# A simpler approach to obtaining an O(1/t) convergence rate for the projected stochastic subgradient method

**0816173 Ying-Tu Chen, 0680838 Ramanjaneyulu Lavanuru**
Department of Computer Science
National Yang Ming Chiao Tung University
{st993201.cs08, ramanjaneyulueecs.ee06}@nycu.edu.tw

## 1 Introduction

In this research paper, the authors propose a new average technique for the projected stochastic subgradient method based on a weighted average for each iteration $w_t$ at iteration $t$ with $t+1$. They demonstrate a convergence rate of $O\left(\frac{1}{t}\right)$ using their new analysis method, which is both easy to prove and implement. The proposed averaging schemes are particularly suitable and convenient for implementation in the online setting. While the main focus of the paper is on the non-smooth case, the proposed method also has implications for the smooth case where the set is strongly convex. In these cases, using the proposed method for averaging with large step sizes leads to better and more robust rates, and is easy to implement. The proposed scheme shows superior performance compared to existing techniques based on experimental results.

## 2 Problem Formulation

We consider a strong convex function $f$ defined on a convex set $K$. We denote by $\mu$ its strong convexity constant. We consider a stochastic approximation scenario where only unbiased estimates of sub-gradients $f$ of $f$ are available, with the projected stochastic sub-gradient method.

More precisely, we assume that we have an increasing sequence of $\sigma$-fields$\left(F_t\right)_t \geqslant 0$,

$$w_t = \Pi_K \left(w_{t-1} - \gamma_t g_t\right) \tag{1}$$

where

$(a)$ $\Pi_K$ is the orthogonal projection on $K$.

$(b)$ $\mathbb{E}\left(g_t \mid F_{t-1}\right)$ is almost surely a subgardient of $f$ at $w_{t-1}$ $\Big($which we denote$f^{'}\left(w_{t-1}\right)\Big)$,

$(c)$ $\mathbb{E}\left(\|g_t\|\right) \leqslant B^2$(finite variance condition).

We denote by $w_*$ the unique minimizer of $f$ on $K$.

Motivation:

Our main motivating example is the support vector machine(SVM), where the pairs $\left(x_t, y_t\right)$ for $t \geqslant 1$ are independent and identically distributed and $f\left(w\right) = \mathbb{E}\, l\left(y, w^\intercal x\right) + \frac{\mu}{2}\|w\|^2$, where $\left(y, u\right)$ is the Lipschitz-continuous convex loss function (with respect to the second variable) and $K$ is the whole space (unconstrained setup). We then have $g_t = l^{'}\left(y_t, w_{t-1}^\intercal x_t\right) x_t + \mu w_{t-1}$, where $l^{'}\left(y, u\right)$ denotes any subgradient with respect to the second variable.

If we make the additional assumption that $\mathbb{E}\,\|x\|^2$ is the finite, then this setup satisfies the assumptions above with $B^2 = 4L_l^2\,\mathbb{E}^2$, where $L_l$ is the Lipschitz constant for $l$.

Alternatively, we can consider $K$ to be a compact convex subset. This is used in particular in a projected version of the stochastic subgradient method for SVM. In this case we can take

$$B^2 = \left(L_l\sqrt{\mathbb{E}\left\|x\right\|^2} + \mu max_{x\in K}\|w\|\right)^2$$

To test the empirical performance of the averaging scheme, we performed a series of experiments using the support vector machine optimization problem as shown in below mathematical expression.

$$\min_w \frac{\lambda}{2}\|w\|^2 + \frac{1}{n}\sum_{i=1}^{n} max\left\{0, 1 - y_i w^\intercal x_i\right\}$$

where $x_i$ is in an Euclidean space and $y_i \in \{-1, 1\}$.

## 3 Algorithms

This paper does not introduce any new algorithms. Rather, it discusses how making small changes to the learning rate and using average methods for the weight can affect model performance. Hence, we will focus on how to implement those average methods. As we will discuss in Section 4 and Section 5, there are methods that use the (t+1) as its weight to average the weight for all iterations. Therefore, we can write the equations as follows:

$$\bar{w}_T = \frac{2}{T\left(T+1\right)}\sum_{t=0}^{T-1}\left(t+1\right)w_t$$

This can be simplified as:

$$\bar{w}_T = \left(1 - \rho_T\right)\bar{w}_{T-1} + \rho_T w_T.$$

Where $\rho_T$ is $\frac{2}{T+2}$.

