Parameter-free Regret in High Probability with Heavy Tails

Jiujia Zhang
Electrical and Computer Engineering
Boston University
jiujiaz@bu.edu

Ashok Cutkosky
Electrical and Computer Engineering
Boston University
ashok@cutkosky.com

Abstract

We present new algorithms for online convex optimization over unbounded domains that obtain parameter-free regret in high-probability given access only to potentially heavy-tailed subgradient estimates. Previous work in unbounded domains considers only in-expectation results for sub-exponential subgradients. Unlike in the bounded domain case, we cannot rely on straight-forward martingale concentration due to exponentially large iterates produced by the algorithm. We develop new regularization techniques to overcome these problems. Overall, with probability at most δ, for all comparators $\mathbf{u}$ our algorithm achieves regret $\tilde{O}(\|\mathbf{u}\|T^{1/\mathfrak{p}}\log(1/\delta))$ for subgradients with bounded $\mathfrak{p}^{th}$ moments for some $\mathfrak{p} \in (1, 2]$.

1 Introduction

In this paper, we consider the problem of online learning with convex losses, also called online convex optimization, with heavy-tailed stochastic subgradients. In the classical online convex optimization setting, given a convex set $\mathcal{W}$, a learning algorithm must repeatedly output a vector $\mathbf{w}_t \in \mathcal{W}$, and then observe a convex loss function $\ell_t : \mathcal{W} \to \mathbb{R}$ and incur a loss of $\ell_t(\mathbf{w}_t)$. After T such rounds, the algorithm's quality is measured by the *regret* with respect to a fixed competitor $\mathbf{u} \in \mathcal{W}$:

$$R_T(\mathbf{u}) = \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \sum_{t=1}^{T} \ell_t(\mathbf{u})$$

Online convex optimization is widely applicable, and has been used to design popular stochastic optimization algorithms ([Duchi et al., 2010a, Kingma and Ba, 2014, Reddi et al., 2018]), for control of linear dynamical systems [Agarwal et al., 2019], or even building concentration inequalities [Vovk, 2007, Waudby-Smith and Ramdas, 2020, Orabona and Jun, 2021].

A popular approach to this problem reduces it to *online linear optimization* (OLO): if $\mathbf{g}_t$ is a subgradient of ℓ_t at $\mathbf{w}_t$, then $R_T(\mathbf{u}) \le \sum_{t=1}^{T}\langle \mathbf{g}_t, \mathbf{w}_t - \mathbf{u}\rangle$ so that it suffices to design an algorithm that considers only linear losses $\mathbf{w} \mapsto \langle \mathbf{g}_t, \mathbf{w}\rangle$. Then, by assuming that the domain $\mathcal{W}$ has some finite diameter D, standard arguments show that online gradient descent [Zinkevich, 2003] and its variants achieve $R_T(\mathbf{u}) \le O(D\sqrt{T})$ for all $\mathbf{u} \in \mathcal{W}$. See the excellent books Cesa-Bianchi and Lugosi [2006], Shalev-Shwartz [2011], Hazan [2019], Orabona [2019] for more detail.

Deviating from the classical setting, we study the more difficult case in which, (1) the domain $\mathcal{W}$ may have *infinite* diameter (such as $\mathcal{W} = \mathbb{R}^d$), and (2) instead of observing the loss ℓ_t, the algorithm is presented only with a potentially heavy-tailed stochastic subgradient estimate $\mathbf{g}_t$ with $\mathbb{E}[\mathbf{g}_t|\mathbf{w}_t] \in \partial\ell_t(\mathbf{w}_t)$. Our goal is to develop algorithms that, with high probability, obtain essentially the same regret bound that would be achievable even if the full information was available.

Considering only the setting of infinite diameter $\mathcal{W}$ with *exact* subgradients $\mathbf{g}_t \in \partial\ell_t(\mathbf{w}_t)$, past work has achieved bounds of the form $R_T(\mathbf{u}) \le \tilde{O}(\epsilon + \|\mathbf{u}\|\sqrt{T})$ for all $\mathbf{u} \in \mathcal{W}$ simultaneously for any

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

user-specified ϵ, directly generalizing the $O(D\sqrt{T})$ rate available when $D < \infty$ [Orabona and Pál, 2016, Cutkosky and Orabona, 2018, Foster et al., 2017, Mhammedi and Koolen, 2020, Chen et al., 2021]. As such algorithms do not require knowledge of the norm $\|\mathbf{u}\|$ that is usually used to specify a learning rate for gradient descent, we will call them *parameter-free*. Note that such algorithms typically guarantee constant $R_T(0)$, which is not achieved by any known form of gradient descent.

