Abstract

We introduce a technique for tuning the learning rate scale factor of any base optimization algorithm and schedule automatically, which we call MECHANIC. Our method provides a practical realization of recent theoretical reductions for accomplishing a similar goal in online convex optimization. We rigorously evaluate MECHANIC on a range of large scale deep learning tasks with varying batch sizes, schedules, and base optimization algorithms. These experiments demonstrate that depending on the problem, MECHANIC either comes very close to, matches or even improves upon manual tuning of learning rates.

1 Introduction

Modern deep learning is driven by first-order stochastic optimization algorithms. These are algorithms that are designed to solve the classical stochastic optimization problem:

$$\min F(x) = \min \mathbb{E}[f(x,z)]$$

where $z$ is a minibatch of examples, $x \in \mathbb{R}^d$ is the model parameters, and $f$ is the loss incurred by using weights $x$ on the minibatch $z$. A first-order algorithm follows the protocol:

1. Output a $t$th iterate $x_t$.
2. Sample a random minibatch $z_t$.
3. Compute $g_t = \nabla f(x_t, z_t)$ (the gradient is taken with respect to $x_t$ only).
4. Possibly update some internal algorithm state based upon $g_t$ in preparation for computing the next iterate $x_{t+1}$.

The prototypical such optimization algorithm is stochastic gradient descent (SGD), which exemplifies the attractive features of this approach: it is computationally cheap (running in $O(d)$ time per update), and with proper tuning obtains minimax optimal convergence guarantees \cite{Hazan2016,Bottou2016}. Modern practice makes use of a wide range of variants of SGD, such SGD with momentum, AdaGrad \cite{Duchi2011}, Adam \cite{Kingma2014}, AdamW \cite{Loshchilov2017} or Lion \cite{Lion2021}. Interest in such improvements to SGD is driven by the increasing computational demands of training large neural networks: better optimization means cheaper training, which translates to significant savings in terms of time, cost, and environmental impact.

Most modern algorithms for training neural networks are equipped with a scalar “scale factor” or “learning rate” hyperparameter $s \in \mathbb{R}$. Roughly speaking, these algorithms produce iterates of the form $x_{t+1} = x_t + s \cdot u_t$ where $u_t$ is some update vector produced as a function of the observed gradients $g_1, \ldots, g_t$ (we will use bold font for vectors in $\mathbb{R}^d$ like $u$ and normal font for all other quantities like $s$). As an example, the classical SGD algorithm sets $u_t = -\eta_t g_t$ for some sequence of scalars $\{\eta_t\}$ typically called the schedule. The formula for the SGD update is:

$$x_{t+1} = x_1 - s \sum_{i=1}^{t} \eta_i g_i. \quad (1)$$
The process of selecting the optimal \( s \) is called “tuning”, and is a key resource sink in machine learning. The typical approach is simply to try many possibilities to find the empirically optimal \( s \), which requires multiple expensive training runs. This paper introduces a technique for choosing \( s \) automatically on-the-fly in order to avoid this expense.

Our procedure, which we call \textsc{mechanic}, is a generic wrapper around any base optimization algorithm (\textsc{base}) that produces a new optimizer which does not require tuning of the scalar \( s \). The base optimization algorithm is allowed to make any kind of update (for example, it may use any kind of schedule, preconditioner or weight decay). If the base optimization algorithm is allowed to make any kind of update (for example, it may use any algorithm (e.g. \cite{20, 21, 16, 15}), then the wrapper will produce a scalar \( s_t \) and set the \( t \)th iterate of the wrapped algorithm to be \( x_t = x_t^{\text{base}} + s_t (x_t^{\text{base}} - x_t^{\text{base}}) \). As an example, suppose that \textsc{base} is the classical SGD algorithm with update equation (\ref{eq:1}). Then, given \( s_t \), we would set \( x_t = x_t^{\text{base}} - s_t \sum_{i=1}^{t-1} \eta_i g_i \). Disregarding for now the fact that the gradients \( g_i \) actually depend on the iterates \( x_i \), we see that \( x_t \) is what the \( t \)th iterate of SGD would have been if the schedule were scaled by \( s_t \).

Removing tuning of learning rate scalars is already a well-studied problem. One of the main attractions of early work in “adaptive” optimization such as AdaGrad and Adam \cite{3, 7, 4} is that these algorithms require less tuning than ordinary SGD. Over the last decade, a number of works have aimed to tackle this problem from both an empirical and theoretical perspective \cite{8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19}. An intuitive approach might take the route of “hypergradient descent”: that is, differentiating the update step of the optimization algorithm itself (e.g. \cite{8, 9}). Strangely, it appears to be difficult to prove that such schemes behave well: theory-based approaches often adopt rather different strategies. Instead, we start from known theory and propose a few important modifications to produce a simple and effective practical implementation. We then rigorously evaluate our algorithm on a variety of datasets. We emphasize that our primary contribution is not new theoretical development, but instead the translation between theory and practice, which involves fusing several known analytical techniques as well as subtle departures from theory.

Previous works that investigate deep learning performance of “learning-rate free” optimization inspired by theory (e.g. \cite{20, 21, 16, 15}) have already demonstrated impressive results. However, these works typically build “hand-crafted” algorithms that often blend theoretical analysis with specific empirically successful algorithms such as Adam. In contrast, our wrapper works well with any base algorithm and so can seamlessly integrate new empirical advances in optimization: one does not need intimate familiarity with the analysis of our approach to apply it to a new algorithm.

2 Background: Online Convex Optimization

We develop our formalism via \textit{online convex optimization} (\textsc{oco}) \cite{22, 23, 24}. \textsc{oco} is a popular framework for design and analysis of stochastic optimization algorithms. In brief, for each of \( T \) rounds (corresponding to \( T \) iterations of optimization), the \textsc{oco} algorithm must first output a \( t \)th iterate \( x_t \), after which the algorithm is presented with a \( t \)th loss function \( \ell_t \). Typically, one envisions the case \( \ell_t(x) = \ell(x, z_t) \) for some fixed loss \( \ell \) and new data point \( z_t \). The goal of an algorithm \textsc{alg} is to minimize the regret \( R^{\text{alg}}(\hat{x}) \):

\[
R^{\text{alg}}(\hat{x}) \triangleq \sum_{t=1}^{T} \ell_t(x_t) - \ell_t(\hat{x}).
\]

Many references focus primarily on the case \( \hat{x} = \arg\min_{x} \sum_{t=1}^{T} \ell_t(x) \) in order to consider the single scalar value \( \sup_{\hat{x}} R_T(\hat{x}) \) \cite{25, 26}, but we will employ the formulation of regret as a function above instead as it is strictly more general. When \( \ell_t \) is convex, then with \( g_t \triangleq \nabla \ell_t(x_t) \) (or, more generally when \( g_t \) is a subgradient of \( \ell_t \) at \( x_t \)), we have:

\[
R^{\text{alg}}(\hat{x}) \leq \sum_{t=1}^{T} \langle g_t, x_t - \hat{x} \rangle \triangleq R^{\text{alg}}_{\text{linear}}(\hat{x}).
\]