Furthermore, we can use the same methods to simplify another model that uses $(t+1)^2$ as its weight to average the weight for all iterations. We will have the same equation:

$$\bar{w}_T = \left(1 - \rho_T\right)\bar{w}_{T-1} + \rho_T w_T.$$

but $\rho_T$ is $\frac{6(T+1)}{(2T+3)(T+2)}$.

## 4 Theoretical Analysis

Convergence analysis:

$$\begin{aligned}\|w_t - w^*\|^2 &\leqslant \|w_{t-1} - \gamma_t g_t - w^*\|^2 \\ &= \|w_{t-1} - w^*\|^2 + \gamma_t^2\|g_t\|^2 - 2\gamma_t\left(w_{t-1} - w^*\right)^\intercal g_t\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left(\|w_t - w^*\|^2 \mid F_{t-1}\right) &\leqslant \|w_{t-1} - w^*\|^2 + \gamma_t^2\,\mathbb{E}\left(\|g_t\|^2 \mid F_{t-1})\right) - 2\gamma_t\left(w_{t-1} - w^*\right)^\intercal f^{'}\left(w_{t-1}\right) \\ &\leqslant \|w_{t-1} - w^*\|^2 + \gamma_t^2\,\mathbb{E}\left(\|g_t\|^2 \mid F_{t-1}\right) - 2\gamma_t\left[f\left(w_{t-1}\right) - f\left(w^*\right) + \frac{\mu}{2}\|w_{t-1} - w^*\|^2\right]\end{aligned}$$

The last inequality is obtained from the $\mu-$strong convexity of f. Thus, by re-arraigning the function values on the LHS and taking exceptions on both sides, we get as shown in below:

$$2\gamma_t\left[\mathbb{E}\,f\left(w_{t-1}\right) - f\left(w^*\right)\right] \leqslant \gamma_t^2\,\mathbb{E}\,\|g_t\|^2 + \left(1 - \mu\gamma_t\right)\mathbb{E}\,\|w_{t-1} - w^*\|^2 - \mathbb{E}\,\|w_t - w^*\|^2$$

$$\mathbb{E}\, f\left(w_{t-1}\right)-f\left(w^{*}\right) \leqslant \frac{\gamma_{t} B^{2}}{2}+\frac{\gamma_{t}^{-1}-\mu}{2} \mathbb{E}\left\|w_{t}-w^{*}\right\|^{2}-\frac{\gamma_{t}^{-1}}{2} \mathbb{E}\left\|w_{t}-w^{*}\right\|^{2} \tag{2}$$

Classical Analysis:

Let us consider $\gamma_t = \frac{1}{\mu t}$, then inequality $(2)$ becomes as shown in below,

$$\mathbb{E}\, f\left(w_{t-1}\right)-f\left(w^{*}\right) \leqslant \frac{B^{2}}{2 \mu t}+\frac{\mu(t-1)}{2} \mathbb{E}\left\|w_{t-1}-w^{*}\right\|^{2}-\frac{\mu t}{2} \mathbb{E}\left\|w_{t}-w^{*}\right\|^{2}$$

and by summing from $t = 1 tot = T$, we obtain the following mathematical expression.

$$\begin{aligned}
\mathbb{E}\, f\left(\frac{1}{T} \sum_{t=1}^{T} w_{t-1}\right)-f\left(w^{*}\right) &\leqslant \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\, f\left(w_{t-1}\right)-f\left(w^{*}\right) \\
&\leqslant \frac{B^{2}}{2 \mu T} \sum_{t=1}^{T} \frac{1}{t}+\frac{\mu}{2 T}\left[0-T\, \mathbb{E}\left\|w_{T}-w^{*}\right\|^{2}\right] \\
&\leqslant \frac{B^{2}}{2 \mu T}(1+\ln T)
\end{aligned}$$

The first line used the convexity of f; the second line is obtained from a telescoping sum.

