Saint-Petersburg State University

Artem Beliakov

Graduation work

Polynomial lookahead. Reformulation of convex optimization problems in terms of hessian

Educational level: bachelor
Field of study 01.03.01 «Mathematics»
Main education program CB.5000.2018 «Mathematics»

Research supervisor:
assosiate professor, Candidate of Physical and Mathematical Sciences
Alexander Avdyushenko

Reviewer
professor, D.Sc. (Mathematics)
Alexander Gasnikov

Saint-Petersburg
2022

Санкт-Петербургский государственный университет

Беляков Артем Алексеевич

Выпускная квалификационная работа

Полиномиальный лукэхед.
Переформулировка задач выпуклой оптимизации в терминах гессиана

Уровень образования: бакалавриат
Направление 01.03.01 «Математика»
Основная образовательная программа СВ.5000.2018 «Математика»

Научный руководитель:
доцент, к.ф-м.н.
А.Ю. Авдюшенко

Рецензент
профессор, доктор физико-математических наук
А.В. Гасников

Санкт-Петербург
2022

Contents

Abstract

For any stochastic gradient method one have to come to a tradeoff between rate of convergence and limit noise level near the solution. In this paper we present a new stochastic gradient type method, which solves this problem more efficiently. In particular, we show that for quadratic problems with a fixed desired noise level, this algorithm converges faster than known stochastic methods. Moreover, we give the convergence of the new technique for a general class of convex problems. The practical results confirm the theoretical analysis.

Аннотация

Для произвольного градиентного метода мы сталкиваемся с компромиссом между скоростью сходимости и предельным уровнем шума вокруг решения. В данной статье мы вводим новый стохастический градиентный метод, который справляется с данной задачей лучше. В частности, мы покажем, что для квадратичной задачи с заданным фиксированным желаемым шумом наш алгоритм сходится быстрее, чем известные стохастические методы. Более того, мы выводим сходимость нового метода для общего класса задач выпуклой оптимизации. Теоретические результаты подкрепляются экспериментами.

1. Introduction

We focus on the optimization problem

$$\min_{x \in \mathbb{R}^n} f(x) \tag{1}$$

These kinds of problems arise in various fields of applied science. By now deterministic algorithms for solving problem (1) have been extensively investigated (Nesterov ((2018))). Meanwhile, with the emergence of new fields of interest (Vapnik ((1999))), one often has to deal with the stochastic formulation of the problem (1):

$$f(x) = \mathbb{E}_{\xi \sim \mathcal{D}} \left[f_\xi(x) \right], \tag{2}$$

where ξ is a random variable. Stochastic formulations come up when we do not have access to (or do not want to use due to the high computational cost) the deterministic value of $f(x)$. For example, in the case of statistical supervised learning theory (Shalev-Shwartz and Ben-David ((2014))). In this case, x is a machine learning

model with n weights/parameters (it can be a regression problem or a deep neural network), $\mathcal{D}$ represents an unknown data distribution, f_ξ is the loss of model x on datapoint ξ, and f is the generalization error. We assume that we do not know the distribution $\mathcal{D}$ from which the data come, but have some samples ξ from $\mathcal{D}$. But we want to have a suitable model fit to whole $\mathcal{D}$. Deterministic methods are no longer suitable, we need stochastic modifications of classical methods. One of the most popular such methods is the stochastic gradient descent (SGD) (Robbins and Monro ((1951)); Nemirovski et al. ((2009)); Vaswani et al. ((2019)); Gorbunov et al. ((2020))).

In the paper we compare different stochastic gradient methods in the regime of fixed step size. In contrast to deterministic methods, stochastic methods in such a situation do not lead to an exact solution of the problem, but begin to oscillate around the solution, having achieved some accuracy. Then, one comes to tradeoff between rate of convergence and limit noise level near the solution. It is proposed to compare stochastic gradient methods by their rate of convergence with desired limit noise being fixed. We use SGD, an efficient method of convex optimization, as a reasonable baseline to compare with. Also we introduce new algorithm called Polynomial Lookahead or alternatively Polynomial Dynamic Modification(PDM) which takes its inspiration in already studied in the literature Lookahead algorithm firstly introduced in Michael R. Zhang and Ba ((2019)) and analysed in quadratic stochastic case. Also Lookahead was studied in non-convex regime in Jianyu Wang and Rabbat ((2020)). These techniques allow to come closer to the solution more efficiently and combat the limit radius of oscillation around the solution.

2. Problem setup and assumptions

We study three methods: SGD, PDM and Lookahead in two type of problems: unidimensional quadratic stochastic problem and strongly convex stochastic problem. In the first part of the paper, we provide an analysis of quadratic case:

$$\min_{x\in\mathbb{R}} L(x) \quad \text{with,} \quad L(x) = \mathbb{E}\widehat{L}(x) \tag{3}$$

$$\widehat{L}(x) = \frac{1}{2}(x-c)^2$$

where $c \sim N(x^*, \sigma^2)$. Without loss of the generality we suppose $x^* = 0$.

We show that with desired limit noise level being fixed PDM experiences the best rate of convergence, while Lookahead the worst, rate of convergence of SGD being just in the middle.

In the second part we analyze the strongly-convex stochastic problem (1) + (2) under assumptions:

Assumption 1. f lies in $F^2_{\mu,L}$, which is the family of μ-strongly convex twice continuously differentiable functions with L-Lipschitz gradients. Which means:

- $f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \frac{\mu}{2}\|y - x\|^2 \, \forall x, y \in \mathbb{R}^n$ - strong convexity
- $\|\nabla f(x) - \nabla f(y)\| \leq L\|x - y\| \, \forall x, y \in \mathbb{R}^n$ - L-Lipschitzness of the gradients.

This two conditions under twice continuous differentiability can be equivalently formulated as:

$$\mu I \preceq H(x) \preceq LI \; \forall x \in \mathbb{R}^n$$

where $H(x)$ - Hessian of f at point x.

Assumption 2. We have unbiased estimate of gradient $\nabla f(x)$ with uniformly bounded noise:

$$\nabla F(x) = \nabla f(x) + \xi$$

$$\mathbb{E}\xi = 0$$

$$\mathbb{E}\xi^2 \leq \sigma^2$$

In this setting we deduce rate of convergence of PDM and Lookahead via new technique that we call reformulation in terms of hessian. For Lookahead there were no estimates in strongly convex case before. We show that Lookahead still experiences worse rate of convergence than SGD in the worst case sense with limit noise being fixed.