As a result, the vast majority of \textsc{oco} algorithms provide analysis that bounds only the linearized regret \( R^{\text{alg}}_{\text{linear}}(\hat{x}) \). Such algorithms do not need to observe the entire function \( \ell_t \) instead, they only make use of the gradients \( g_t \). That is, the \( t \)th output of \textsc{alg} (i.e. \( x_t \)) is purely a function of the previous sequence of gradients \( g_1, \ldots, g_{t-1} \) so that \textsc{alg} is a first-order algorithm.\footnote{This seems like a significant issue to disregard, but we will provide mathematical justification shortly.}
2.1 Learning the Scale in OCO

Just like stochastic optimization algorithms, most OCO algorithms also require a scale factor \( s \). In fact, many stochastic optimization algorithms (such as SGD and AdaGrad) are also OCO algorithms. Setting \( \eta_t = \eta \) for all \( t \), SGD ensures the regret bound:

\[
R^{\text{SGD}}(\bar{x}) \leq R^{\text{SGD}}_{\text{linear}}(\bar{x}) \leq O\left( \frac{\|\bar{x} - x_1\|^2}{\eta} \right).
\]

From this equation, one can deduce in hindsight that for any given \( \bar{x} \), the optimal value for \( s \) is

\[
\sqrt{\frac{n\eta}{\sum_{t=1}^{T} \|g_t\|^2}} - \frac{\|\bar{x} - x_1\|}{n\eta},
\]

which would provide the bound:

\[
R^{\text{SGD}}_{\text{WITH TUNED } s}(\bar{x}) \leq O\left( \|\bar{x} - x_1\|\sqrt{\sum_{t=1}^{T} \|g_t\|^2} \right).
\]

This result is minimax optimal \([27]\), but requires knowledge of the unknown optimal \( s \). Very recently, \([14, 16, 15]\) have produced algorithms that estimate the value of \( \|x_1^{\text{BASE}} - \bar{x}\| \) on-the-fly and use this estimate to quickly identify the optimal scaling value \( s \). These algorithms achieve impressive practical performance, but they require an understanding of the closed-form solution for the optimal \( s \) value above. Our goal is to learn the correct scaling regardless of the base algorithm.

To this end, we will leverage a scheme recently developed by \([28]\) that allows one to automatically tune the scale of a base OCO algorithm using another “meta” OCO algorithm. We reproduce their result below (with notation altered to suit our application) along with the short proof:

**Theorem 1** \([28]\). Suppose \( \text{BASE} \) and \( \text{TUNER} \) are both OCO algorithms. Let \( \{x_t^{\text{BASE}}\} \subset \mathbb{R}^d \) indicate the iterates of \( \text{BASE} \) in response to an arbitrary sequence of gradients \( \{g_t\} \), and let \( \{s_t\} \subset \mathbb{R} \) indicate the iterates of \( \text{TUNER} \) in response to the sequence of scalars \( \{\eta_t\} \). Define a new online algorithm MECHANIC via:

\[
x_t^{\text{MECHANIC}} = x_t^{\text{BASE}} + s_t \cdot (x_t^{\text{BASE}} - x_1^{\text{BASE}}).
\]

Then \( x_t^{\text{MECHANIC}} \) ensures regret:

\[
R^{\text{MECHANIC}}_{\text{linear}}(\bar{x}) \leq \inf_{\hat{s}} R^{\text{TUNER}}_{\text{linear}}(\hat{s}) + \hat{s} R^{\text{BASE}}_{\text{linear}}(\bar{x} - x_1^{\text{BASE}})/\hat{s}.
\]

**Proof.** By definition, for any \( \hat{s} \), we have:

\[
R^{\text{MECHANIC}}_{\text{linear}}(\bar{x}) = R^{\text{TUNER}}_{\text{linear}}(\hat{s}) + \hat{s} R^{\text{BASE}}_{\text{linear}}(\bar{x} - x_1^{\text{BASE}})/\hat{s}.
\]

With this result, the job of finding the optimal \( s \) can usually be completely delegated to TUNER. Although the value \( \hat{s} \) appears in both terms of the sum \( R^{\text{TUNER}}_{\text{linear}}(\hat{s}) + \hat{s} R^{\text{BASE}}_{\text{linear}}(\bar{x} - x_1^{\text{BASE}})/\hat{s} \), it turns out that for essentially all plausible \( \text{BASE} \) algorithms, there is a particular value \( \hat{s} \) that causes \( \hat{s} R^{\text{BASE}}_{\text{linear}}(x_1^{\text{BASE}} + (\bar{x} - x_1^{\text{BASE}})/\hat{s}) \) to obtain the optimal regret bound \([3]\). Thus, by setting \( \hat{s} \) to be this value, which is unknown \textit{a priori}, we need only ensure that \( R^{\text{TUNER}}_{\text{linear}}(\hat{s}) \) is small enough to not significantly affect the overall regret bound. Note that this setting of \( \hat{s} \) is done entirely in the analysis. For example, if \( \text{BASE} \) is actually SGD with a learning rate \( \eta \) and \( s = 1 \) as in \([2]\), we have

\[
R^{\text{MECHANIC}}_{\text{linear}}(\bar{x}) \leq R^{\text{MECHANIC}}_{\text{linear}}(\bar{x}) \leq \inf_{\hat{s}} R^{\text{TUNER}}_{\text{linear}}(\hat{s}) + O\left( \frac{\|\bar{x} - x_1\|^2}{\hat{s} \eta} + \frac{\sum_{t=1}^{T} \|g_t\|^2}{\hat{s} \eta} \right).
\]
setting \( \hat{s} = \frac{\|\hat{x} - x_1\|}{\eta \sqrt{\sum_{t=1}^{T} ||g_t||^2}} \):

\[
\leq R_{\text{linear}}^{\text{TUNER}} \left( \frac{\|\hat{x} - x_1\|}{\eta \sqrt{\sum_{t=1}^{T} ||g_t||^2}} \right) + O \left( \frac{\|\hat{x} - x_1\|}{\sqrt{\sum_{t=1}^{T} ||g_t||^2}} \right).
\]

Thus, if TUNER obtains low regret, then we will obtain the same regret bound as if we had optimally tuned the scaling factor for SGD. Intuitively, the gradient \( h_t \) provided to TUNER approximates the gradient over the entire course of the base optimizer rather than just at the most recent iterate. That is, for SGD, \( h_t \approx \frac{d f(x_t, x)}{d s} \) where \( x_t = x_1 - s \sum_{k=1}^{t-1} \eta_k g_k \).

### 2.2 Parameter-Free Online Optimization

The problem with the above result is that we seem to have simply pushed the problem off to TUNER: what if TUNER itself requires us to set a scale factor? Solving this problem has been the focus of substantial effort in the online optimization community \([29, 30, 10, 11, 28, 12]\). The most advanced such algorithms are able to ensure for all \( \hat{s} \) simultaneously:

\[
R_{\text{linear}}(\hat{s}) = \sum_{t=1}^{T} h_t(s_t - \hat{s}) \leq \tilde{O} \left( \|\hat{s}\| \sum_{t=1}^{T} h_t^2 \right).
\]

Thus, if we set \( h_t = \langle g_t, x_t^{\text{BASE}} - x_1^{\text{BASE}} \rangle \), we obtain:

\[
R_{\text{linear}}^{\text{TUNER}}(\hat{s}) \leq \tilde{O} \left( \|\hat{s}\| \sum_{t=1}^{T} (\langle g_t, x_t^{\text{BASE}} - x_1^{\text{BASE}} \rangle)^2 \right).
\]