We also obtain the following mathematical expression.

$$\mathbb{E}\left\|w_{T}-w^{*}\right\|^{2} \leqslant \frac{B^{2}}{\mu^{2} T}(1+\ln T)$$

New analysis:

Let us consider $\gamma_t = \frac{2}{\mu(t+1)}$ and multiplying inequality$(2)$ by t, then we obtain the following mathematical expression.

$$\begin{aligned}
t\left[\mathbb{E}\, f\left(w_{t-1}\right)-f\left(w^{*}\right)\right] &\leqslant \frac{t B^{2}}{\mu(t+1)}+\frac{\mu}{4}\left[t(t-1)\, \mathbb{E}\left\|w_{t-1}-w^{*}\right\|^{2}-t(t+1)\, \mathbb{E}\left\|w_{t}-w^{*}\right\|^{2}\right] \\
&\leqslant \frac{B^{2}}{\mu}+\frac{\mu}{4}\left[t(t-1)\, \mathbb{E}\left\|w_{t-1}-w^{*}\right\|^{2}-t(t+1)\, \mathbb{E}\left\|w_{t}-w^{*}\right\|^{2}\right]
\end{aligned}$$

By summing from $t = 1$ to $t = T$ these t-weighted inequalities, we obtain a similar telescoping sum, but this time the term with $B^2$ stays constant across the sum as shown in the below equation $(3)$.

$$\sum_{t=1}^{T} t\left[\mathbb{E}\, f\left(w_{t-1}\right)-f\left(w^{*}\right)\right] \leqslant \frac{T B^{2}}{\mu}+\frac{\mu}{4}\left[0-T(T+1)\, \mathbb{E}\left\|w_{T}-w^{*}\right\|^{2}\right] \tag{3}$$

Thus

$$\mathbb{E}\, f\left(\frac{2}{T(T+1)} \sum_{t=0}^{T-1}(t+1) w_{t}\right)-f\left(w^{*}\right)+\frac{\mu}{2} \mathbb{E}\left\|w_{T}-w^{*}\right\|^{2} \leqslant \frac{2 B^{2}}{\mu(T+1)}$$

which implies that

$$\mathbb{E}\, f\left(\frac{2}{T(T+1)} \sum_{t=0}^{T-1}(t+1) w_{t}\right)-f\left(w^{*}\right) \leqslant \frac{2 B^{2}}{\mu(T+1)}$$

and

$$\mathbb{E}\left\|w_T - w^*\right\|^2 \leqslant \frac{4B^2}{\mu^2 \left(T+1\right)}$$

So by using the weighted average$\bar{w}_T = \frac{2}{(T+1)(T+2)} \sum_{t=0}^{\intercal} \left(t+1\right) w_t$ instead of a uniform average, we get a $O\left(\frac{1}{T}\right)$rate instead of $O\left(\frac{logT}{T}\right)$. Note that these averaging schemes are efficiently implemented in an online faction as shown in the below equation $(4)$.

$$\bar{w}_T = \left(1-\rho_T\right)\bar{w}_{T-1} + \rho_T w_T. \tag{4}$$

For the proposed weighted averaging scheme $\rho_T = \frac{2}{(T+2)}$ and compare with $\rho_T = \frac{1}{(T+1)}$ for the uniform averaging scheme.