3. Lookahead and Polynomial Dynamic Modification

In this section we are going to introduce Lookahead and PDM algorithms. Lookahead algorithm have slow and fast iterations, and slow and fast weights. At

Algorithm 1 Lookahead($\gamma, x_0, T, k, \alpha$)

input: γ - step size, x_0 - start point, T - number of slow iterations
$t \leftarrow 0$
$\phi_0 \leftarrow x_0$
while $t \leq T$ **do**
  $\theta_{t+1,0} = \phi_t$
  for $i = 1, i \leq k$ **do**
    $\theta_{t+1,i} = \theta_{t+1,i-1} - \gamma \nabla F(\theta_{t+1,i-1})$
  end for
  $\phi_{t+1} = (1 - \alpha)\phi_t + \alpha\theta_{t+1,k}$
end while
return ϕ_T

Algorithm 2 PDM($\gamma, x_0, T, k, a_0, \ldots, a_k$)

input: γ - step size, x_0 - start point, T - number of slow iterations
$t \leftarrow 0$
$\phi_0 \leftarrow x_0$
while $t \leq T$ **do**
  $\theta_{t+1,0} = \phi_t$
  for $i = 1, i \leq k$ **do**
    $\theta_{t+1,i} = \theta_{t+1,i-1} - \gamma \nabla F(\theta_{t+1,i-1})$
  end for
  $\phi_{t+1} = \sum_{i=0}^{k} a_i \theta_{t+1,i}$
end while
return ϕ_T

each slow iteration it k times performs fast iterations which is just a gradient descent. At the end of the slow iteration it updates slow weight as average of fast weight at k-th fast iteration and slow weight at the previous slow iteration. See algorithm 1.

PDM has the same principle, but applies different averaging scheme at the end of slow iteration. It has positive coefficients $a_0, \dots, a_k$ such that $\sum_{i=0}^{k} a_i = 1$ and to gain slow weight it averagies all the fast weights at the slow iteration with coefficients $a_0, \dots, a_k$.

On the matter of fact, Lookahead and PDM algorithms can apply any inner solver instead of SGD. We can formulate these algorithms alternatively in terms of dynamical systems. Let $F_\sharp \colon \mathbb{R}^n \to \big((\Omega, \mathcal{A}, \mathcal{P}) \to \mathbb{R}^n\big)$ be arbitrary discrete stochastic dynamical system where $(\Omega, \mathcal{A}, \mathcal{P})$ - probability space. Then, dynamical system corresponding to one slow iteration of Lookahead can be gained as:

$$F_{LA}(x) = p(F_\sharp), \qquad \text{where } p(x) = (1-\alpha) + \alpha x^k$$

And for PDM:

$$F_{PDM}(x) = p(F_\sharp), \qquad \text{where } p(x) = \sum_{i=0}^{k} a_i x^i$$

Thus, Lookahead and Polynomial Dynamic Modification can be considered to be modification of arbitrary dynamical system. From here the name of the second method takes roots. Here, in case of SGD:

$$F_\sharp(x) = F_{SGD}(x) = x - \gamma \nabla F(x)$$

Also note that condition $\sum_{i=0}^{k} a_i = 1$ is necessary for $\overline{F}_{PDM}(x) = EF_{PDM}(x)$ to save stationary points of $\overline{F_\sharp}(x) = EF_\sharp(x)$. Thus, if x^* stationary point of $\overline{F_\sharp}$, i.e. $\overline{F_\sharp}(x^*) = x^*$, then $\overline{F}_{PDM}(x^*) = x^*$.

4. Unidimensional quadratic problem

In the section we analyse behaviour of all three methods on problem 3. Through out this section we suppose $\gamma \leq 1$ where γ is step size. Firstly, con-

sider SGD algorithm:

$$\phi_{t+1} = \phi_t - \gamma \nabla \widehat{L}(\phi_t)$$

Denote by $V^*_{SGD} = \lim_{t\to\infty} E\|\phi_t\|^2$ - limit noise. Then following lemma shows linear convergence of SGD to its limit noise:

Lemma 4.1.

$$E\|\phi_{t+1}\|^2 = (1-\gamma)^2 E\|\phi_t\|^2 + \gamma^2\sigma^2$$

$$V^*_{SGD} = \frac{\gamma^2\sigma^2}{1-(1-\gamma)^2}$$

And from the first two equalities linear convergence to V^*_{SGD} follows:

$$E\|\phi_{t+1}\|^2 - V^*_{SGD} = (1-\gamma)^2(E\|\phi_t\|^2 - V^*_{SGD})$$

Proof.

$$\phi_{t+1} = \phi_t - \gamma \nabla \widehat{L}(\phi_t) = (1-\gamma)\phi_t + \gamma c$$

and we immediately get:

$$E\|\phi_{t+1}\|^2 = (1-\gamma)^2 E\|\phi_t\|^2 + \gamma^2\sigma^2$$

By taking a limit in this equality when t goes to ∞ we get:

$$V^*_{SGD} = (1-\gamma)^2 V^*_{SGD} + \gamma^2\sigma^2$$

$$V^*_{SGD} = \frac{\gamma^2\sigma^2}{1-(1-\gamma)^2}$$

□

Denote by $s = 1-\gamma$ - rate of convergence of SGD [1] which we can see from lemma 4.1. Finally, rewrite V^*_{SGD} in terms of s:

Lemma 4.2.

$$V^*_{SGD} = \sigma^2 \frac{1-s}{1+s}$$

[1] We call s - rate of convergence, though it is a square root of actual rate of convergence. Further, we will do the same for PDM and Lookahead.

We want to find algorithm with fastest convergence with limit noise being fixed. But instead we can actually find algorithm with the best limit noise with rate of convergence being fixed. The equivalence of these two tasks comes from the fact that function $v^*(t)$ where v^* denotes limit noise and t denotes rate of convergence is decreasing for all three algorithms which we will see later on.