While parameter-free algorithms appear to fully generalize the finite-diameter case, they fall short when $\mathbf{g}_t$ is a stochastic subgradient estimate. In particular, lower-bounds suggest that parameter-free algorithms must require Lipschitz ℓ_t [Cutkosky and Boahen, 2017], which means that care must be taken when using $\mathbf{g}_t$ with unbounded noise as this may make ℓ_t "appear" to be non-Lipschitz. In the case of *sub-exponential* $\mathbf{g}_t$, Jun and Orabona [2019], van der Hoeven [2019] provide parameter-free algorithms that achieve $\mathbb{E}[R_T(\mathbf{u})] \le \tilde{O}(\epsilon + \|\mathbf{u}\|\sqrt{T})$, but these techniques do not easily extend to heavy-tailed $\mathbf{g}_t$ or to high-probability bounds. The high-probability statement is particularly elusive (even with sub-exponential $\mathbf{g}_t$) because standard martingale concentration approaches appear to fail spectacularly. This failure may be counterintuitive: for *finite diameter* $\mathcal{W}$, one can observe that $\langle \mathbf{g}_t - \mathbb{E}[\mathbf{g}_t], \mathbf{w}_t - \mathbf{u}\rangle$ forms a martingale difference sequence with variance determined by $\|\mathbf{w}_t - \mathbf{u}\| \le D$, which allows for relatively straightforward high-probability bounds. However, parameter-free algorithms typically exhibit *exponentially growing* $\|\mathbf{w}_t\|$ in order to compete with all possible scales of $\|\mathbf{u}\|$, which appears to stymie such arguments.

Our work overcomes these issues. Requiring only that $\mathbf{g}_t$ have a bounded $\mathfrak{p}^{th}$ moment for some $\mathfrak{p} \in (1, 2]$, we devise a new algorithm whose regret with probability at least $1 - \delta$ is $R_T(\mathbf{u}) \le \tilde{O}(\epsilon + \|\mathbf{u}\|T^{1/\mathfrak{p}} \log(1/\delta))$ for all $\mathbf{u}$ simultaneously. The $T^{1/\mathfrak{p}}$ dependency is unimprovable Bubeck et al. [2013], Vural et al. [2022]. Moreover, we achieve these results simply by adding novel and carefully designed regularizers to the losses ℓ_t in a way that converts any parameter-free algorithm with sufficiently small regret into one with the desired high probability guarantee.

Motivation: *High-probability* analysis is appealing since it provides a confidence guarantee for an algorithm over a single run. This is crucially important in the online setting in which we must make irrevocable decisions. It is also important in the standard stochastic optimization setting encountered throughout machine learning as it ensures that even a single potentially very expensive training run will produce a good result. (See Harvey et al. [2019], Li and Orabona [2020], Madden et al. [2020], Kavis et al. [2022] for more discussion on the importance of high-probability bounds.) This goal naturally synergizes with the overall objective of *parameter-free* algorithms, which attempt to provide the best-tuned performance after a single pass over the data. In addition, we consider the presence of *heavy-tailed* stochastic gradients, which are empirically observed in large neural network architectures Zhang et al. [2020], Zhou et al. [2020].

Contribution and Organization: After formally introducing and discussing our setup in Sections 2, we then proceed to conduct an initial analysis for the 1-D case $\mathcal{W} = \mathbb{R}$ in 3. First (Section 4), we introduce a parameter-free algorithm for *sub-exponential* g_t that achieves regret $\tilde{O}(\epsilon + |u|\sqrt{T})$ in high probability. This already improves significantly on prior work, and is accomplished by introducing a novel regularizer that "cancels" some unbounded martingale concentration terms, a technique that may have wider application. Secondly (Section 5), we extend to *heavy-tailed* g_t by employing clipping, which has been used in prior work on optimization [Bubeck et al., 2013, Gorbunov et al., 2020, Zhang et al., 2020, Cutkosky and Mehta, 2021] to convert heavy-tailed estimates into sub-exponential ones. This clipping introduces some bias that must be carefully offset by yet another novel regularization (which may again be of independent interest) in order to yield our final $\tilde{O}(\epsilon + |u|T^{1/\mathfrak{p}})$ parameter-free regret guarantee. Finally (Section 6), we extend to arbitrary dimensions via the reduction from Cutkosky and Orabona [2018].

2 Preliminaries

Our algorithms interact with an adversary in which for $t = 1 \dots T$ the algorithm first outputs a vector $\mathbf{w}_t \in \mathcal{W}$ for $\mathcal{W}$ a convex subset of some real Hilbert space, and then the adversary chooses a convex and G-Lipschitz loss function $\ell_t : \mathcal{W} \to \mathbb{R}$ and a distribution P_t such that for $\mathbf{g}_t \sim P_t$, $\mathbb{E}[\mathbf{g}_t] \in \partial\ell_t(\mathbf{w}_t)$ and $\mathbb{E}[\|\mathbf{g}_t - \mathbb{E}[\mathbf{g}_t]\|^{\mathfrak{p}}] \le \sigma^{\mathfrak{p}}$ for some $\mathfrak{p} \in (1, 2]$. The algorithm then observes a random sample $\mathbf{g}_t \sim P_t$. After t rounds, we compute the *regret*, which is a function $R_t(\mathbf{u}) = \sum_{i=1}^t \ell_i(\mathbf{w}_i) - \ell_i(\mathbf{u})$. Our goal is to guarantee $R_T(\mathbf{u}) \le \epsilon + \tilde{O}(\|\mathbf{u}\|T^{1/\mathfrak{p}})$ for all $\mathbf{u}$ simultaneously with high probability.