# 5 Experiments

## 5.1 Architecture of the System

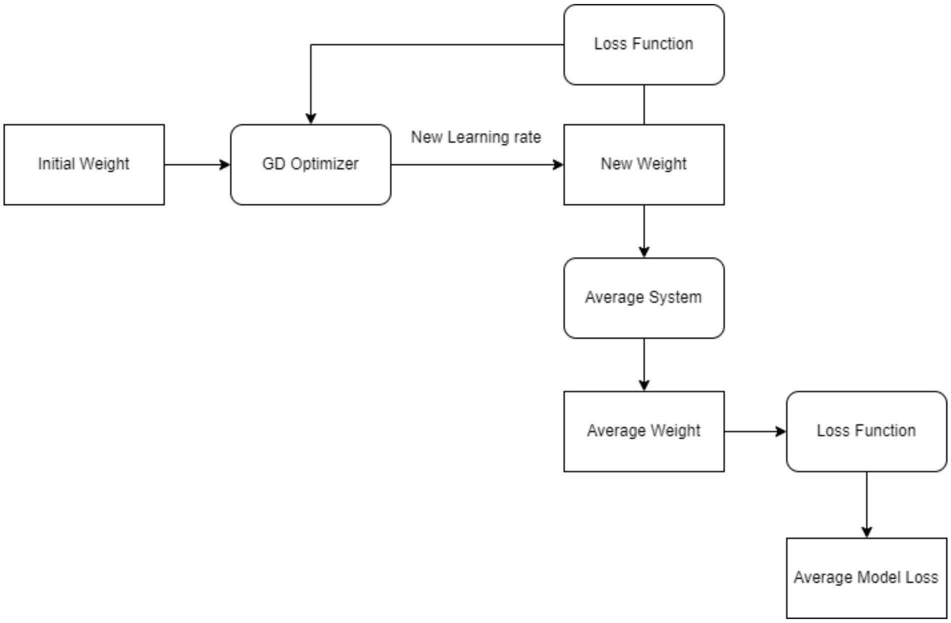

Figure 1: Architecture of the System

The system begins by working with the initial weights and uses a gradient descent optimizer to update the weights on our support vector machine (SVM) model. To accelerate the convergence speed, the author of this paper made some simple changes. Firstly, they continually updated the learning rate at each iteration. Then, they used an average system to calculate the average weight, which was used to calculate the loss for evaluating the model's performance.

## 5.2 Objective Function

The objective function used to evaluate the model's performance includes the Hinge loss and an L2 regularization term.

$$\min_w \frac{\lambda}{2}||w||^2 + \frac{1}{N}\sum_{i=1}^{N} \max(0, 1 - y_i \cdot w^{\intercal} x_i)$$

where $x_i$ is in an Euclidean space and $y_i \in \{-1, 1\}$. As we have the whole space as our K, so we do not have to do the projection.

### 5.3 Learning Rate type

In this experimental section, we use the two methods mentioned in the section 4. to dynamically adjust the learning rate at each iteration.

- $\frac{1}{t\mu}$
- $\frac{2}{(t+1)\mu}$

### 5.4 Average System type

Our experiments compare the following type of average system:

- $0$ : We use the original model weight.
- $1$ : We average all iterates with uniform weight.
- $0.5$ : We average the second half of the iterates with uniform weights.
- $D$ : We average all iterates since the last iteration that was a power of 2 with uniform weight.
- $W$ : We average all iterates with a weight of $t + 1$, as discussed in the section 4.
- $W^2$ : We average all iterates with a weight of $(t + 1)^2$.

### 5.5 Result and Discussion

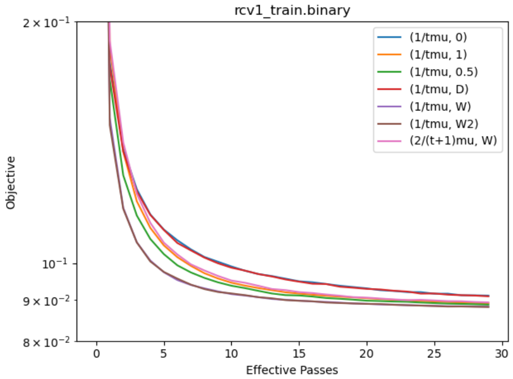

Figure 2: rcv1

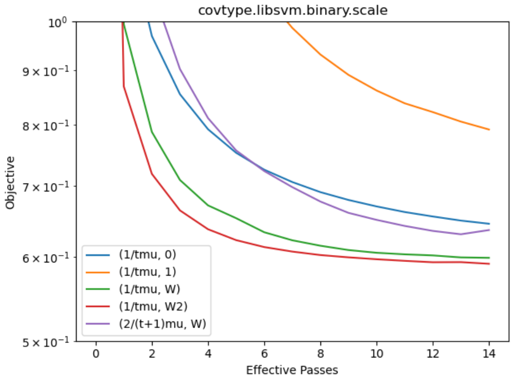

Figure 3: covertype

In these two images, we can see that the model with the learning rate $\frac{1}{t\mu}$ performs better than the model with a different learning rate. When evaluating the performance of the average methods, we can see that some methods have worse performance than the original model, as shown in Figure 3. Additionally, it is clear that the model with $W^2$ consistently performs better than the model with $W$."