4.1. PDM analysis

Consider PDM algorithm:

$$\phi_{t+1} = \sum_{i=0}^{k} a_i \theta_{t+1,i}$$

$$\theta_{t+1,i} = \theta_{t+1,i-1} - \gamma \nabla \widehat{L}(\theta_{t+1,i-1})$$

$$\theta_{t+1,0} = \phi_t$$

Lemma 4.3. *For PDM*

$$E\|\phi_{t+1}\|^2 = (\sum_{i=0}^{k} a_i (1-\gamma)^i)^2 E\|\phi_t\|^2 + \gamma^2\sigma^2 \sum_{i=0}^{k-1} (\sum_{j=0}^{k-1-i} a_{i+j+1}(1-\gamma)^j)^2$$

$$V^*_{PDM} = \gamma^2\sigma^2 \frac{\sum_{i=0}^{k-1} (\sum_{j=0}^{k-1-i} a_{i+j+1}(1-\gamma)^j)^2}{1 - (\sum_{i=0}^{k} a_i (1-\gamma)^i)^2}$$

Proof.

$$\theta_{t+1,i} = \theta_{t+1,i-1} - \gamma \nabla \widehat{L}(\theta_{t+1,i-1}) = (1-\gamma)\theta_{t+1,i-1} + \gamma c_{t+1,i-1} = (1-\gamma)^i \phi_t + \gamma (\sum_{j=0}^{i-1} c_{t+1,j} (1-\gamma)^{i-1-j})$$

$$\phi_{t+1} = \sum_{i=0}^{k} a_i \theta_{t+1,i} = (\sum_{i=0}^{k} a_i (1-\gamma)^i)\phi_t + \gamma \sum_{i=0}^{k-1} (\sum_{j=0}^{k-1-i} a_{i+j+1}(1-\gamma)^j) c_{t+1,i}$$

Then

$$E\|\phi_{t+1}\|^2 = (\sum_{i=0}^{k} a_i(1-\gamma)^i)^2 E\|\phi_t\|^2 + \gamma^2\sigma^2 \sum_{i=0}^{k-1}(\sum_{j=0}^{k-1-i} a_{i+j+1}(1-\gamma)^j)^2$$

By taking limit we get:

$$V^*_{PDM} = \gamma^2\sigma^2 \frac{\sum_{i=0}^{k-1}(\sum_{j=0}^{k-1-i} a_{i+j+1}(1-\gamma)^j)^2}{1-(\sum_{i=0}^{k} a_i(1-\gamma)^i)^2}$$

and lemma follows. □

For convenience denote $c = 1-\gamma$. Also denote $t = \sum_{i=0}^{k} a_i(1-\gamma)^i$ - rate of convergence of PDM as we can see from lemma 4.3. What is more, we suppose that $t = s^k$ where s - rate of convergence of SGD as by the time PDM takes one step SGD can make k steps.

In terms of c and t we have:

Lemma 4.4.

$$V^*_{PDM} = (1-c)^2\sigma^2 \frac{\sum_{i=0}^{k-1}(\sum_{j=0}^{k-1-i} a_{i+j+1}c^j)^2}{1-t^2}$$

Now, we want to find optimal values of coefficients $a_0, \ldots, a_k$. So, we aim to solve the following problem:

$$V^*_{PDM} \longrightarrow \min_{\substack{a_i \geq 0 \, \forall i \\ \sum_{i=0}^{k} a_i = 1 \\ \sum_{i=0}^{k} a_i c^i = t}} \tag{4}$$

Lemma 4.5. *Under condition:*

$$\frac{1-t}{1+(k-1)t} \leq \gamma \leq 1$$

Solution of 4 is attained by:

$$\begin{cases} a_0 = t - \frac{(1-t)c}{(1-c)k} \\ a_i = \frac{1-t}{k}, \quad i = 1, \cdots, k-1 \\ a_k = \frac{1-t}{(1-c)k} \end{cases}$$

And in this case:

$$V^*_{PDM} = \frac{\sigma^2(1-t)}{k(1+t)}$$

doesn't depend on γ.

Proof. Problem 4 is equivalent to:

$$\sum_{i=0}^{k-1} \Big(\sum_{j=0}^{k-1-i} a_{i+j+1} c^j \Big)^2 \longrightarrow \min_{\substack{a_i \geq 0 \, \forall i \\ \sum_{i=0}^k a_i = 1 \\ \sum_{i=0}^k a_i c^i = t}}$$

Consider change of variables $d_i = \sum_{j=i}^k a_j c^{j-i}$. Inverse to this change of variables will be:

$$\begin{cases} a_i = d_i - c d_{i+1} \text{ for } i = 0, \ldots, k-1 \\ a_k = d_k \end{cases}$$

Condition $\sum_{i=0}^k a_i = 1$ is transformed to $\sum_{i=0}^{k-1} d_i - c d_{i+1} + d_k = 1$ or $\sum_{i=1}^k d_i = \frac{1-t}{1-c}$. Then the problem takes following form:

$$\sum_{i=1}^k d_i^2 \longrightarrow \min_{\substack{d_i - c d_{i+1} \geq 0 \\ d_k \geq 0 \\ d_0 = t \\ \sum_{i=1}^k d_i = \frac{1-t}{1-c}}}$$

Note that problem

$$\sum_{i=1}^k d_i^2 \longrightarrow \min_{\sum_{i=1}^k d_i = \frac{1-t}{1-c}}$$

is standard and its minimum is attained when $d_1 = \ldots = d_k$ which gives the following solution:

$$\begin{cases} d_0 = t \\ d_i = \frac{1-t}{(1-c)k}, \quad i = 1, \ldots, k \end{cases} \tag{5}$$

Note, though, that conditions $d_i - c d_{i+1} \geq 0, d_k \geq 0$ must be satisfied. All conditions are satisfied automatically, except $d_0 - c d_1 \geq 0$ which adds following

restriction:

$$t - \frac{c(1-t)}{(1-c)k} \geq 0$$

$$\frac{c}{1-c} \leq \frac{kt}{1-t}$$

$$c \leq \frac{kt}{1+(k-1)t}$$

$$\frac{1-t}{1+(k-1)t} \leq \gamma$$

Note that for $\gamma < \frac{1-t}{1+(k-1)t}$ we get worse V^*_{PDM}. Finally, from system 5 we get optimal $\{a_i\}_{i=0}^k$:

$$\begin{cases} a_0 = t - \frac{(1-t)c}{(1-c)k} \\ a_i = \frac{1-t}{k}, \quad i = 1, \cdots, k-1 \\ a_k = \frac{1-t}{(1-c)k} \end{cases}$$

Now, let's deduce formula for V^*_{PDM}:

$$V^*_{PDM} = (1-c)^2\sigma^2 \frac{\sum_{i=0}^{k-1}(\sum_{j=0}^{k-1-i} a_{i+j+1}c^j)^2}{1-t^2} = (1-c)^2\sigma^2 \frac{\sum_{i=1}^{k} d_i^2}{1-t^2} = \frac{\sigma^2(1-t)}{k(1+t)}$$