Throughout this paper we will employ the notion of a *sub-exponential* random sequence:

Definition 1. *Suppose $\{X_t\}$ is a sequence of random variables adapted to a filtration $\mathcal{F}_t$ such that $\{X_t, \mathcal{F}_t\}$ is a martingale difference sequence. Further, suppose $\{\sigma_t, b_t\}$ are random variables such that σ_t, b_t are both $\mathcal{F}_{t-1}$-measurable for all t. Then, $\{X_t, \mathcal{F}_t\}$ is $\{\sigma_t, b_t\}$ sub-exponential if*

$$\mathbb{E}[\exp(\lambda X_t)|\mathcal{F}_{t-1}] \le \exp(\lambda^2\sigma_t^2/2)$$

almost everywhere for all $\mathcal{F}_{t-1}$-measurable λ satisfying $\lambda < 1/b_t$.

We drop the subscript t when we have uniform (not time-varying) sub-exponential parameters (σ, b). We use bold font $(\mathbf{g}_t)$ to refer to vectors and normal font (g_t) to refer to scalars. Occasionally, we abuse notation to write $\nabla \ell_t(\mathbf{w}_t)$ for an arbitrary element of $\partial \ell_t(\mathbf{w}_t)$.

We present our results using $O(\cdot)$ to hide constant factors, and $\tilde{O}(\cdot)$ to hide $\log$ factors (such as some power of $\log T$ dependence) in the main text, the exact results are left at the last line of the proof for interested readers.

Finally, observe that by the unconstrained-to-constrained conversion of Cutkosky and Orabona [2018], we need only consider the case that $\mathcal{W}$ is an entire vector space. By solving the problem for this case, the reduction implies a high-probability regret algorithm for any convex $\mathcal{W}$.

3 Challenges

A reader experienced with high probability bounds in online optimization may suspect that one could apply fairly standard approaches such as gradient clipping and martingale concentration to easily achieve high probability bounds with heavy tails. While such techniques do appear in our development, the story is far from straightforward. In this section, we will outline these non-intuitive difficulties. For a further discussion, see Section 3 of Jun and Orabona [2019].

For simplicity, consider $w_t \in \mathbb{R}$. Before attempting a high probability bound, one may try to derive a regret bound in expectation with heavy-tailed (or even light-tailed) gradient g_t via the following calculation:

$$\mathbb{E}[R_T(u)] = \mathbb{E}\left[\sum_{t=1}^T \ell_t(w_t) - \ell_t(u)\right] \le \sum_{t=1}^T \mathbb{E}\left[\langle g_t, w_t - u\rangle\right] + \sum_{t=1}^T \mathbb{E}\left[\langle \nabla \ell_t(w_t) - g_t, w_t - u\rangle\right]$$

The second sum from above vanishes, so one is tempted to send g_t directly to some existing parameter-free algorithm to obtain low regret. Unfortunately, most parameter-free algorithms require a uniform bound on $|g_t|$ - even a *single* bound-violating g_t could be catastrophic [Cutkosky and Boahen, 2017]. With heavy-tailed g_t, we are quite likely to encounter such a bound-violating g_t for any reasonable uniform bound. In fact, the issue is difficult even for light-tailed g_t, as described in detail by Jun and Orabona [2019].

A natural approach to overcome this uniform bound issue is to incorporate some form of clipping, a commonly used technique controlling for heavy-tailed subgradients. The clipped subgradient $\hat{g}_t$ is defined below with a positive clipping parameter τ as:

$$\hat{g}_t = \frac{g_t}{|g_t|}\min(\tau, |g_t|)$$

If we run algorithms on uniformly bounded $\hat{g}_t$ instead, the expected regret can now be written as:

$$\mathbb{E}[R_T(u)] \le \underbrace{\sum_{t=1}^T \mathbb{E}\left[\langle \hat{g}_t, w_t - u\rangle\right]}_{\text{parameter-free regret}} + \underbrace{\sum_{t=1}^T \mathbb{E}\left[\langle \mathbb{E}[\hat{g}_t] - \hat{g}_t, w_t - u\rangle\right]}_{\text{martingale concentration?}} + \underbrace{\sum_{t=1}^T \mathbb{E}\left[\langle \nabla \ell_t(w_t) - \mathbb{E}[\hat{g}_t], w_t - u\rangle\right]}_{\text{bias}} \tag{1}$$

Since $|\hat{g}_t| \le \tau$, the first term can in fact be controlled for appropriate τ at a rate of $\tilde{O}(\epsilon + |u|\sqrt{T})$ using sufficiently advanced parameter-free algorithms (e.g. Cutkosky and Orabona [2018]). However, now bias accumulates in the last term, which is difficult to bound due to the dependency on w_t. On the surface, understanding this dependency appears to require detailed (and difficult) analysis of the dynamics of the parameter-free algorithm. In fact, from naive inspection of the updates for standard parameter-free algorithms, one expects that $|w_t|$ could actually grow exponentially fast in t, leading to a very large bias term.