In a theoretical development of this technique, it is necessary to prevent the terms \( \langle g_t, x_t^{\text{BASE}} - x_1^{\text{BASE}} \rangle^2 \) from becoming too large (as otherwise \( R_{\text{linear}}^{\text{TUNER}} \) is too large). Typically, this is accomplished by constraining the base algorithm to satisfy \( ||x_t^{\text{BASE}} - x_1^{\text{BASE}}|| \leq \rho \) for some user-specified arbitrary \( \rho \). Enforcing such a constraint means that the regret bound \( (4) \) would only apply to \( ||\hat{x}|| \leq \rho \), but ensures that \( ||g_t, x_t^{\text{BASE}} - x_1^{\text{BASE}}||^2 \leq \rho^2 ||g_t||^2 \). Thus, by setting \( \hat{s} = ||\hat{x} - x_1^{\text{BASE}}|| / \rho \), the combined algorithm obtains the optimal regret bound of \( O(\|\hat{x} - x_1^{\text{BASE}}|| \sqrt{\sum_{t=1}^{T} ||g_t||^2}) \) (amazingly, the value of \( \rho \) is irrelevant!). In practice however, we do not attempt to explicitly enforce any such constraints and simply rely on the intuition that any non-diverging algorithm is unlikely to produce excessively large iterates.

At no point in this process do we need access to the internal state of the base algorithm BASE. This means that improvements to BASE will automatically be reflected in improvements to the overall algorithm. In this paper, we investigate the performance of MECHANIC on deep learning tasks. We consider a variety of settings for the base algorithm BASE (i.e. AdamW, Lion, SGD, with various batch sizes and learning rate schedules of various shapes), and employ a parameter-free algorithm as the TUNER to automatically find the best scale factor for the base algorithm.

### 3 The MECHANIC algorithm

Our MECHANIC algorithm is specified in Algorithm [1]. The algorithm is built by applying Theorem 1 to a parameter-free TUNER algorithm presented in Algorithm 2, which is described along with theoretical analysis in Appendix D. However, when building MECHANIC, we modify the “pure” theoretically tractable Algorithm 2 to simplify the implementation while still capturing the essential intuition and maintaining the same performance. In the remainder of this section we will provide some intuition behind the TUNER update as used in MECHANIC as well as describing some potentially unfamiliar subtleties relating to our use of exponentially weighted moving averages.

MECHANIC takes as input a base algorithm that generates update vectors \( u_t \) as described in the previous sections. We then set \( \Delta_{t+1} = \sum_{k=1}^{t} u_k = x_t^{\text{BASE}} - x_1^{\text{BASE}} \). The majority of the algorithm contains our TUNER method, which is a variant of the analytically tractable Algorithm 2 with a
to prevent instabilities. In fact, the dependence of rate scaled by $\epsilon = 10^{-8}$: small value for numerical stability). To understand this update strategy, we can re-write the update as:

$$s_{t+1} = s_t \cdot \frac{\sqrt{v_{t-1}}}{\sqrt{v_t}} - \frac{s_t h_t}{\sqrt{v_t}} \approx \left(1 - \frac{h_t^2}{2v_t}\right) s_t - \frac{s_t h_t}{\sqrt{v_t}}.$$  

Thus, the update looks like a combination of an AdaGrad-esque gradient descent step with learning rate scaled by $s_t$ and a kind of “adaptive decay” (multiplication by $1 - \frac{h_t^2}{2v_t}$). The adaptive decay is very important for stabilizing the algorithm: without it the values for $s_t$ are prone to unchecked exponential growth due to scaling by $s_t$ in $\frac{s_t h_t}{\sqrt{v_t}}$. Intuitively, this decay is the minimum amount required to prevent instabilities.

In Appendix we provide a formal theorem bounding the regret of a variant of the procedure described above. Roughly speaking, for $\beta = 1$ this result suggests:

$$\sum_{t=1}^{T} h_t (s_t - \hat{s}) \leq O \left( (\hat{s} + \max_i s_i) \cdot m_T + \hat{s} \cdot \log(T \hat{s} / s_{init}) \right) \left( \sum_{t=1}^{T} h_t^2 \right). \quad (5)$$

In fact, the dependence of $O(\log(T))$ in equation (5) can be improved to $O(\sqrt{\log(T)})$ via more complicated algorithms (e.g. [28] [12] [31] [32]). However, we favor the simpler update and pleasing resemblance to familiar algorithms like AdaGrad via the Taylor expansion analysis above. Of note, the dependence on $s_{init}$ is very mild: this suggests that we should be able to set $s_{init}$ to a very small
value without damaging performance. In practice, we choose $s_{\text{init}} = 10^{-8}$, which we expect to dramatically under-estimate the optimal value in all scenarios.

We hypothesize that the simplified TUNER we use in MECHANIC in fact possesses a rigorous theoretical analysis (although perhaps only with respect to simpler non-fully-worst-case adversaries), but demonstrating such a bound appears to involve difficult technical hurdles. In particular, our implementation is designed to be “scale-free”: rescaling the values of $g_t$ by any constant scalar will have no effect on $s_t$. This property was first achieved only recently in theoretical analysis of parameter-free algorithms [12], and as-yet requires significantly more involved algorithms [12][33].

3.1 The use of $\beta$

We include $\beta$ to introduce some recency bias in the statistics recorded by MECHANIC, a common feature of practical optimizers. Mathematically, we accomplish this by up-weighting the $t$th feedback to TUNER by $\beta^{-t}$: $h_t \to h_t \beta^{-t}$. Thus, for example, we have $v_t = \sum_{k=1}^{T} h_t \beta^{-2kt}$ and $r_t = -\sum_{k=1}^{T} h_k s_{k-1} \beta^{-k}$.

Using these weights directly results in numerical stability issues as the weights become exponentially large. Instead, since we only need to maintain the correct ratio we can cancel a factor of $\beta^{-t}$ from both sides, giving the update equations in Algorithm 2.

We found that tuning the value of $\beta$ can significantly improve performance on different tasks. Thus, we incorporate multiple $\beta$ values simultaneously in a way that obviates the need for such tuning.

Our approach is inspired by work on “combining” parameter free algorithms [34]. The idea is simple: parameter-free algorithms typically ensure that $R_{\text{linear}}(0) \leq \epsilon$ for some constant $\epsilon$ set by the user. So, if $s_{t,1}, \ldots, s_{t,n}$ are the outputs of $n$ parameter-free algorithms with regret bounds $R_{\text{linear}}^1(\hat{s}), \ldots, R_{\text{linear}}^n(\hat{s})$, we have for any $j$:

$$
\sum_{t=1}^{T} h_t \left( \sum_{i=1}^{n} s_{t,i} - \hat{s} \right) = \sum_{t=1}^{T} h_t (s_{t,j} - \hat{s}) + \sum_{i \neq j} \sum_{t=1}^{T} h_t (s_{t,i} - 0) = R_{\text{linear}}^j(\hat{s}) + \sum_{i \neq j} R_{\text{linear}}^i(0) \leq R_{\text{linear}}^j(\hat{s}) + (n - 1) \epsilon.
$$

So, with small constant additive overhead in the regret, the sum of all the outputs $s_{t,1} + \cdots + s_{t,n}$ achieves the same regret as the best of all the outputs. Motivated by this observation, we instantiate $n = 6$ copies of TUNER with different $\beta$ values and add their iterates to produce a final scaling.

