leq F(x_0) - F^* + \frac{L}{2\beta_0} (\beta_{t_0-1} - \beta_0) \tag{35}
\]

**Proof.** Same as equation (32).

\[
F(x_{t+1}) \leq F(x_t) - \frac{1}{\beta_{t+1}} \left( 1 - \frac{L}{2\beta_{t+1}} \right) \| g_t \|^2_2.
\]

For \( t \geq t_0 - 1, \beta_{t+1} \geq L \), so

\[
F(x_{t+1}) \leq F(x_t) - \frac{1}{2\beta_{t+1}} \| g_t \|^2_2.
\]

Hence, for all \( k \geq 0, \)

\[
F(x_{t_0+k}) \leq F(x_{t_0-1}) - \frac{1}{2} \sum_{i=0}^{k} \| g_{t_0+i-1} \|^2_2 \beta_{t_0+i}^{-1}, \tag{36}
\]

i.e.,

\[
\sum_{i=0}^{k} \| g_{t_0+i-1} \|^2_2 \beta_{t_0+i}^{-1} \leq 2(F(x_{t_0-1}) - F^*). \tag{37}
\]

Note that \( \varphi \) is \( \gamma \)-strongly convex and

\[
\frac{\| g_t \|^2_2}{\beta_{t+1}} = \beta_{t+1} \varphi'(\frac{\beta_{t+1}}{\beta_t}) \geq \gamma \beta_t \left( \frac{\beta_{t+1}}{\beta_t} - 1 \right),
\]

and

\[
\beta_{t+1} - \beta_t \leq \frac{1}{\gamma} \frac{\| g_t \|^2_2}{\beta_{t+1}}. \tag{38}
\]

Combining equation (37) and equation (38), we have

\[
\beta_{t_0+k} \leq \beta_{t_0-1} + \frac{1}{\gamma} \sum_{i=0}^{k} \frac{\| g_{t_0+i-1} \|^2_2}{\beta_{t_0+i}} \leq \beta_{t_0-1} + \frac{2}{\gamma} (F(x_{t_0-1}) - F^*).
\]
We remain to give an upper bound for \( F(x_{t_0-1}) \) in the case \( t_0 > 1 \). Using equation (32) again, we get
\[
F(x_{t_0-1}) - F(x_0) \leq \sum_{i=0}^{t_0-2} - \frac{1}{\beta_{i+1}} \left( 1 - \frac{L}{2\beta_{i+1}} \right) \|g_i\|_2^2 \leq \frac{L}{2} \sum_{i=0}^{t_0-2} \frac{\|g_i\|_2^2}{\beta_{i+1}^2}
= \frac{L}{2} \sum_{i=0}^{t_0-2} \varphi \left( \frac{\beta_{i+1}}{\beta_i} \right) \leq \frac{L}{2} t_0 \sum_{i=0}^{t_0-2} \left( \frac{\beta_{i+1}}{\beta_i} - 1 \right)
\leq L \frac{t_0}{2} \sum_{i=0}^{t_0-2} \left( \frac{\beta_{i+1} - \beta_i}{\beta_0} \right) = L \frac{t_0}{2\beta_0} (\beta_{t_0-1} - \beta_0).
\]

In the above, the second inequality follows from the assumed \( \ell \)-smoothness of \( \varphi \), and the last inequality follows from \( \beta_t \geq \beta_0 \) for all \( t \geq 0 \).

**proof of Theorem 23.** If \( t_0 = 1 \), by equation (36), for all \( t \geq 1 \), we have
\[
F(x_t) \leq F(x_0) - \frac{1}{2} \sum_{i=0}^{t-1} \frac{\|g_i\|_2^2}{\beta_{i+1}} \leq F(x_0) - \frac{1}{2} \sum_{i=0}^{t-1} \frac{\|g_i\|_2^2}{\beta_0 + \frac{2}{\gamma} (F(x_0) - F^*)}.
\]
Then after \( T = 1 + \left\lceil \frac{2(\beta_0 + 2F(x_0) - F^*)/\gamma)(F(x_0) - F^*)}{\varepsilon} \right\rceil \) steps,
\[
\min_{t=0:T-1} \|g_t\|_2^2 \leq \frac{1}{T} \sum_{t=0}^{T-1} \|g_t\|_2^2 \leq \frac{2}{T} (F(x_0) - F^*) (\beta_0 + \frac{2}{\gamma} (F(x_0) - F^*)) \leq \varepsilon.
\]

Otherwise, if \( t_0 > 1 \), we have \( \beta_{t_0-1} < L \). Then for all \( t \geq t_0 \),
\[
\beta_t \leq L + \frac{2}{\gamma} \left( F(x_0) - F^* + \frac{L(L - \beta_0)}{2\beta_0} \right) (39)
\]
Denote the right hand of equation (39) as \( \beta_{\max} \). Using equation (36) again, for we have
\[
F(x_{t_0+M}) \leq F(x_{t_0-1}) - \frac{1}{2} \sum_{i=0}^{M} \frac{\|g_{t_0+i-1}\|_2^2}{\beta_{t_0+i}} \leq F(x_{t_0-1}) - \frac{1}{2\beta_{\max}} \sum_{i=0}^{M} \|g_{t_0+i-1}\|_2^2.
\]

Hence,
\[
\min_{t=t_0:M-1} \|g_t\|_2^2 \leq \min_{t=t_0-1:t_0+M-1} \|g_t\|_2^2 \leq \frac{M}{M+1} \sum_{i=0}^{M} \|g_{t_0+i-1}\|_2^2 \leq \frac{1}{M+1} 2\beta_{\max} (F(x_{t_0-1}) - F^*) \leq \frac{2\beta_{\max}}{M+1} \left( F(x_0) - F^* + \frac{L(L - \beta_0)}{2\beta_0} \right).
\]

At last, with recalling the conclusion of Lemma 23, after
\[
T = \left\lceil \frac{\log \left( \frac{\beta_0}{\varepsilon} \right)}{\log \left( L/\left( L + 1 \right) \right)} \right\rceil + \frac{2\beta_{\max}}{\varepsilon} \left( F(x_0) - F^* + \frac{L(L - \beta_0)}{2\beta_0} \right) + 1\]
steps, we have \( \min_{t=0:T-1} \|g_t\|_2^2 \leq \varepsilon. \)
Next we prove Theorem 24.