Finally, disregarding these challenges faced even in expectation, to derive a high-probability bound the natural approach is to bound the middle sum in (1) via some martingale concentration argument. Unfortunately, the variance process for this martingale depends on w_t just like the bias term. In fact, this issue appears even if the original g_t already have bounded norm, which is the most extreme version of *light* tails! Thus, we again appear to encounter a need for small w_t, which may instead grow exponentially. In summary, the unbounded nature of w_t makes dealing with any kind of stochasticity in the g_t very difficult. In this work we will develop techniques based on regularization that intuitively force the w_t to behave well, eventually enabling our high-probability regret bounds.

4 Bounded Sub-exponential Noise via Cancellation

In this section, we describe how to obtain regret bound in high probability for stochastic subgradients g_t for which $\mathbb{E}[g_t^2] \le \sigma^2$ and $|g_t| \le b$ for some σ and b (in particular, g_t exhibits $(\sigma, 4b)$ sub-exponential noise). We focus on the 1-dimensional case with $\mathcal{W} = \mathbb{R}$. The extension to more general $\mathcal{W}$ is covered in Section 6. Our method involves two coordinated techniques. First, we introduce a carefully designed regularizer ψ_t such that *any algorithm* that achieves low regret with respect to the losses $w \mapsto g_t w + \psi_t(w)$ will automatically ensure low regret with high probability on the original losses ℓ_t. Unfortunately, ψ_t is not Lipschitz and so it is still not obvious how to obtain low regret. We overcome this final issue by an "implicit" modification of the optimistic parameter-free algorithm of Cutkosky [2019]. Our overall goal is a regret bound of $R_T(u) \le \tilde{O}(\epsilon + |u|(\sigma + G)\sqrt{T} + b|u|)$ for all u with high probability. Note that with this bound, b can be $O(\sqrt{T})$ before it becomes a significant factor in the regret.

Let us proceed to sketch the first (and most critical) part of this procedure: Define $\epsilon_t = \nabla \ell_t(w_t) - g_t$, so that ϵ_t captures the "noise" in the gradient estimate g_t. In this section, we assume that ϵ_t is $(\sigma, 4b)$ sub-exponential for all t for some given σ, b and $|g_t| \le b$. Then we can write:

$$
\begin{aligned}
R_T(u) &\le \sum_{t=1}^T \langle \nabla \ell_t(w_t), w_t - u\rangle = \sum_{t=1}^T \langle g_t, w_t - u\rangle + \sum_{t=1}^T \langle \epsilon_t, w_t\rangle - \sum_{t=1}^T \langle \epsilon_t, u\rangle \\
&\le \sum_{t=1}^T \langle g_t, w_t - u\rangle + \underbrace{\left|\sum_{t=1}^T \epsilon_t w_t\right| + |u| \left|\sum_{t=1}^T \epsilon_t\right|}_{\text{"noise term", NOISE}} \qquad (2)
\end{aligned}
$$

Now, the natural strategy is to run an OLO algorithm $\mathcal{A}$ on the observed g_t, which will obtain some regret $R_T^{\mathcal{A}}(u) = \sum_{t=1}^T \langle g_t, w_t - u\rangle$, and then show that the remaining NOISE terms are small. To this end, from sub-exponential martingale concentration, we might hope to show that with probability $1-\delta$, we have an identity similar to:

$$
\text{NOISE} \le \sigma \sqrt{\sum_{t=1}^T w_t^2 \log(1/\delta)} + b \max_t |w_t| \log(1/\delta) + |u|\sigma\sqrt{T\log(1/\delta)} + |u| b \log(1/\delta)
$$

The dependency of $|u|$ above appears to be relatively innocuous as it only contributes $\tilde{O}(|u|\sigma\sqrt{T} + |u|b)$ to the regret. The w_t-dependent term is more difficult as it involves a dependency on the algorithm $\mathcal{A}$. This captures the complexity of our unbounded setting: in a *bounded domain*, the situation is far simpler as we can uniformly bound $|w_t| \le D$, ideally leaving us with an $\tilde{O}(D\sqrt{T})$ bound overall.

Unfortunately, in the unconstrained case, $|w_t|$ could grow exponentially ($|w_t| \sim 2^t$) even when u is very small, so we cannot rely on a uniform bound. In fact, even in the finite-diameter case, if we wish to guarantee $R_T(0) \le \epsilon$, the bound $|w_t| \le D$ is still too coarse. The resolution is to instead feed the algorithm $\mathcal{A}$ a *regularized* loss $\hat{\ell}_t(w) = \langle g_t, w\rangle + \psi_t(w)$, where ψ_t will "cancel" the w_t dependency in the martingale concentration. That is, we now define $R_T^{\mathcal{A}}(u) = \sum_{t=1}^T \hat{\ell}_t(w_t) - \hat{\ell}_t(u)$

and rearrange:

$$\sum_{t=1}^T \langle g_t, w_t - u\rangle \le R_T^{\mathcal{A}}(u) - \sum_{t=1}^T \psi_t(w_t) + \sum_{t=1}^T \psi_t(u) \tag{3}$$