3.2 Weight decay

Finally, we found that an addition of a peculiar weight-decay-esque term helped significantly on certain tasks, including vision tasks with smaller datasets, multi-objective NLP tasks and especially with reducing the variance in final results for all tasks. Specifically, rather than providing $h_t = \langle g_t, \Delta_t \rangle$ as the input to the TUNER algorithm, we instead provide $h_t = \langle g_t + \lambda \| g_t \| \sum_{i=1}^{n} s_{t,i} \| x_t \|, \Delta_t \rangle$. We conjecture that this term is helpful in the common case that the base algorithm itself is incorporating regularization or weight-decay.

This extra term is the derivative of the regularizer $x \mapsto \lambda \| g_t \| \sum_{i=1}^{n} s_{t,i} \| x \|$. From a standard theoretical perspective, this regularization may seem overly large. However, it may not have as big an impact as one might imagine because the base algorithm does not see this regularization. Instead, the base algorithm may (or may not) perform weight decay using another method that MECHANIC has no insight into. That said, we do not propose an analytical explanation for this modification. We simply observed that in practice it performed well with a fixed $\lambda = 0.01$.

3.3 Runtime and Memory Cost

MECHANIC incurs little additional cost over that of BASE. In Algorithm 1 we denote $d$-dimensional vectors with bold font, and $n$-dimensional vectors and scalars with normal font (note that typically $n = 6$). We use 1 additional $O(d)$ memory slot, and four $O(d)$-time steps in lines 8, 9, 10 and 17. All other steps are $O(1)$ or $O(n)$ time and so have negligible cost.
<table>
<thead>
<tr>
<th>Model</th>
<th>Size</th>
<th>Pre Opt</th>
<th>MLM</th>
<th>Optimizer</th>
<th>MNLI-m/mm</th>
<th>QNLI</th>
<th>SST-2</th>
<th>QQP</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-B</td>
<td>110M</td>
<td>AdamW</td>
<td>71.5</td>
<td>AdamW</td>
<td>84.3/84.8</td>
<td>91.0</td>
<td>92.4</td>
<td>90.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M-AdamW</td>
<td>83.7/83.5</td>
<td>90.6</td>
<td>91.9</td>
<td>90.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M-AdamW</td>
<td>84.7/85.1</td>
<td>91.2</td>
<td>93.3</td>
<td>90.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lion</td>
<td>71.7</td>
<td>Lion</td>
<td>83.4/83.5</td>
<td>86.8</td>
<td>89.7</td>
<td>89.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M-Lion</td>
<td>83.1/83.8</td>
<td>89.9</td>
<td>91.0</td>
<td>90.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M-Lion</td>
<td>84.5/84.2</td>
<td>89.0</td>
<td>91.2</td>
<td>90.8</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-L</td>
<td>340M</td>
<td>AdamW</td>
<td>75.4</td>
<td>AdamW</td>
<td>86.2/86.4</td>
<td>92.2</td>
<td>93.9</td>
<td>91.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M-AdamW</td>
<td>86.1/86.4</td>
<td>92.5</td>
<td>93.7</td>
<td>91.5</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M-AdamW</td>
<td>86.3/86.5</td>
<td>92.7</td>
<td>94.4</td>
<td>91.4</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Lion</td>
<td>75.3</td>
<td>Lion</td>
<td>86.7/86.6</td>
<td>90.7</td>
<td>92.9</td>
<td>91.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>M-Lion</td>
<td>86.0/86.2</td>
<td>90.3</td>
<td>93.4</td>
<td>91.2</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M-Lion</td>
<td>87.4/87.4</td>
<td>92.9</td>
<td>93.3</td>
<td>91.7</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>M-Lion</td>
<td>87.2/87.1</td>
<td>91.5</td>
<td>92.3</td>
<td>91.6</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1: Comparing MECHANIC on BERT. 5 largest datasets from GLUE. Results reported are peak validation scores averaged over 3 runs, both for the baseline and MECHANIC tuned models.

4 Experiments

In this section we describe our experiments using MECHANIC to tune various base optimizers on different tasks. Note that almost all base optimizer implementations require a user-specified scale factor which is not directly visible to MECHANIC. We set this value to 1.0 before applying MECHANIC. Since MECHANIC multiplies the base update by $s$, setting the base scale factor to 1.0 allows us to interpret $s$ as the “correct” value for the base scale.

4.1 Masked Language Modeling

We perform BERT pre-training on the Wikibooks dataset following the procedure from [35] with a few minor changes, most notably, we omit the Next Sentence Prediction (NSP) loss following [36]. Masked language modeling (MLM) requires reconstructing randomly masked tokens given an input sequence of tokens. As shown in Table 1, using MECHANIC leads to a noticeable improvement in MLM accuracy.

Varying batch size and model size: Past works observe that the scale factor $s$ should decrease as either batch size is decreased or model size is increased [37, 38]. To inspect the scale factor that MECHANIC learns, we vary the batch size and model size while pre-training BERT using MECHANIC. As shown in Figure 1, in both cases, MECHANIC learns to decrease the scale factor $s$ when decreasing the batch size and when increasing the model size.

Addition Ablations: Ablation studies on the effects of $n$, $\lambda$, $s_{init}$ can be found in Appendix B.

Finetuning pre-trained models: In addition to pre-training, we evaluate our models on the 5 largest datasets from the GLUE suite [39]. One possible failure mode of MECHANIC tuned pre-trained models could have been that, even though they lead to high accuracy at pre-training time, transfer learning may fail at finetuning time.

To ensure that standard transfer learning pipelines still work with MECHANIC pre-trained models, we finetune them without a learning rate tuner using the AdamW optimizer and find that MECHANIC pre-trained models lead to higher accuracy at pre-training time, and they also outperform in finetuning more often than not. We finetune BERT-B (110M) and BERT-L (340M) models for at most 10 epochs on each of the GLUE datasets and report results on the GLUE dev set in Table 1.

Using MECHANIC for finetuning: We also investigated using MECHANIC for finetuning. Typically, to not erase the progress already made, a much lower base learning rate is employed at finetuning.
time. This could easily have been a potential failure mode of any kind of automatic learning rate tuner as such strategies might “explore” a high learning rate at the beginning of the optimization procedure. Fortunately, we observed that this inductive bias typically baked at finetuning time is still maintained when using MECHANIC.

4.2 Image Classification

In this Section, we present results on popular Image Classification tasks. Apart from training from scratch, we also perform transfer learning experiments where we pre-train on the JFT-300M \cite{10} dataset and finetune on ImageNet, Cifar-10 and Cifar-100 datasets. We follow the exact setting employed in \cite{5} for both pre-training and finetuning.

As shown in Table 2, MECHANIC is quite competitive across the board and produces results either very close to the baseline or better. Since MECHANIC optimizes for the train loss, in general, we observe that it results in better test performance on tasks with large amounts of data where the model is unable to overfit to the train set. For instance, we see that MECHANIC beats the baseline substantially when pre-training ViT models on JFT-300M, whereas it lags slightly behind on smaller datasets like ImageNet-1k or CIFAR-10/100. Even though we fix $\lambda$ to 0.01 as default for all our reported experiments, we find that for small datasets like CIFAR-10, increasing it led to better test performance.