□

4.2. Lookahead analysis

For Lookahead from lemma 4.3 by substituting $a_0 = 1-\alpha, a_k = \alpha, a_1 = \ldots = a_{k-1} = 0$ next lemma follows:

Lemma 4.6. *For Lookahead*

$$E\|\phi_{t+1}\|^2 = ((1-\alpha) + \alpha(1-\gamma)^k)^2 E\|\phi_t\|^2 + \sigma^2\alpha^2\gamma^2 \frac{1-(1-\gamma)^{2k}}{1-(1-\gamma)^2}$$

$$V^*_{LA} = \alpha^2\sigma^2\gamma^2 \frac{1-(1-\gamma)^{2k}}{(1-(1-\gamma)^2)(1-((1-\alpha)+\alpha(1-\gamma)^k)^2)}$$

Next we formulate auxiliary lemma necessary for our analysis:

Lemma 4.7.

$$f(x) = \frac{(1-x)(1+x^k)}{(1+x)(1-x^k)}$$

is decreasing for $0 \leq x \leq 1$.

Proof. Consider taking derivative $f'(x)$ and prove that $f'(x) \leq 0$:

$$f'(x)(1+x)^2(1-x^k)^2 = -(1+x^k)(1+x)(1-x^k)+kx^{k-1}(1-x)(1+x)(1-x^k)-$$
$$(1-x)(1+x^k)(1-x^k)+kx^{k-1}(1-x)(1+x^k)(1+x) =$$
$$2(-1+kx^{k-1}-kx^{k+1}+x^{2k})$$

Now, we need to prove that $-1+kx^{k-1}-kx^{k+1}+x^{2k} \leq 0$. For this it is enough to prove that this function is increasing as its value at $x=1$ equals 0. Let's show that its derivative is positive:

$$k(k-1)x^{k-2}-k(k+1)x^k+2kx^{2k-1} \geq 0$$

Or

$$k-1-(k+1)x^2+2x^{k+1} \geq 0$$

For this it is enough to prove that the last function is decreasing as its value at $x=1$ is 0. Its derivative is obviously negative:

$$2(k+1)x(-1+x^{k-1}) \leq 0$$

□

Denote by $c = 1-\gamma$ and $t = (1-\alpha)+\alpha c^k$ - rate of convergence of Lookahead. Then from lemmas 4.6 and 4.7 next lemma follows:

Lemma 4.8.

$$V^*_{LA} = \frac{\sigma^2(1-c)(1+c^k)(1-t)}{(1+c)(1-c^k)(1+t)}, \textit{ with } \quad t \geq c^k$$

Here we can vary c from 0 to $\sqrt[k]{t}$ by varying α.

$$\min_{c^k \leq t} V_{LA}^* = \frac{\sigma^2(1-s)}{1+s}$$

where $s^k = t$ and minimum is attained at $c^k = t$ and $\alpha = 1$. This shows that Lookahead attains worse limit noise than SGD.

Proof. By lemma 4.6 we get:

$$V_{LA}^* = \alpha^2\sigma^2\gamma^2 \frac{1-(1-\gamma)^{2k}}{(1-(1-\gamma)^2)(1-((1-\alpha)+\alpha(1-\gamma)^k)^2)}$$

Then, we use $c = 1-\gamma$ and $t = (1-\alpha)+\alpha c^k$. From second one follows $\alpha = \frac{1-t}{1-c^k}$. Then we get following form of V_{LA}^*:

$$V_{LA}^* = \sigma^2 \frac{(1-t)^2(1-c)^2(1-c^{2k})}{(1-c^k)^2(1-c^2)(1-t^2)} = \frac{\sigma^2(1-c)(1+c^k)(1-t)}{(1+c)(1-c^k)(1+t)}$$

From lemma 4.7

$$\min_{c^k \leq t} V_{LA}^* = \frac{\sigma^2(1-s)}{1+s}$$

follows. □

4.3. Summary

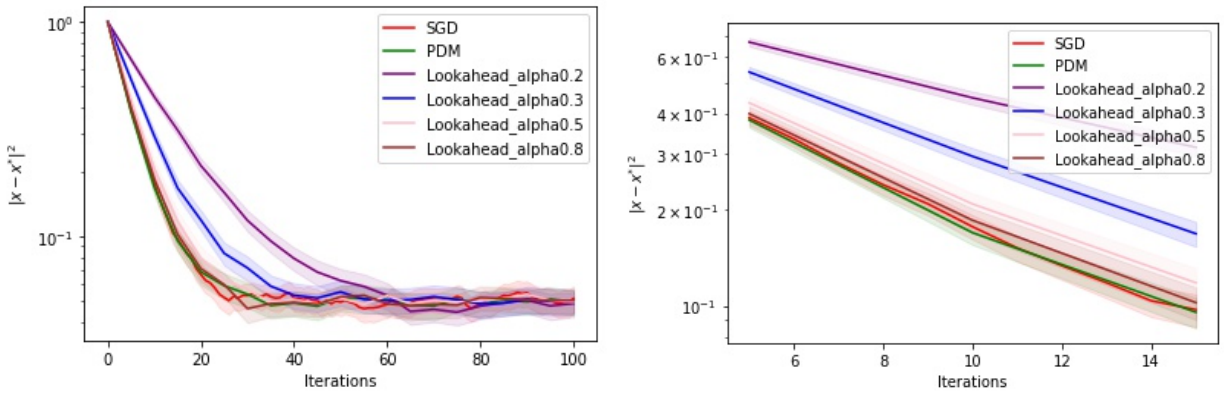

Figure 1: Right figure is just enlarged version of the left one. Here, SGD, PDM and Lookahead were trained on quadratic stochastic function. Parameters of all algorithms are picked up according to the theory so as they all have the same limit noise $v^* = 0.05$. For PDM and Lookahead $k = 5$.

Finally, let's compare SGD and PDM:

Lemma 4.9.

$$\frac{V^*_{PDM}}{V^*_{SGD}} = \frac{(1-t)(1+s)}{k(1+t)(1-s)} \leq 1$$

Proof. $\frac{(1-t)(1+s)}{k(1+t)(1-s)}$ is increasing in s which follows from lemma 4.7 and its value at $s = 1$ equals 1. □

Lemmas 4.8 and 4.9 show exactly that PDM attains the best limit noise and Lookahead the worst. In the Figure 1 empirical results on quadratic function are depicted. It can be seen that rate of convergence of Lookahead is the worst. As for SGD and PDM their convergence rates here are indistinguishable which we prescribe to the high noise level.