And now combine equations (2) and (3):

$$\begin{aligned} R_T(u) &\le R_T^{\mathcal{A}}(u) - \sum_{t=1}^T \psi_t(w_t) + \sum_{t=1}^T \psi_t(u) + \text{NOISE} \\ &\le R_T^{\mathcal{A}}(u) + \sigma\sqrt{\sum_{t=1}^T w_t^2 \log(1/\delta)} + b\max_t |w_t|\log(1/\delta) - \sum_{t=1}^T \psi_t(w_t) \\ &\quad + |u|\sigma\sqrt{T\log(1/\delta)} + |u|b\log(1/\delta) + \sum_{t=1}^T \psi_t(u) \end{aligned} \tag{4}$$

From this, we can read off the desired properties of ψ_t: (1) ψ_t should be large enough that $\sum_{t=1}^T \psi_t(w_t) \ge \sigma\sqrt{\sum_{t=1}^T w_t^2 \log(1/\delta)} + b\max_t |w_t|\log(1/\delta)$, (2) ψ_t should be small enough that $\sum_{t=1}^T \psi_t(u) \le \tilde{O}(|u|\sqrt{T})$, and (3) ψ_t should be such that $R_T^{\mathcal{A}}(u) = \tilde{O}(\epsilon + |u|\sqrt{T})$ for an appropriate algorithm $\mathcal{A}$. If we can exhibit a ψ_t satisfying all three properties, we will have developed a regret bound of $\tilde{O}(\epsilon + |u|\sqrt{T})$ in high probability.

It turns out that the modified Huber loss $r_t(w)$ defined in equation (5) and (6) with appropriately chosen constants $c_1, c_2, p_1, p_2, \alpha_1, \alpha_2$ satisfies criterion (1) and (2).

$$r_t(w; c, p, \alpha_0) = \begin{cases} c\left(p|w| - (p-1)|w_t|\right)\frac{|w_t|^{p-1}}{(\sum_{i=1}^t |w_i|^p + \alpha_0^p)^{1-1/p}}, & |w| > |w_t| \\ c|w|^p \frac{1}{(\sum_{i=1}^t |w_i|^p + \alpha_0^p)^{1-1/p}}, & |w| \le |w_t| \end{cases} \tag{5}$$

$$\psi_t(w) = r_t(w; c_1, p_1, \alpha_1) + r_t(w; c_2, p_2, \alpha_2) \tag{6}$$

Let us take a moment to gain some intuition for these functions r_t and ψ_t. First, observe that r_t is always continuously differentiable, and that r_t's definition requires knowledge of w_t. This is acceptable because online learning algorithms must be able to handle even adaptively chosen losses. In particular, consider the $p = 2$ case, $r_t(w; c, 2, \alpha)$ for some positive constants c and α. We plot this function in Figure 1, where one can see that r_t grows quadratically for $|w| \le |w_t|$, but grows only linearly afterwards so that r_t is Lipschitz.

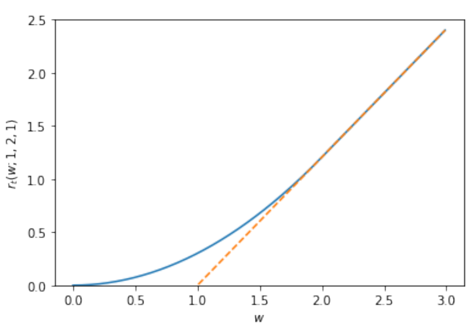

Figure 1: $r_t(w; 1, 2, 1)$ when $\sum_{i=1}^t w_i^2 = 10$ and $w_t = 2$. The dashed line has slope $cp\frac{|w_t|^{p-1}}{(\sum_{i=1}^t |w_i|^p + \alpha_0^p)^{1/p}}$, so that r_t is quadratic for $|w| \le |w_t|$ and linear otherwise. Notice that w_t is a constant used to define r_t - it is not the argument of the function.

Eventually, in Lemma 13 we will show that this functions satisfies

$$\sum_{t=1}^T r_t(w_t; c, 2, \alpha) \ge c\sqrt{\sum_{t=1}^T w_t^2} - \alpha$$

$$\sum_{t=1}^T r_t(u : c, 2, \alpha) \le \tilde{O}(u\sqrt{T})$$

so that for appropriate choice of c and α, $r_t(w; c, 2, \alpha)$ will cancel the $O(\sqrt{\sum_{t=1}^T w_t^2})$ martingale concentration term while not adding too much to the regret - it satisfies criteria (1) and (2). The lower-bound follows from the standard inequality $\sqrt{a+b} \le \sqrt{a} + \frac{b}{\sqrt{a+b}}$ since $r_t(w_t) = c\frac{w_t^2}{\sqrt{\alpha^2 + \sum_{i=1}^t w_i^2}}$. The upper-bound is more subtle, and involves the piece-wise definition. For simplicity, suppose it were true that either $|w_t| < |u|$ for all t or $|w_t| \ge |u|$ for all t. In the former case,