**Lemma 27.** In update rule (12), suppose $F \in C^{1,1}_L(\mathbb{R}^d)$, $\varphi \in C^{1,1}_L(\mathbb{R}^+)$, and $\varphi$ is $\gamma$-strongly convex function. Denote $F^* = \inf_x F(x)$. Let $t_0 \geq 1$ be the first index such that $\beta_{t_0} \geq \bar{L}$. Then for all $t \geq t_0$,

$$\beta_t \leq \beta_{t_0} + \frac{8}{\gamma} (F(x_{t_0-1}) - F^*),$$

and moreover,

$$F(x_{t_0-1}) - F^* \leq F(x_0) - F^* + \frac{Ll}{2\beta_0} (\beta_{t_0-1} - \beta_0),$$

$$\beta_{t_0} \leq \begin{cases} 
\beta_0 + \frac{\|g_0\|^2_2}{\gamma \beta_0} & \text{if } t_0 = 1, \\
L + \frac{2}{\gamma \beta_0} L^2 + \frac{2L}{\gamma} & \text{if } t_0 \geq 2,
\end{cases}$$

**Proof.** Same as the proof of Lemma 26, we first get for all $k \geq 0$,

$$\sum_{i=0}^{k} \frac{\|g_{t_0+i-1}\|^2_2}{\beta_{t_0+i}} \leq 2(F(x_{t_0-1}) - F^*).$$

Note that in update rule (12), $\beta^2 \varphi'(\beta_{t+1}/\beta_t) = \|g_t\|^2_2$. So

$$\beta_{t_0+k+1} = \beta_{t_0+k} + \beta_{t_0+k} \left( \frac{\beta_{t_0+k+1}}{\beta_{t_0+k}} - 1 \right)$$

$$\leq \beta_{t_0+k} + \frac{\beta_{t_0+k}}{\gamma} \varphi' \left( \frac{\beta_{t_0+k+1}}{\beta_{t_0+k}} \right) = \beta_{t_0+k} + \frac{1}{\gamma} \frac{\|g_{t_0+k}\|^2_2}{\beta_{t_0+k}}$$

$$\leq \beta_{t_0+k} + \frac{2L^2}{\gamma} \frac{\|x_{t_0+k} - x_{t_0+k-1}\|^2_2 + \|g_{t_0+k-1}\|^2_2}{\beta_{t_0+k}}$$

$$\leq \beta_{t_0+k} + \frac{4}{\gamma} \frac{\|g_{t_0+k-1}\|^2_2}{\beta_{t_0+k}} \leq \beta_{t_0} + \frac{4}{\gamma} \sum_{i=0}^{k} \frac{\|g_{t_0+i-1}\|^2_2}{\beta_{t_0+i}}$$

$$\leq \beta_{t_0} + \frac{8}{\gamma} (F(x_{t_0-1}) - F^*).$$

If $t_0 = 1$, then

$$\beta_{t_0} \leq \beta_0 + \frac{\|g_0\|^2_2}{\gamma \beta_0},$$

and if $t_0 \geq 2$, then

$$\beta_{t_0} \leq \beta_{t_0-1} + \frac{\|g_{t_0-1}\|^2_2}{\gamma \beta_{t_0-1}} = \beta_{t_0-1} + \frac{2L^2}{\gamma} \frac{\|g_{t_0-2}\|^2_2}{\beta_{t_0-1}} + \frac{2}{\gamma} \frac{\|g_{t_0-2}\|^2_2}{\beta_{t_0-2}}$$

$$\leq \beta_{t_0-1} + \frac{2L^2}{\gamma} \beta_{t_0-2} + \frac{2L}{\gamma} \beta_{t_0-2}$$

$$\leq L + \frac{2L}{\gamma} \beta_{t_0-2} + \frac{2L}{\gamma} L.$$
At last, for $t_0 > 0$, we have

\[
F(x_{t_0-1}) - F(x_0) \leq \sum_{i=0}^{t_0-2} -\frac{1}{\beta_i+1} \left( 1 - \frac{L}{2\beta_i+1} \right) \|g_i\|_2^2 \\
\leq \frac{L}{2} \sum_{i=0}^{t_0-2} \frac{\|g_i\|_2^2}{\beta_i^2} \leq \frac{L}{2} \sum_{i=0}^{t_0-2} \frac{\|g_i\|_2^2}{\beta_i^2} \\
= \frac{L}{2} \sum_{i=0}^{t_0-2} \phi' \left( \frac{\beta_i+1}{\beta_i} \right) \leq \frac{L}{2} \sum_{i=0}^{t_0-2} \left( \frac{\beta_i+1}{\beta_i} - 1 \right) \\
\leq \frac{Lt}{2} \sum_{i=0}^{t_0-2} \left( \frac{\beta_i+1}{\beta_0} - \frac{\beta_i}{\beta_0} \right) = \frac{Lt}{2\beta_0} (\beta_{t_0-1} - \beta_0).
\]

\[\square\]

**proof of Theorem 24** The proof is completely similar to the proof of Theorem 23.

\[\square\]