5. Strongly convex case

Through out this section we suppose $\gamma \leq \frac{2}{L+\mu}$.

5.1. Hessian reformulation

Now, we consider problem 1 + (2) under assumptions 1 and 2.

On the matter of fact, we can reformulate our problem and algorithms in terms of Hessian and forget that we were given function f. Assumption 1 in terms of hessian can be equivalently reformulated as two assumptions:

Assumption 3.

$$H\colon \mathbb{R}^n \to \mathbb{R}^{n\times n} \text{ is hessian function}$$

Assumption 4.

$$\mu I \preceq H(x) \preceq LI \ \forall x \in \mathbb{R}^n$$

Assumption 3 is indeed a similar assumption to gradient function to be a conservative vector field. In our case also some conservative properties for some integrals over closed paths must be satisfied. Though, we leave analysis of assumption 3 and its corollaries for future work.

Firstly, let's formulate Sylvester's criterion which we will use in this section:

Lemma 5.1. *Sylvester's criterion*

Symmetric matrix A is positive-semidefinite if and only if all its principal minors are non-negative.

Corollary 5.2. *Symmetric matrix A is negative-semidefinite if and only if*

- *all its even principal minors are non-negative*
- *all its odd principal minors are non-positive*

Further we formulate auxiliary lemmas relevant for our analysis:

Lemma 5.3. *For symmetric matrix S such that*

$$-cI \preceq S \preceq cI$$

where $c > 0$.

1. $v = Se \in B(0, c)$, *where* $e = (1, 0, \ldots, 0)$.
2. *For any $v \in B(0, c)$ there exists symmetric S, $-cI \preceq S \preceq cI : v = Se$.*

Proof. 1. Let $v = Se$. Then $\|v\|^2 = v^T v = e^T S^2 e \leq c^2$ as $S^2 \preceq c^2 I$.

2. We can prove for $c = 1$ and general statement will follow. Let $v = (v_1, \ldots, v_n)$ be in $B(0, 1)$. Define $\widetilde{v} = p(v - v_1 e)$, where $p = (v_2^2 + \ldots v_n^2)^{-\frac{1}{2}}$. Complete to orthogonal basis $e, \widetilde{v}, u_3, \cdots, u_n$. Define map:

$$\widetilde{S} = \begin{bmatrix} v_1 & p^{-1} & 0 & \cdots & 0 \\ p^{-1} & -v_1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \cdots & 1 \end{bmatrix}$$

$\widetilde{S}$ is symmetric. Also,

$$-I \preceq \begin{bmatrix} v_1 & p^{-1} \\ p^{-1} & -v_1 \end{bmatrix} \preceq I$$

due to Sylvester's criterion and $\|v\| \leq 1$. Then $-I \preceq \widetilde{S} \preceq I$. Finally put $S = U\widetilde{S}U^T$, where U - orthogonal matrix of transition from basis $e_1, \ldots, e_n$ to $e, \widetilde{v}, u_3, \ldots, u_n$. It is obvious that $v = Se$.

□

Lemma 5.4. *For any symmetric S such that*

$$\mu I \preceq S \preceq LI$$

and $v \in \mathbb{R}^n$

$$Sv \in B(\frac{\mu + L}{2}v, \frac{L - \mu}{2}\|v\|)$$

and for any u in such sphere exists symmetric S, $\mu I \preceq S \preceq LI$ such that $u = Sv$.

Proof. $v = \|v\|Ue$, where U - orthogonal matrix. Then

$$Sv = \|v\|UU^TSUe = \|v\|U((U^TSU - \frac{L+\mu}{2}I) + \frac{L+\mu}{2}I)e$$

Here $-\frac{L-\mu}{2}I \preceq \widetilde{S} = U^TSU - \frac{L+\mu}{2}I \preceq \frac{L-\mu}{2}I$ any such symmetric matrix. By lemma 5.3 we obtain:

$$Sv = \|v\|\frac{L - \mu}{2}w + \frac{L + \mu}{2}v$$

where $\|w\| \leq 1$. □

Using auxiliary lemmas we can rewrite gradient of the function in the following form:

Lemma 5.5. *For f in $F^2_{\mu,L}$*

$$-\nabla f(x) = -\frac{L + \mu}{2}(x - x^*) + \frac{L - \mu}{2}\|x - x^*\|v(x)$$

where $v(x) \in B(0, 1)$ and x^* - optimum of f.

Proof.

$$\nabla f(x) = \nabla f(x) - \nabla f(x^*) = \int_0^1 H(x^* + t(x - x^*))dt(x - x^*)$$

Here

$$\mu I \preceq S = \int_0^1 H(x^* + t(x - x^*)) \preceq LI$$

and S - symmetric. And indeed we can gain any such matrix and lemma follows by lemma 5.4. □

Note, that lemma 5.5 shows that at any point $x \in \mathbb{R}^n$ vector $v(x) \in B(0,1)$ can be chosen arbitrary. Indeed, for any such v there exists a function f under assumption 1 for which $\nabla f(x)$ takes form as in lemma 5.5 with specified v. Though, there is some dependence of vectors $v(x), v(y)$ at any two given points, which takes place from assumption 3 and which we do not study here.

5.2. PDM analysis

Consider PDM algorithm:

$$\phi_{t+1} = \sum_{i=0}^{k} a_i \theta_{t+1,i}$$

$$\theta_{t+1,i} = \theta_{t+1,i-1} - \gamma F(\theta_{t+1,i-1})$$

$$\theta_{t+1,0} = \phi_t$$

Consider expression:

$$\theta_{t+1,i} - \phi^* = (1 - \gamma\frac{L+\mu}{2})(\theta_{t+1,i-1} - \phi^*) + \gamma\frac{L-\mu}{2}\|\theta_{t+1,i-1} - \phi^*\| v(\theta_{t+1,i-1}) + \gamma\xi_{t+1,i-1} \quad (6)$$

which follows from lemma 5.5.

Let's define a filtration $F_0 \preceq F_1 \preceq \ldots \preceq F_k$ with $F_i = \sigma(\theta_{t+1,i})$.