$\sum_{t=1}^{T} r_t(u) = O\left(|u| \sum_{t=1}^{T} \frac{|w_t|}{\sqrt{\alpha + \sum_{i=1}^{t} w_i^2}}\right)$, which via algebraic manipulation can be bounded as $\tilde{O}(|u|\sqrt{T})$. In the latter case, we have $\sum_{t=1}^{T} \frac{|u|^2}{\sqrt{\alpha^2 + \sum_{i=1}^{t} w_t^2}} \le \sum_{t=1}^{T} \frac{|u|^2}{\sqrt{\alpha^2 + tu^2}} = \tilde{O}(|u|\sqrt{T})$ so that both cases result in the desired bound on $\sum_{t=1}^{T} r_t(u)$. The general setting is handled by partitioning the sum into two sets depending on whether $|w_t| \le |u|$. In order to cancel the $\max_t |w_t|$ term in the martingale concentration, we employ $p = \log T$. This choice is motivated by the observation that $\|\mathbf{v}\|_{\log T} \in [\|\mathbf{v}\|_\infty, \exp(1)\|\mathbf{v}\|_\infty]$ for all $\mathbf{v} \in \mathbb{R}^{\log T}$. With this identity in hand, the argument is very similar to the $p = 2$ case.

The correct values for the constants are provided in Theorem 3. Again, at a high level, the important constants are p_1 and p_2. With $p_1 = 2$, we allow $\sum_t r_t(w_t; p = 2)$ to cancel out the $\sqrt{\sum_t w_t^2}$ martingale concentration term, while with $p_2 = \log T$, $\sum_t r_t(w_t; p = \log T)$ cancels that $\max_t |w_t|$ term.

It remains to show that ψ_t also allows for small $R_T^{\mathcal{A}}(u)$ and so satisfies criterion (3). Unfortunately, our setting for c_2 in the definition of ψ_t is $\tilde{O}(b)$, which means that ψ_t is $\tilde{O}(b)$-Lipschitz. Since we wish to allow for $b = \Theta(\sqrt{T})$, this means that we cannot simply let $\mathcal{A}$ linearize ψ_t and apply an arbitrary OLO algorithm. Instead, we must exploit the fact that ψ_t is known *before* g_t is revealed. That is, algorithm $\mathcal{A}$ is chosen to exploit the structure composite loss $\hat{\ell}_t(w)$. Intuitively, the regret of a composite loss should depend only on the non-composite g_t terms (as in e.g. Duchi et al. [2010b]). Our situation is slightly more complicated as ψ_t depends on w_t as well, but we nevertheless achieve the desired result via a modification of the parameter-free optimistic reduction in Cutkosky [2019]. For technical reasons, this algorithm still requires $|g_t| \le b$ with probability 1, but obtains regret only $R_T^{\mathcal{A}}(u) \le \tilde{O}(\epsilon + |u|\sigma\sqrt{T} + |u|b)$. This technical limitation is lifted in the following section.

We display the method as Algorithm 1, which provides a regularization that cancels the $|w_t|$ dependent part of the NOISE term in (7). It also allows us to control $R_T^{\mathcal{A}}(u)$ to order $\tilde{O}(\epsilon + |u|\sigma\sqrt{T} + b|u|)$ by taking account into the predictable structure of regularizer $\psi_t(w)$. The algorithm requires black-box access to two base online learning algorithms, which we denote $\mathcal{A}_1$ and $\mathcal{A}_2$ with domains $(-\infty, \infty)$ and $[0, \infty)$ respectively. These can be any algorithms that obtain so-called "second-order" parameter-free regret bounds, such as available in Cutkosky and Orabona [2018], van der Hoeven [2019], Kempka et al. [2019], Mhammedi and Koolen [2020]. Roughly speaking, the role of $\mathcal{A}_1$ is to provide an initial candidate ouput x_t that is then "corrected" by $\mathcal{A}_2$ using the regularization to obtain the final w_t.

Following the intuition previously outlined in this section, We first provide a deterministic regret guarantee on the quantity $R_T^{\mathcal{A}}(u) = \sum_{t=1}^{T} \hat{\ell}_t(w_t) - \hat{\ell}_t(u)$ as an intermediate result (Theorem 2). Then, we provide the analysis of the full procedure of Algorithm 1 for the final high probability result (Theorem 3). Missing proofs are provided in the Appendix A and B.

Algorithm 1 Sub-exponential Noisy Gradients with Optimistic Online Learning

---

Require: $E[g_t] = \nabla \ell_t(w_t)$, $|g_t| \le b, \mathbb{E}[g_t|w_t] \le \sigma^2$ almost surely. Two online learning algorithms (e.g. copies of Algorithm 1 from Cutkosky and Orabona [2018]) labelled as $\mathcal{A}_1, \mathcal{A}_2$ with domains $\mathbb{R}$ and $\mathbb{R}_{\ge 0}$ respectively. Time horizon T, $0 < \delta \le 1$.