4.3 Comparison with D-adaptation

Recently, \cite{16} introduced the D-adaptation algorithm, with the same goal of learning the correct scale $s$ for SGD and Adam base optimizers. D-adaptation showed impressive empirical results on a range of popular deep learning tasks, so we compare MECHANIC with D-adaptation on a selection of tasks that D-adaptation worked well on, using code provided by the authors. Hyper-parameter settings were kept the same to ensure a fair comparison. In contrast to D-adaptation, MECHANIC does not require modification for different base optimizers and, as shown in Figure 2, it remains quite
Table 2: Comparing MECHANIC on vision models. All fine-tuning results are averaged over 3 independent runs with different seeds.

competitive on small datasets like CIFAR-10/100 while outperforming both a manually tuned baseline and D-adaptation on bigger tasks like IWSLT14 and language modeling on BookWiki dataset. We present additional results in Appendix C.3 including a comparison on a suite of 12 logistic regression problems.

5 Conclusion

MECHANIC is a new technique for scaling the updates of any base optimization algorithm. Our approach provides a practical implementation of recent developments in optimization theory, and is able to match the performance of tuned baselines on large-scale machine learning tasks. This work suggests several natural future directions. First, is there a theoretical motivation for our weight-decay term? Next, is it possible to leverage similar techniques to learn a per-layer scale factor? Such a capacity would not significantly increase computation cost, but by allowing more degrees of freedom may yield a method that significantly outperforms baselines since it is infeasible to manually tune a scale factor for every layer.

Acknowledgments

Ashok Cutkosky acknowledges funding support from NSF grant CCF-2211718, an Amazon research award, and Google.

References

Figure 2: Comparing MECHANIC with D-adaptation and Adam or SGD with manually tuned learning rates on vision and language tasks.


[43] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Yao Liu, Hieu Pham, Xuanryi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V. Le. Symbolic discovery of optimization algorithms, 2023.


A Limitations

In this paper, we introduce a technique for automatically learning the right scale of the learning rate called MECHANIC and evaluate it on a broad range of practical deep learning problems and settings. We find that, depending on the problem, MECHANIC can be quite effective and even surpass performance of manual tuning of learning rates at a fraction of the cost. We also find that in addition to training from scratch, MECHANIC also works for finetuning.

While the initial set of results are encouraging, many challenges remain. Firstly, we found that MECHANIC does not seem to work well with dropout [42]. While MECHANIC is effective against noise from sampling, we believe there may be a more fundamental reason why dropout does not work well with MECHANIC. Second, MECHANIC re-purposes the gradient coming from the train set for learning the learning rate, which means it optimizes for train loss. This is different from manual tuning of learning rates where researchers tune it based on performance on the validation set. A principled way to handle this discrepancy is also an interesting avenue of future research.

B Additional Ablations

B.1 Setting BASE learning rate with Mechanic

Even though we recommend setting peak learning rate of Base optimizer to 1.0, to make sure that Mechanic truly is insensitive to the peak learning rate and robust to the choice of this hyperparameter, we conduct a study where we vary peak learning rate of AdamW on BERT-B masked language modeling task. As shown in Table 3, Mechanic is largely robust against choice of peak LR set for AdamW on BERT-B MLM task.

<table>
<thead>
<tr>
<th>Peak LR of BASE with Mechanic</th>
<th>MLM Acc</th>
</tr>
</thead>
<tbody>
<tr>
<td>1e-2</td>
<td>71.7</td>
</tr>
<tr>
<td>1e-1</td>
<td>71.6</td>
</tr>
<tr>
<td>1e0</td>
<td>71.6</td>
</tr>
<tr>
<td>1e1</td>
<td>71.4</td>
</tr>
<tr>
<td>1e2</td>
<td>71.6</td>
</tr>
</tbody>
</table>

Table 3: MECHANIC is largely robust against choice of peak LR set for AdamW on BERT-B MLM task.

B.2 Robustness to $s_{init}$

We have observed that while $\lambda$ is helpful in stabilizing MECHANIC on some problems, as long as it is set to a reasonable small value it does not affect performance by a lot.
B.4 Number of $\beta$ values

<table>
<thead>
<tr>
<th>Number of $\beta$ values ($n$)</th>
<th>2</th>
<th>4</th>
<th>6</th>
<th>8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td>48.9</td>
<td>49.5</td>
<td>49.9</td>
<td>49.6</td>
</tr>
</tbody>
</table>

Table 6: Accuracy on JFT-300M using model ViT-B/16 and optimizer $\mathcal{M}$-AdamW as a function $n$, the number of $\beta$ values. The $\beta$ values are set as $1 - 0.1^i$ for $i$ from 1 to $n$. This is the most sensitive parameter in MECHANIC. For smaller $n$ we see some significant performance degradation, while for larger $n$ we see milder degradation. Theory suggests that larger $n$ should result in a degradation that is roughly logarithmic in $n$.

C Additional Experimental Details

C.1 Hyperparams for BERT

<table>
<thead>
<tr>
<th>Model</th>
<th>Aug</th>
<th>Optimizer</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>lr sweep</th>
<th>best lr</th>
<th>Weight decay</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-B</td>
<td>AdamW</td>
<td>0.9</td>
<td>0.999</td>
<td></td>
<td>[5e-4, 1e-3, 2e-3, 5e-3, 1e-2]</td>
<td>5e-3</td>
<td></td>
</tr>
<tr>
<td>BERT-L</td>
<td>AdamW</td>
<td>0.9</td>
<td>0.999</td>
<td></td>
<td>[5e-4, 1e-3, 2e-3, 5e-3, 1e-2]</td>
<td>1e-3</td>
<td>0.1</td>
</tr>
<tr>
<td>BERT-B/L</td>
<td>$\mathcal{M}$-AdamW</td>
<td>0.9</td>
<td>0.999</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BERT-B</td>
<td>Lion</td>
<td>0.9</td>
<td>0.999</td>
<td></td>
<td>[5e-5, 1e-4, 2e-4, 5e-4, 1e-3]</td>
<td>5e-4</td>
<td></td>
</tr>
<tr>
<td>BERT-L</td>
<td>Lion</td>
<td>0.9</td>
<td>0.999</td>
<td></td>
<td>[5e-5, 1e-4, 2e-4, 5e-4, 1e-3]</td>
<td>2e-4</td>
<td></td>
</tr>
<tr>
<td>BERT-B/L</td>
<td>$\mathcal{M}$-Lion</td>
<td>0.9</td>
<td>0.999</td>
<td></td>
<td></td>
<td></td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 7: Critical hyperparameters we used for BERT pre-training. For the baselines we grid searched the learning rate as shown in the table. Batch size 2k, trained for 150k steps on original Wikibooks dataset w/o NSP loss (similar to Roberta). We found that a small amount of weight decay makes MECHANIC slightly more effective.

<table>
<thead>
<tr>
<th>Model</th>
<th>AdamW LR</th>
<th>MECHANIC-AdamW</th>
</tr>
</thead>
<tbody>
<tr>
<td>BERT-B</td>
<td>71.1</td>
<td>71.3</td>
</tr>
<tr>
<td>BERT-L</td>
<td>75.0</td>
<td>74.6</td>
</tr>
<tr>
<td>BERT-B</td>
<td>70.8</td>
<td>71.8</td>
</tr>
<tr>
<td>BERT-L</td>
<td>75.1</td>
<td>74.7</td>
</tr>
</tbody>
</table>

Table 8: As detailed in Table 7, we performed a grid search of learning rate values for each base optimizer. Here, we present the resulting accuracy values.
Table 9: BERT GLUE finetuning hparams with AdamW.