For PDM consider equality:

$$E\|\phi_{t+1} - \phi^*\|^2 \mid F_0 = \sum_{i=0}^{k} a_i^2 E\|\theta_{t+1,i} - \phi^*\|^2 \mid F_0 + \sum_{i<j} 2a_i a_j E(\theta_{t+1,i} - \phi^*, \theta_{t+1,j} - \phi^*) \mid F_0 \quad (7)$$

Then we evaluate first and second part of the equality 7 separately:

Lemma 5.6.

$$E\|\theta_{t+1,i} - \phi^*\|^2 \leq (1 - \gamma\mu)^2 E\|\theta_{t+1,i-1} - \phi^*\|^2 + \gamma^2\sigma^2$$

Proof. From expression 6 we get:

$$E\|\theta_{t+1,i} - \phi^*\|^2 \mid F_{i-1} \leq (1 - \gamma\frac{L+\mu}{2})^2\|\theta_{t+1,i-1} - \phi^*\|^2 + \gamma^2\frac{(L-\mu)^2}{4}\|\theta_{t+1,i-1} - \phi^*\|^2 + \gamma(L-\mu)(1 - \gamma\frac{L+\mu}{2})\|\theta_{t+1,i-1} - \phi^*\|(\theta_{t+1,i-1} - \phi^*, v(\theta_{t+1,i-1})) + \gamma^2\sigma^2 \leq (1 - \gamma\mu)^2\|\theta_{t+1,i-1} - \phi^*\|^2 + \gamma^2\sigma^2$$

where we used the inequality $(\theta_{t+1,i-1} - \phi^*, v(\theta_{t+1,i-1})) \leq \|\theta_{t+1,i-1} - \phi^*\|$. □

Lemma 5.7.

$$E\|\theta_{t+1,j} - \phi^*\|^2 \mid F_i \leq (1 - \gamma\mu)^{2(j-i)}\|\theta_{t+1,i} - \phi^*\|^2 + \gamma^2\sigma^2\frac{2(j-i)}{2-\gamma\mu}$$

Proof. Firstly, by applying lemma 5.6 multiple times we get:

$$E\|\theta_{t+1,j} - \phi^*\|^2 \mid F_i \leq (1 - \gamma\mu)^{2(j-i)}\|\theta_{t+1,i} - \phi^*\|^2 + \gamma^2\sigma^2\frac{1 - (1-\gamma\mu)^{2(j-i)}}{1 - (1-\gamma\mu)^2}$$

Then by Bernoulli inequality:

$$(1 + x)^n \geq 1 + nx \text{ for } x \geq -1$$

lemma follows. □

Lemma 5.8. *For $j > i$*

$$E(\theta_{t+1,i} - \phi^*, \theta_{t+1,j} - \phi^*) \mid F_i \leq \|\theta_{t+1,i} - \phi^*\|^2 + \gamma^3\sigma^2\frac{L-\mu}{4}\frac{(j-i)(j-i-1)}{2-\gamma\mu}$$

Proof. By expression 6 we get:

$$E(\theta_{t+1,i} - \phi^*, \theta_{t+1,j} - \phi^*) \mid F_{j-1} = (1 - \gamma\frac{L+\mu}{2})(\theta_{t+1,i} - \phi^*, \theta_{t+1,j-1} - \phi^*) + \gamma\frac{L-\mu}{2}\|\theta_{t+1,j-1} - \phi^*\|(\theta_{t+1,i} - \phi^*, v(\theta_{t+1,j-1}))$$

Further, we get:

$$\begin{aligned} &E(\theta_{t+1,i} - \phi^*, \theta_{t+1,j} - \phi^*) \mid F_{j-1} \\ &\leq (1 - \gamma\frac{L+\mu}{2})(\theta_{t+1,i} - \phi^*, \theta_{t+1,j-1} - \phi^*) + \gamma\frac{L-\mu}{2}\|\theta_{t+1,i} - \phi^*\|\|\theta_{t+1,j-1} - \phi^*\| \\ &\leq (1 - \gamma\frac{L+\mu}{2})(\theta_{t+1,i} - \phi^*, \theta_{t+1,j-1} - \phi^*) + \gamma\frac{L-\mu}{4}\|\theta_{t+1,j-1} - \phi^*\|^2 \\ &\qquad + \gamma\frac{L-\mu}{4}\|\theta_{t+1,i} - \phi^*\|^2 \quad (8) \end{aligned}$$

Then applying inequality 8 multiple times we gain:

$$\begin{aligned} E(\theta_{t+1,i} - \phi^*, \theta_{t+1,j} - \phi^*) \mid F_i \leq \big((1 - \gamma\frac{L+\mu}{2})^{j-i} \\ + \gamma\frac{L-\mu}{4}\sum_{l=i+1}^{j}(1 - \gamma\frac{L+\mu}{2})^{j-l}\big)\|\theta_{t+1,i} - \phi^*\|^2 \\ + \gamma\frac{L-\mu}{4}\sum_{l=i+1}^{j}(1 - \gamma\frac{L+\mu}{2})^{j-l}E\|\theta_{t+1,l-1} - \phi^*\|^2 \mid F_i \end{aligned}$$

Further, by applying lemma 5.7 to $E\|\theta_{t+1,l-1} - \phi^*\|^2 \mid F_i$:

$$E(\theta_{t+1,i} - \phi^*, \theta_{t+1,j} - \phi^*) \mid F_i \leq \big((1 - \gamma\frac{L+\mu}{2})^{j-i} + \gamma\frac{L-\mu}{2}\sum_{l=i+1}^{j}(1 - \gamma\frac{L+\mu}{2})^{j-l}\big)\|\theta_{t+1,i} - \phi^*\|^2 + \gamma\frac{L-\mu}{4}\sum_{l=i+1}^{j}(1 - \gamma\frac{L+\mu}{2})^{j-l}\gamma^2\sigma^2\frac{2(l-i-1)}{2-\gamma\mu}$$

Continuing:

$$\begin{aligned}
&E(\theta_{t+1,i} - \phi^*, \theta_{t+1,j} - \phi^*) \mid F_i \leq \\
&((1 - \gamma\frac{L+\mu}{2})^{j-i} + \gamma\frac{L-\mu}{2}\sum_{l=i+1}^{j}(1 - \gamma\frac{L+\mu}{2})^{j-l})\|\theta_{t+1,i} - \phi^*\|^2 + \\
&\gamma^3\sigma^2\frac{L-\mu}{4}\sum_{l=i+1}^{j}\frac{2(l-i-1)}{2-\gamma\mu} \\
&\leq ((1 - \gamma\frac{L+\mu}{2})^{j-i} + \frac{L-\mu}{L+\mu}(1 - (1 - \gamma\frac{L+\mu}{2})^{j-i}))\|\theta_{t+1,i} - \phi^*\|^2 + \\
&\gamma^3\sigma^2\frac{L-\mu}{4}\frac{(j-i)(j-i-1)}{2-\gamma\mu} \leq \\
&\|\theta_{t+1,i} - \phi^*\|^2 + \gamma^3\sigma^2\frac{L-\mu}{4}\frac{(j-i)(j-i-1)}{2-\gamma\mu}
\end{aligned}$$