Initialize:
  Constants $\{c_1, c_2, p_1, p_2, \alpha_1, \alpha_2\}$ from Theorem 3. ▷ for defining ψ_t in equation (6)
  $H = c_1 p_1 + c_2 p_2$
for $t = 1$ to T **do**
  Receive x_t' from $\mathcal{A}_1$, y_t' from $\mathcal{A}_2$
  Rescale $x_t = x_t'/(b + H)$, $y_t = y_t'/(H(b + H))$
  Solve for w_t: $w_t = x_t - y_t \nabla \psi_t(w_t)$ ▷ The solution exists by Lemma 6
  Play w_t to, suffer loss $\ell_t(w_t)$
  Receive g_t with $\mathbb{E}[g_t] \in \partial \ell_t(w_t)$
  Compute $\psi_t(w) = r_t(w; c_1, p_1, \alpha_1) + r_t(w; c_2, p_2, \alpha_2)$ and $\nabla \psi_t(w_t)$ ▷ equations (5), (6)
  Send $(g_t + \nabla \psi_t(w_t))/(b + H)$ to $\mathcal{A}_1$
  Send $-\langle g_t + \nabla \psi_t(w_t), \nabla \psi_t(w_t) \rangle / H(b + H)$ to $\mathcal{A}_2$
end for

Theorem 2. *Suppose $\mathcal{A}_1$ ensure that given some $\epsilon > 0$ and a sequence c_t with $|c_t| \le 1$:*

$$\sum_{t=1}^{T}\langle c_t, w_t - u\rangle \le \epsilon + A|u|\sqrt{\sum_{t=1}^{T}|c_t|^2\left(1+\log\left(\frac{|u|^2T^C}{\epsilon^2}+1\right)\right)} + B|u|\log\left(\frac{|u|T^C}{\epsilon}+1\right)$$

for all u for some positive constants A, B, C, and that $\mathcal{A}_2$ obtains the same guarantee for all $u \ge 0$, then for $|g_t| \le b$, $|\nabla\psi_t(w_t)| \le H$, we have the following guarantee from Algorithm 1,

$$R_T^{\mathcal{A}}(u) \le O\left[\epsilon + |u|\left(\sqrt{\max\left(0, \sum_{t=1}^{T}|g_t|^2 - |\nabla\psi_t(w_t)|^2\right)} + (b+H)\log T\right)\right]$$

Although this Theorem 2 is rather technical, the overall message is not too complicated. If we ignore the negative $|\nabla\psi_t(w_t)|^2$ terms, the bound simply says that the regret on the "composite" loss $\langle g_t, w\rangle + \psi_t(w)$ only increases with the apriori-unknown g_t, and *not* with $\nabla\psi_t(w_t)$. With this result, we can formalize the intuition in this section to provide the following high probability regret bound:

Theorem 3. *Suppose $\{g_t\}$ are stochastic subgradients such that $\mathbb{E}[g_t] \in \partial\ell_t(w_t)$, $|g_t| \le b$ and $\mathbb{E}[g_t^2|w_t] \le \sigma^2$ almost surely for all t. Set the following constants for $\psi_t(w)$ shown in equation (6) for any $0 < \delta \le 1$, $\epsilon > 0$,*

$$c_1 = 2\sigma\sqrt{\log\left(\tfrac{32}{\delta}\left[\log\left(2^{T+1}\right)+2\right]^2\right)}, \quad c_2 = 32b\log\left(\tfrac{224}{\delta}\left[\log\left(1+\tfrac{b}{\sigma}2^{T+2}\right)+2\right]^2\right),$$
$$p_1 = 2, \qquad p_2 = \log T, \qquad \alpha_1 = \epsilon/c_1, \qquad \alpha_2 = \epsilon\sigma/(4b(b+H))$$

where $H = c_1p_1 + c_2p_2$, $|\nabla\psi_t(w_t)| \le H$. Then, with probability at least $1-\delta$, algorithm 1 guarantees

$$R_T(u) \le \tilde{O}\left[\epsilon\log\frac{1}{\delta} + |u|b\log\frac{1}{\delta} + |u|\sigma\sqrt{T\log\frac{1}{\delta}}\right]$$

Note that this result is *already* of interest: prior work on parameter-free algorithms with sub-exponential noise only achieve in-expectation rather than high probability results. Of course, there is a caveat: our bound requires that $|g_t|$ be uniformly bounded by b. Even though b could be as large as $\sqrt{T}$, this is still a mild restriction. In the next section, we remove both this restriction as well as the light tail assumption all together.

5 Heavy tails via Truncation

In this section, we aim to give a high probability bound for heavy-tailed stochastic gradients $\mathbf{g}_t$. Our approach builds on Section 4 by incorporating gradient clipping with a clipping parameter $\tau \in \mathbb{R}^+$.

$$\hat{\mathbf{g}}_t = \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|}\min(\tau, \|\mathbf{g}_t\|)$$