<table>
<thead>
<tr>
<th>Hyperparam</th>
<th>Optimizer</th>
<th>Without MECHANIC pre-training</th>
<th>With MECHANIC pre-training</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning Rate</td>
<td>Adam</td>
<td>[5e-5, 1e-4, 2e-4, 3e-4, 5e-4]</td>
<td>[1e-5, 2e-5, 3e-5, 5e-5, 1e-4]</td>
</tr>
<tr>
<td>Learning Rate</td>
<td>Lion</td>
<td>[5e-6, 1e-5, 2e-5, 3e-5, 5e-5]</td>
<td>[1e-6, 2e-6, 3e-6, 5e-6, 1e-5]</td>
</tr>
<tr>
<td>Batch Size</td>
<td>[16, 32]</td>
<td>16</td>
<td>16</td>
</tr>
<tr>
<td>Weight Decay</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Max Epochs</td>
<td>10</td>
<td>10</td>
<td>10</td>
</tr>
<tr>
<td>Learning Rate Decay</td>
<td>Linear</td>
<td>Linear</td>
<td>Linear</td>
</tr>
<tr>
<td>Warmup Ratio</td>
<td>0.06</td>
<td>0.06</td>
<td>0.06</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
<tr>
<td>Attention Dropout</td>
<td>0.1</td>
<td>0.1</td>
<td>0.1</td>
</tr>
</tbody>
</table>

Table 10: BERT GLUE finetuning hparams when using mechanic at finetuning time. We found a limitations of MECHANIC that it does not perform well in combination to dropout, so we set dropout rate to 0 for these experiments.

### C.2 Hyperparams for Image Classification

<table>
<thead>
<tr>
<th>Model</th>
<th>Aug</th>
<th>Optimizer</th>
<th>$\beta_1$</th>
<th>$\beta_2$</th>
<th>lr</th>
<th>Weight decay</th>
<th>Num epochs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B/16</td>
<td>AdamW</td>
<td>0.9</td>
<td>0.999</td>
<td>8e-4</td>
<td>0.1</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>ViT-B/16</td>
<td>$\mathcal{M}$-AdamW</td>
<td>0.9</td>
<td>0.999</td>
<td>0.1</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViT-L/16</td>
<td>AdamW</td>
<td>0.9</td>
<td>0.999</td>
<td>4e-4</td>
<td>0.1</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>ViT-L/16</td>
<td>$\mathcal{M}$-AdamW</td>
<td>0.9</td>
<td>0.999</td>
<td>0.1</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViT-B/16</td>
<td>Lion</td>
<td>0.9</td>
<td>0.99</td>
<td>1e-4</td>
<td>0.3</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>ViT-B/16</td>
<td>$\mathcal{M}$-AdamW</td>
<td>0.9</td>
<td>0.99</td>
<td>0.3</td>
<td>7</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViT-L/16</td>
<td>Lion</td>
<td>0.9</td>
<td>0.99</td>
<td>1e-4</td>
<td>0.3</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>ViT-L/16</td>
<td>$\mathcal{M}$-AdamW</td>
<td>0.9</td>
<td>0.99</td>
<td>0.3</td>
<td>7</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 11: Critical hyperparameters we used for all the experiments, most of them directly repurposed from [41]. For each baseline we repurposed a well-tuned base learning from previous work [41, 43]. Trained on JFT-300M with batch size 4k with LR cosine decay schedule.
Figure 3: Scaling values $s$ learned by MECHANIC during ResNet18 training on CIFAR10 and WideResNet training on CIFAR100. Shaded area represents max/min value over 3 runs. Dark line is average.

<table>
<thead>
<tr>
<th>Hyperparam</th>
<th>ImageNet</th>
<th>Cifar100</th>
<th>Cifar10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Learning rate sweep</td>
<td>{0.003, 0.01, 0.03, 0.06}</td>
<td>{0.001, 0.003, 0.01, 0.03}</td>
<td>{0.001, 0.003, 0.01, 0.03}</td>
</tr>
<tr>
<td>Batch size</td>
<td>512</td>
<td>512</td>
<td>512</td>
</tr>
<tr>
<td>Weight decay</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Num steps</td>
<td>20k</td>
<td>10k</td>
<td>10k</td>
</tr>
<tr>
<td>Warmup steps</td>
<td>500</td>
<td>500</td>
<td>500</td>
</tr>
<tr>
<td>Learning rate decay</td>
<td>Cosine</td>
<td>Cosine</td>
<td>Cosine</td>
</tr>
<tr>
<td>Dropout</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Clipping norm</td>
<td>1.0</td>
<td>1.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

Table 12: We directly use ViT finetuning hyperparams recommended by [41]. For MECHANIC we also use same hyperparameters, omitting just the learning rate sweep, since we don’t need it now. We use finetuning resolution of 384.

<table>
<thead>
<tr>
<th>Hyperparam</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>weight decay</td>
<td>[0.001, 0.0005, 0.0001]</td>
</tr>
<tr>
<td>lr</td>
<td>[0.3, 0.1, 0.03]</td>
</tr>
<tr>
<td>SGD Momentum $\beta$</td>
<td>0.9</td>
</tr>
<tr>
<td>batch size</td>
<td>128</td>
</tr>
<tr>
<td>num epochs</td>
<td>200</td>
</tr>
<tr>
<td>schedule</td>
<td>Step decay by 0.2 at 60, 120, 160 epochs</td>
</tr>
<tr>
<td>augmentations</td>
<td>Random Crop, Random Horizontal Flip</td>
</tr>
<tr>
<td>Gradient clip by global norm $\lambda$</td>
<td>1.0</td>
</tr>
<tr>
<td></td>
<td>Kept at default 0.01</td>
</tr>
</tbody>
</table>

Table 13: Hyperparameters for tuning ResNet18 on CIFAR10 and WideResNet on CIFAR100

C.3 Hyperparams and additional results on comparisons with D-adaptation

D Theoretical Analysis

Here we provide the theoretically tractable version of TUNER as well as its analysis.

D.1 Algorithmic Simplifications

To simplify the implementation of MECHANIC, we replaced all of the red text in Algorithm 2 with a single line $r_t \leftarrow \max(0, r_t)$ right after the definition of $r_t$. This is motivated by two ideas.
Figure 4: Comparing MECHANIC with D-adaptation and manually tuned learning rates on a suite of convex tasks.

Figure 5: Complementary train set results of MECHANIC with D-adaptation and manually tuned learning rates on vision and language tasks.
Algorithm 2 TUNER (theoretically tractable version)

1: **Input:** $\beta \in [0, 1]$, $W_0$: initial “wealth”.
2: $v_0 \leftarrow 0 \in \mathbb{R}$, $r_0 \leftarrow 0 \in \mathbb{R}$, $q_t \leftarrow 0 \in \mathbb{R}$, $m_0 \leftarrow 0 \in \mathbb{R}$, $s_1 \leftarrow 0 \in \mathbb{R}$
3: for $t = 1 \ldots T$ do
4: Output $s_t$.
5: Receive $h_t$.
6: $m_t \leftarrow \max(\beta m_t-1, h_t)$
7: $\hat{h}_t = \text{clip}(h_t, -m_t-1, m_t-1)$ // Red text are steps that we omit in practice (see Algorithm 1)
8: $q_t \leftarrow \beta q_{t-1} + \hat{h}_t$
9: $r_t \leftarrow \beta r_{t-1} - s_{t-1} \hat{h}_t$
10: $W_t \leftarrow W_0 + r_t$ // Blue text is changed in practice (see Algorithm 1)
11: $v_t \leftarrow \beta^2 v_{t-1} + \hat{h}_t^2$
12: $s_{t+1} \leftarrow \frac{W_t}{\sqrt{4m_t^2 + v_t + \epsilon}} \cdot \text{clip}(q_t / \sqrt{4m_t^2}, v_t, 0, 1)$
13: end for

First, we conjecture that the clip operation using $q_t / \sqrt{v_t}$ may even be unnecessary in theory\footnote{Removing the clip in theory may requiring some additional non-worst-case assumption.}, we observed no change from removing this operation in practice, and observe that the update has an intuitive interpretation via the Taylor expansion discussed in Section 3.