□

Then main theorem follows:

Theorem 5.9. *For* $\gamma \leq \frac{2}{L+\mu}$

$$E\|\phi_{t+1} - \phi^*\|^2 \leq (\sum_{i=0}^{k} a_i^2(1-\gamma\mu)^{2i} + \sum_{i<j} 2a_ia_j(1-\gamma\mu)^{2i})E\|\phi_t - \phi^*\|^2 + \gamma^2\sigma^2(\sum_{i=0}^{k} a_i^2\frac{2i}{2-\gamma\mu} + \sum_{i<j} 2a_ia_j(\frac{2i}{2-\gamma\mu} + \gamma\frac{L-\mu}{4}\frac{(j-i)(j-i-1)}{2-\gamma\mu}))$$

Note that

$$\sum_{i=0}^{k} a_i^2(1-\gamma\mu)^{2i} + \sum_{i<j} 2a_i a_j(1-\gamma\mu)^{2i} < (\sum_{i=0}^{k} a_i)^2 = 1$$

This implies linear convergence to limit noise.

Proof.

$$E\|\phi_{t+1}-\phi^*\|^2 \mid F_0 = \sum_{i=0}^{k} a_i^2 E\|\theta_{t+1,i}-\phi^*\|^2 \mid F_0 + \sum_{i<j} 2a_i a_j E(\theta_{t+1,i}-\phi^*, \theta_{t+1,j}-\phi^*) \mid F_0 \leq$$

$$\sum_{i=0}^{k} a_i^2(1-\gamma\mu)^{2i}\|\phi_t-\phi^*\|^2 + \sum_{i<j} 2a_i a_j E\|\theta_{t+1,i}-\phi^*\|^2 \mid F_0 +$$

$$\gamma^2\sigma^2(\sum_{i=0}^{k} a_i^2 \frac{2i}{2-\gamma\mu} + \sum_{i<j} 2a_i a_j \gamma \frac{L-\mu}{4}\frac{(j-i)(j-i-1)}{2-\gamma\mu}) \leq$$

$$(\sum_{i=0}^{k} a_i^2(1-\gamma\mu)^{2i} + \sum_{i<j} 2a_i a_j(1-\gamma\mu)^{2i})E\|\phi_t-\phi^*\|^2$$

$$+\gamma^2\sigma^2(\sum_{i=0}^{k} a_i^2 \frac{2i}{2-\gamma\mu} + \sum_{i<j} 2a_i a_j(\frac{2i}{2-\gamma\mu} + \gamma\frac{L-\mu}{4}\frac{(j-i)(j-i-1)}{2-\gamma\mu}))$$

□

5.3. Lower bounds

In this section we support our statement that Lookahead works worse than SGD in strongly convex case as well. Next lemma shows tight estimates for SGD:

Lemma 5.10. *For SGD*

$$E\|\phi_{t+1}-\phi^*\|^2 \leq (1-\gamma\mu)^2 E\|\phi_t-\phi^*\|^2 + \gamma^2\sigma^2$$

$$V^*_{SGD} \leq \frac{\sigma^2}{\mu^2}\frac{1-s}{1+s}$$

where $s = 1 - \gamma\mu$. This estimate is indeed precise and inequalities turns into equalities for function $f(x) = \frac{\mu}{2}\|x - \phi^*\|^2$.

Proof. Due to lemma 5.5 we have:

$$\phi_{t+1} - \phi^* = (1 - \gamma\frac{L+\mu}{2})(\phi_t - \phi^*) + \gamma\frac{L-\mu}{2}\|\phi_t - \phi^*\|v(\phi_t) + \gamma\xi$$

Then:

$$E\|\phi_{t+1} - \phi^*\|^2|\sigma(\phi_t) \leq (1 - \gamma\frac{L+\mu}{2})^2\|\phi_t - \phi^*\|^2 + \gamma^2\frac{(L-\mu)^2}{4}\|\phi_t - \phi^*\|^2 + \gamma(L-\mu)(1 - \gamma\frac{L+\mu}{2})\|\phi_t - \phi^*\|(\phi_t - \phi^*, v(\phi_t)) + \gamma^2\sigma^2 \leq (1 - \gamma\mu)^2\|\phi_t - \phi^*\|^2 + \gamma^2\sigma^2 \quad (9)$$

where the last inequality follows from $(\phi_t - \phi^*, v(\phi_t)) \leq \|\phi_t - \phi^*\|$. Note that these two inequalities turn into equalities on the function $f(x) = \frac{\mu}{2}\|x - \phi^*\|^2$ when noise under assumption 2 has exactly variance σ^2. Taking limit in inequality 9 we get:

$$V^*_{SGD} \leq \frac{\sigma^2}{\mu^2}\frac{1-s}{1+s}$$

□

Also for Lookahead we show following lower bound:

Lemma 5.11. *For Lookahead*

On function $f(x) = \frac{\mu}{2}\|x - \phi^*\|^2$ and when noise under assumption 2 has exactly variance σ^2

$$E\|\phi_{t+1} - \phi^*\|^2 = ((1-\alpha) + \alpha(1-\gamma\mu)^k)^2 E\|\phi_t - \phi^*\|^2 + \gamma^2\sigma^2\alpha^2\frac{1-(1-\gamma\mu)^{2k}}{1-(1-\gamma\mu)^2}$$

$$V^*_{LA} = \frac{\sigma^2}{\mu^2}\frac{1-t}{1+t}\frac{(1-c)(1+c^k)}{(1-c^k)(1+c)}$$

with $c = 1 - \gamma\mu$, $t = (1-\alpha) + \alpha c^k$.

Proof. Completely repeat the proof of convergence of Lookahead for quadratic case. See section 4.2. □

Note that V^*_{SGD} and V^*_{LA} in strongly convex case take the same form as in the quadratic case except for the multiplier μ^{-2}. This shows that just like in quadratic case in strongly convex case Lookahead experiences worse rate of convergence than SGD with limit noise being fixed in the worst case sense.