## 6 Conclusion

Based on our review of the proof and experimentation with the simple changes to the gradient descent algorithm, we have reached the following conclusions:

- The learning rate used in the classical analysis performs better than the learning rate used in the new analysis.
- According to the paper, it is stated that models 0 and 1 have the same convergence rate of $O(\frac{\log t}{t})$, while the remaining models ($0.5$, $D$, $W$, and $W^2$) have a convergence rate of $O(\frac{1}{t})$.
- In this paper, no constraints were set for optimization, so we did not implement any projection in our experiments.

- It would be worthwhile to further explore how the weight on averaging can improve model performance in future analyses.

To summarize, the following image provides an intuitive explanation of the concept of averaging methods:

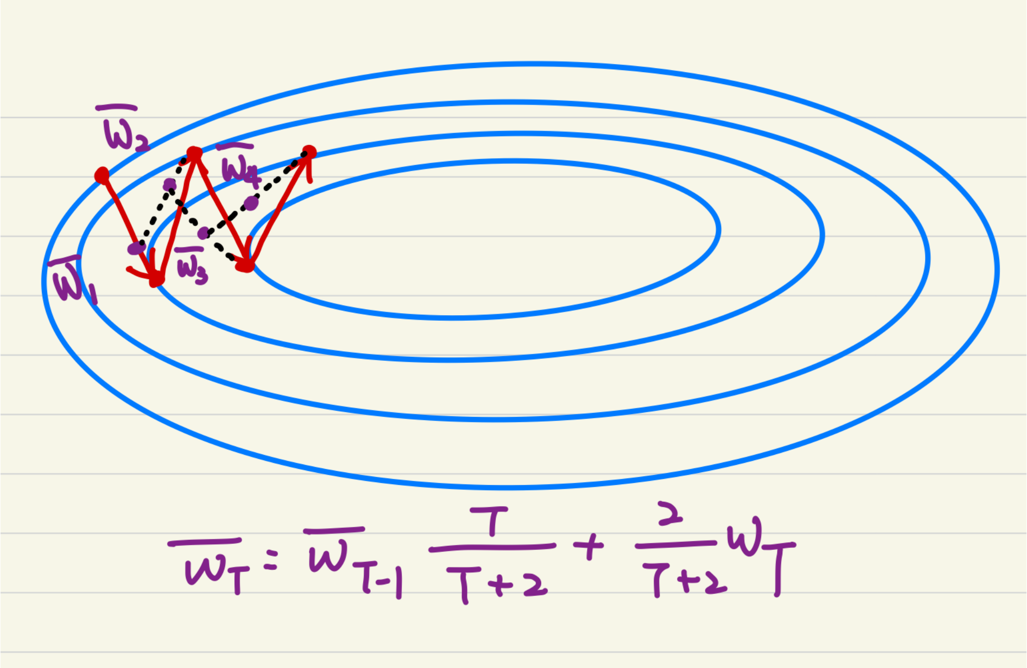
$\overline{W}_2$
$\overline{W}_4$
$\overline{W}_3$
$\overline{W}_1$
$\overline{w}_T = \overline{w}_{T-1} \frac{T}{T+2} + \frac{2}{T+2} w_T$

Figure 4: The 'zig-zag' problem

Gradient descent algorithms can sometimes experience the 'zig-zag' problem, where the optimization process becomes oscillatory instead of steadily converging. However, using the W average method can reduce this problem and therefore speed up the optimization process.

## References

[1] Simon Lacoste-Julien, Mark Schmidt, and Francis Bach. A simpler approach to obtaining an o(1/t) convergence rate for the projected stochastic subgradient method, 2012.