Second, the clip operation on $h_t$ using $m_t-1$ is essentially designed to prevent the wealth $W_t$ from becoming negative or zero using the gradient truncation technique employed by \[44, 45, 12\]. While less well consistent with known theory, we found it simpler to ensure the wealth $W_t$ does become negative simply by clipping $r_t$ directly (we did not clip $W_t$ to be nonnegative as $W_t = 0$ would cause the algorithm to output $s_t = 0$ for all future iterations). We found these changes simplified the algorithm while having no noticeable effect on the performance. Although these deviations technically do not come with guarantees, the accomplish similar intuitive goals and so we expected (and observed) that they simplified the implementation while not damaging performance.

D.2 Eliminating $W_0$ in favor of $s_{\text{init}}$

While TUNER makes use of the “initial wealth” value $W_0$, MECHANIC instead adopts a varying value for $W_0$ proportional to $s_{\text{init}} \cdot m_t$. This makes the first $s$ value proposed by MECHANIC equal to $s_{\text{init}}$, which is more intuitive to set than $W_0$. The exponential growth in $s$ allows us to set $s_{\text{init}}$ to a very small value of $10^{-8}$. It also makes the values for $s$ “scale-free” in the sense that rescaling the updates $u_t$ by any constant will have no effect on the resulting $s_t$.

D.3 Regret Bound

**Theorem 2.** With $\beta = 1$, Algorithm 2 guarantees for all $\hat{s} \geq 0$:

$$\sum_{t=1}^{T} h_t(s_t - \hat{s}) \leq W_0 + (\hat{s} + \max_t s_t) \cdot m_T + O \left( \hat{s} \cdot \log(T \bar{s} m_T / m_1 W_0) \sum_{t=1}^{T} h_t^2 \right).$$

**Proof.** First, we employ an argument developed by \[44\]:

$$\sum_{t=1}^{T} h_t(s_t - \hat{s}) \leq \sum_{t=1}^{T} \hat{h}_t(s_t - \hat{s}) + \sum_{t=1}^{T} |\hat{h}_t - h_t| (|s_t| + |\hat{s}|)$$

$$\leq \sum_{t=1}^{T} \hat{h}_t(s_t - \hat{s}) + (\max_t |s_t| + |\hat{s}|) \sum_{t=1}^{T} |\hat{h}_t - h_t|$$

$$= \sum_{t=1}^{T} \hat{h}_t(s_t - \hat{s}) + m_T \max_t |s_t| + |\hat{s}|.$$
So, in the following, it suffices to bound $\sum_{t=1}^{T} \hat{h}_t (s_t - \hat{s})$. This is helpful because we will be able to use the bound $|\hat{h}_t| \leq m_{t-1}$, and $m_{t-1}$ is known before $\hat{h}_t$ is revealed.

As is typical in the analysis of parameter-free algorithms, the proof proceeds by lower-bounding the wealth. Define a function $a(x)$ piecewise by:

$$a(x) = \begin{cases} 
0 & x \leq 0 \\
x^{2}/2 & x \in [0, 1] \\
x - 1/2 & x \geq 1
\end{cases}$$

Notice that $a(x)$ is differentiable, monotonically increasing and 1-Lipschitz. We are going to roughly show that $W_t \geq \Omega(\exp(a(-\sum_{k=1}^{t} \hat{h}_k / \sqrt{v_t})))$, after which the regret bound will follow from the wealth-regret duality [10].

The key technical inequality in the proof is the following: for any $A, B, m$ with $B \geq 4m^2$, and any $|x| \leq m$, we have:

$$a \left( \frac{-A}{\sqrt{B}} \right) - \frac{x}{\sqrt{B}} \left[ \frac{x}{x^2 + x^2} \right] \geq a \left( \frac{-A - x}{\sqrt{B + x^2}} \right) - \frac{x^2}{B + x^2}.$$  \hspace{1cm} (6)

Once (6) is established, we proceed as follows: defining $c_t = \frac{\exp(\sum_{k=1}^{t-1} \hat{h}_k / \sqrt{4m_{t-1}^2 + v_{t-1}})}{\sqrt{v_t}}$, we have:

$$\log(W_t) = \log(W_{t-1}) + \log(1 - \hat{h}_t c_t) \geq \log(W_{t-1}) - \hat{h}_t c_t - (2\hat{h}_t^2),$$

where we have used $c_t \leq 1/2$ and the identity $\log(1 - x) \geq -x - x^2$ for $x \leq 1/2$ (which applies since $\hat{h}_t \leq m_{t-1}$ by definition). Now, set $A = \sum_{k=1}^{t-1} \hat{h}_k$ and $B = 4m_{t-1}^2 + v_{t-1}$ and $x = \hat{h}_t$ in (6), we see that:

$$\log(W_t) - \log(W_{t-1}) \geq -\frac{x}{\sqrt{B}} \left[ \frac{\exp(\sum_{k=1}^{t-1} \hat{h}_k / \sqrt{4m_{t-1}^2 + v_{t-1}})}{\sqrt{v_t}} \right] - \frac{\hat{h}_t^2}{4m_{t-1}^2 + v_{t-1}} \\
\geq a \left( \frac{-A - x}{\sqrt{B + x^2}} \right) - a \left( \frac{-x^2}{B + x^2} \right) - \frac{\hat{h}_t^2}{4m_{t-1}^2 + v_{t-1}} \\
\geq a \left( \frac{-A - x}{\sqrt{B + x^2}} \right) - a \left( \frac{-x^2}{B + x^2} \right) - \frac{2\hat{h}_t^2}{4m_{t-1}^2 + v_{t-1}} \\
\geq a \left( \frac{-A - x}{\sqrt{B + x^2}} \right) - a \left( \frac{-x^2}{B + x^2} \right) - \frac{2\hat{h}_t^2}{4m_{t-1}^2 + v_{t-1}}.$$ 

Thus by telescoping the sum, we have:

$$\log(W_T) \geq \log(W_0) + a \left( \frac{-\sum_{k=1}^{T} \hat{h}_k}{\sqrt{v_T}} \right) - \sum_{t=1}^{T} \frac{2\hat{h}_t^2}{4m_{t-1}^2 + v_{t-1}}.$$ 