6. Experiments

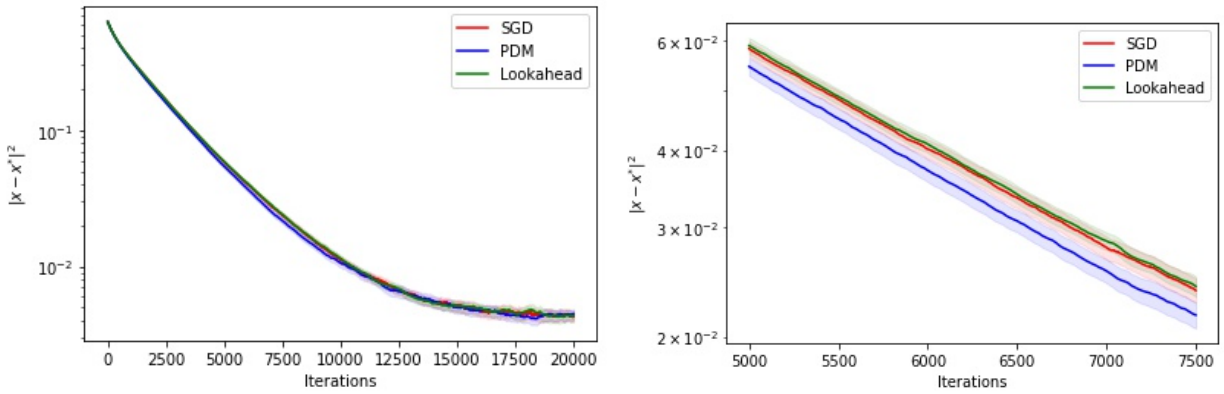

Figure 2: Right figure is just enlarged version of the left one. We train logistic regression with regularization on dataset German. Parameters of the algorithms are picked up so as to their limit noises are approximately the same. Here, their limit noises approximately equal $v^* = 0.0044$. For Lookahead and PDM $k = 20$.

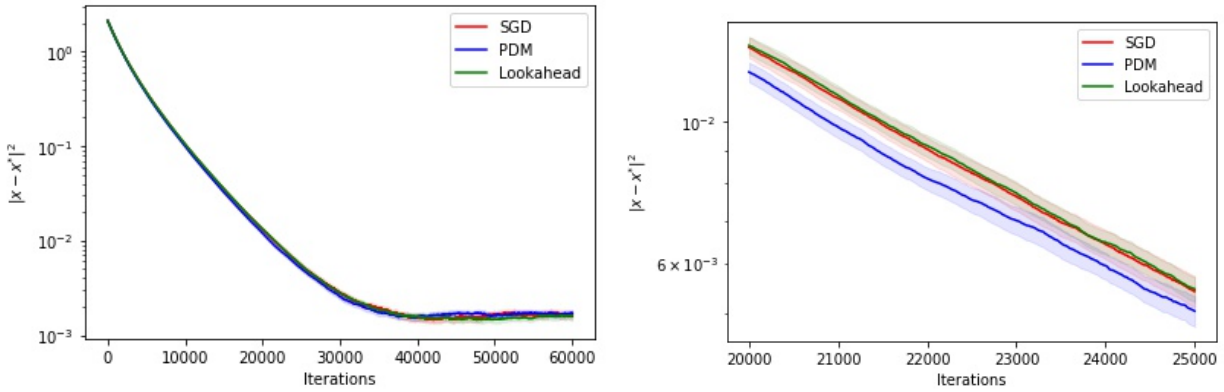

Figure 3: Right figure is just enlarged version of the left one. We train logistic regression with regularization on dataset Mushrooms. Parameters of the algorithms are picked up so as to their limit noises are approximately the same. Here, their limit noises approximately equal $v^* = 0.0016$. For Lookahead and PDM $k = 20$.

We trained logistic regression model with regularization on two datasets Mushrooms and German from Chang and Lin ((2011)) on the task of 2-label classification. Methods' parameters are picked so as they have approximately equal

limit noises. Coefficients for PDM are taken as:

$$\begin{cases} a_0 = (1-\alpha)(t + \frac{1-t}{k}) \\ a_i = \frac{1-t}{k}, \quad i = 1, \cdots, k-1 \\ a_k = \alpha(t + \frac{1-t}{k}) \end{cases}$$

where $\alpha \in [0, 1]$ is new hyperparameter. Actually, coefficients are taken in accordance with optimal coefficients for quadratic case with the only difference that we do not know c in the system from lemma 4.5 so we need new hyperparameter. In the Figures 2 and 3 we see that PDM experiences superior rate of convergence.

7. Conclusion

We introduced new algorithm called Polynomial Dynamic Modification and showed that it experiences faster convergence rate on the quadratic stochastic problem compared to SGD and Lookahead with desired noise near solution being fixed. Moreover, we prove linear convergence of PDM to some limit noise on the class of stongly convex twicely differentiable functions.

8. Acknowlegements

We express gratitude to Alexsandr Beznosikov for his supervision during the work on this thesis.

References

C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM transactions on intelligent systems and technology (TIST)*, 2(3):1–27, 2011.

E. Gorbunov, F. Hanzely, and P. Richtárik. A unified theory of sgd: Variance reduction, sampling, quantization and coordinate descent. In *International Conference on Artificial Intelligence and Statistics*, pages 680–690. PMLR, 2020.

N. B. Jianyu Wang, Vinayak Tantia and M. Rabbat. Lookahead converges to stationary points of smooth non-convex functions. 2020.

G. H. Michael R. Zhang, James Lucas and J. Ba. Lookahead optimizer: k steps forward, 1 step back. 2019.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *Society for Industrial and Applied Mathematics*, 19:1574–1609, 01 2009. doi: 10.1137/070704277.

Y. Nesterov. *Lectures on convex optimization*, volume 137. Springer, 2018.

H. Robbins and S. Monro. A Stochastic Approximation Method. *The Annals of Mathematical Statistics*, 22(3):400 – 407, 1951. doi: 10.1214/aoms/1177729586. URL `https://doi.org/10.1214/aoms/1177729586`.

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

V. Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

S. Vaswani, F. Bach, and M. Schmidt. Fast and faster convergence of sgd for over-parameterized models and an accelerated perceptron. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1195–1204. PMLR, 2019.