Now, observe that $\frac{2\hat{h}_t^2}{4m_{t-1}^2 + v_{t-1}} \leq \frac{2\hat{h}_t^2}{v_t} \leq 2(\log(v_t) - \log(v_{t-1}))$, so we have $\sum_{t=1}^{T} \frac{\hat{h}_t^2}{v_{t-1}} \leq 2\log(T m_T / m_1)$ so that overall:

$$W_T \geq W_0 m_1 \exp \left[ a \left( \frac{-\sum_{k=1}^{T} \hat{h}_k}{\sqrt{v_T}} \right) \right].$$
Thus, if we define $p(H) = \frac{Wm}{Tm} \exp \left[ a \left( \frac{H}{\sqrt{v_T}} \right) \right]$, we have $W_T \geq p(\sum_{k=1}^{T} \hat{h}_k)$. Now, we employ the reward-regret duality:

$$\sum_{t=1}^{T} \hat{h}_t (s_t - \hat{s}) = s_{\text{init}} \cdot m + \sum_{t=1}^{T} (-\hat{h}_t) - W_T \leq W_0 + \sup_G \hat{s} \cdot G - p(G) = W_0 + p^* (\hat{s}) \leq W_0 + O(s \log(sT/s_{\text{init}})/\sqrt{v_T}).$$

Where $p^*$ is the Fenchel conjugate of $p$ and the evaluation of the conjugate follows from direct calculation (see, e.g. [10, 28, 46]).

Thus, to prove the theorem we need only show (6). This is established via casework in a manner similar to [46].

**Case 1.** $\frac{-A \cdot x}{\sqrt{B + x^2}} \leq 0$: In this case, the statement is equivalent to: $\frac{x^2}{B} \geq a \left( \frac{-A - x}{\sqrt{B + x^2}} \right)$. Note that since $\frac{-A \cdot x}{\sqrt{B + x^2}} \leq 0$, we have $A \geq 0$. Therefore:

$$\frac{-A - x}{\sqrt{B + x^2}} = \frac{-A}{\sqrt{B + x^2}} - \frac{x}{\sqrt{B + x^2}} \leq -\frac{x}{\sqrt{B + x^2}}.
$$

Further, we clearly have $-\frac{x}{\sqrt{B + x^2}} \leq 1$ so that:

$$a \left( \frac{-A - x}{\sqrt{B + x^2}} \right) \leq a \left( -\frac{x}{\sqrt{B + x^2}} \right) = \frac{x^2}{2(B + x^2)} \leq \frac{x^2}{B}.
$$

So, in the following we assume $\frac{-A \cdot x}{\sqrt{B + x^2}} \geq 0$.

**Case 2.** $\frac{-A \cdot x}{\sqrt{B + x^2}} \leq 0$: In this case, it suffices to show $\frac{x}{\sqrt{B + x^2}} \cdot \text{clip} \left( \frac{-A \cdot x}{\sqrt{B + x^2}}, 0, 1 \right) \geq -\frac{x^2}{B}$. The case assumption implies $m^2 \geq x \geq -A \geq 0$. Therefore, since $B \geq 4m^2$, clip $\left( \frac{-A \cdot x}{\sqrt{B + x^2}}, 0, 1 \right) = \frac{-A}{\sqrt{B}}$ so that

$$\frac{x}{\sqrt{B + x^2}} \geq \frac{x}{\sqrt{B}} \geq \frac{x}{B} \text{ as desired.}
$$

So, in the following we now further assume $\frac{-A \cdot x}{\sqrt{B + x^2}} \geq 0$.

**Case 3.** $\frac{-A \cdot x}{\sqrt{B + x^2}} \in [0, 1]$: We have $a \left( \frac{-A \cdot x}{\sqrt{B + x^2}} \right) = \frac{A^2}{2B}$, and also since $a(z) \leq \frac{1}{2} z^2$ for all $z$, $a \left( \frac{-A \cdot x}{\sqrt{B + x^2}} \right) \leq \frac{(A+x)^2}{2(B+x^2)} - \frac{x^2}{B}$, but this is equivalent to $\frac{(A+x)^2}{2(B+x^2)} - \frac{x^2}{B}$, which clearly holds.

**Case 4.** $\frac{-A \cdot x}{\sqrt{B + x^2}} \geq 1$ and $\frac{-A \cdot x}{\sqrt{B + x^2}} \geq 1$: In this case it suffices to show $\frac{-A}{\sqrt{B}} - \frac{x}{\sqrt{B + x^2}} \geq \frac{-A \cdot x}{\sqrt{B + x^2}} - \frac{x^2}{B}$. To see this, we have:

$$\frac{-A - x}{\sqrt{B + x^2}} = \frac{-A}{\sqrt{B + x^2}} - \frac{x}{\sqrt{B + x^2}} \leq \frac{-A}{\sqrt{B}} - \frac{x}{\sqrt{B}} + x \left( \frac{1}{\sqrt{B}} - \frac{1}{\sqrt{B + x^2}} \right) \leq \frac{-A}{\sqrt{B}} - \frac{x}{\sqrt{B}} + \frac{2x^3}{3B^{3/2}} \leq \frac{-A}{\sqrt{B}} - \frac{x}{\sqrt{B}} + \frac{x}{B},$$
where in the second-to-last line we have used the fact that $h \mapsto \frac{1}{\sqrt{B} + h}$ is a convex in $h$, and in the last line we have used $\sqrt{B} \geq m \geq x$.

**Case 5:** $\frac{-A}{\sqrt{B}} \geq 1$ and $\frac{-A+x}{\sqrt{B+x^2}} \in [0, 1)$: In this case we need to show $\frac{-A}{\sqrt{B}} - \frac{1}{2} \geq \frac{(A+x)^2}{2(B+x^2)} - \frac{x^2}{B} + \frac{x}{\sqrt{B}}$.

To see this, we first observe that since $\frac{-A+x}{\sqrt{B+x^2}} \in [0, 1)$, we have

\[
A^2 + 2Ax + x^2 \leq B + x^2
\]

Thus, by quadratic formula, $A \geq -x - \sqrt{x^2 + B}$, so that we have $A \in [-x - \sqrt{x^2 + B}, -\sqrt{B}]$.

Next, our target identity can be rearranged into an equivalent form as follows:

\[
\frac{-A}{\sqrt{B}} - \frac{1}{2} \geq \frac{(A+x)^2}{2(B+x^2)} - \frac{x^2}{B} + \frac{x}{\sqrt{B}}
\]

so that it suffices to show the second line above. Notice that the RHS of this expression is convex in $A$ and so is maximized at the boundary of the range $[-x - \sqrt{x^2 + B}, -\sqrt{B}]$. When $A = -\sqrt{B}$ we have:

\[
\frac{(A+x)^2}{2(B+x^2)} + \frac{A}{\sqrt{B}} + \frac{1}{2} - \frac{x^2}{B} + \frac{x}{\sqrt{B}} \leq \frac{(A+x)^2}{2B} + \frac{A}{\sqrt{B}} + \frac{1}{2} - \frac{x^2}{B} + \frac{x}{\sqrt{B}} = -\frac{x^2}{2B} \leq 0.
\]

Alternatively, when $A = -x - \sqrt{x^2 + B}$, we have

\[
\frac{(A+x)^2}{2(B+x^2)} + \frac{A}{\sqrt{B}} + \frac{1}{2} - \frac{x^2}{B} + \frac{x}{\sqrt{B}} = 1 - \frac{\sqrt{x^2 + B}}{\sqrt{B}} - \frac{x^2}{B} \leq 0.
\]

This establishes the claimed inequality (6) and completes the proof.

□