We continue to consider a 1-dimensional problem in this section, replacing the norm $\|\cdot\|$ with absolute value $|\cdot|$ and $\mathbf{g}_t$ with g_t. The key insight is that the clipped $\hat{g}_t$ satisfies $\mathbb{E}[\hat{g}_t^2] \le 2^{\mathfrak{p}-1}\tau^{2-\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})$ and of course $|\hat{g}_t| \le \tau$. Hence, a high probability bound could be obtained by feeding $\hat{g}_t$ into Algorithm 1 from Section 4. Let us formally quantify the effect of this clipping:

$$R_T(u) \le \sum_{t=1}^{T}\langle\nabla\ell_t(w_t), w_t - u\rangle = \underbrace{\sum_{t=1}^{T}\langle\nabla\ell_t(w_t) - \mathbb{E}[\hat{g}_t], w_t - u\rangle}_{\text{bias}} + \underbrace{\sum_{t=1}^{T}\langle\mathbb{E}[\hat{g}_t], w_t - u\rangle}_{\text{Section 4}} \tag{7}$$

Without clipping, we would have $\mathbb{E}[\hat{g}_t] = \nabla\ell_t(w_t)$, and so if we were satisfied with an in-expectation result, the first sum above would vanish. However, with clipping, the first sum actually represents some "bias" that must be controlled even to obtain an in-expectation result, let alone high probability. We control this bias using a cancellation-by-regularization strategy analogous at a high level to the one developed in Section 4, although technically quite distinct. After dealing with the bias, we must handle the second sum. Fortunately, since $\hat{g}_t$ is sub-exponential, bounding the second sum in high probability is precisely the problem solved in Section 4. We introduce the analysis in two elementary steps. For the purpose of bias cancellation, we define a linearized loss $\tilde{\ell}_t(w)$ with regularization function $\phi(w)$

$$\tilde{\ell}_t(w) = \langle \mathbb{E}[\hat{g}_t], w\rangle + \phi(w), \qquad \phi(w) = 2^{\mathfrak{p}-1}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})|w|/\tau^{\mathfrak{p}-1} \tag{8}$$

the regret in equation (7) can be re-written as

$$= \underbrace{\sum_{t=1}^{T} \left(\langle \nabla \ell_t(w_t) - \mathbb{E}[\hat{g}_t], w_t - u\rangle - \phi(w_t) + \phi(u)\right)}_{\text{bias cancellation}} + \underbrace{\sum_{t=1}^{T} \tilde{\ell}_t(w_t) - \tilde{\ell}_t(u)}_{\text{Section 4}} \tag{9}$$

We will be able to show that the w_t-dependent terms of the first summation sum to a negative number and so can be dropped. This leaves only the u-dependent terms, which for appropriate choice of τ will be $\tilde{O}(|u|T^{1/\mathfrak{p}})$.

Note that at this point, if we were satisfied with an *in expectation* bound for heavy-tailed subgradient estimates (which would already be an interesting new result), we would not require the techniques of Section 4: we could instead define $\hat{\ell}_t(w) = \langle \hat{g}_t, w\rangle + \psi(w)$, so that the last sum is equal to $\sum_{t=1}^{T} \hat{\ell}_t(w_t) - \hat{\ell}_t(u)$ in expectation. Then, since $|\hat{\ell}_t(w_t)| \le O(\tau)$ with probability 1, we can control $\sum_{t=1}^{T} \hat{\ell}_t(w_t) - \hat{\ell}_t(u)$ using a parameter-free algorithm obtaining regret $\tilde{O}(|u|\sqrt{\sum_{t=1}^{T} |\nabla \hat{\ell}_t(w_t)|^2} + \tau|u|)$ to bound the total expected regret, yielding a simple way to recover prior work on expected regret with sub-exponential subgradients (up to logs), while extending the results to heavy-tailed subgradients.

However, since we *do* aim for a high probability bound, we need to be more careful with the second summation. Fortunately, given that $\hat{g}_t$ is sub-exponential and bounded, and $\nabla\phi(w_t)$ is deterministic, we can supply $\hat{g}_t + \nabla\phi(w_t)$ to Algorithm 1 and then bound the sum in high probability by Theorem 3. We formalize the procedure as Algorithm 2, and its guarantee is stated in Theorem 4. The exact regret guarantee (including constants) can be found in Appendix C.

Algorithm 2 Gradient clipping for $(\sigma, G)-$Heavy tailed gradients

---

Require: $\mathbb{E}[g_t] = \nabla \ell_t(w_t)$, $|\mathbb{E}[g_t]| \le G$, $\mathbb{E}[|g_t - \mathbb{E}[g_t]|^{\mathfrak{p}}] \le \sigma^{\mathfrak{p}}$ for some $\mathfrak{p} \in (1, 2]$, Time horizon T, gradient clipping parameter τ.
1: Initialize Algorithm 1 using the parameters of Theorem 3.
2: **for** $t = 1$ to T **do**
3: Receive w_t from Algorithm 1.
4: Suffer loss $\ell_t(w_t)$, receive g_t
5: Truncate $\hat{g}_t = \frac{g_t}{|g_t|}\min(\tau, |g_t|)$.
6: Compute $\tilde{g}_t = \hat{g}_t + \nabla\phi_t(w_t)$ ▷ $\phi(w)$ is defined in (8), $\mathbb{E}[\tilde{g}_t] \in \partial\tilde{\ell}_t(w_t)$.
7: Send $\tilde{g}_t$ to Algorithm 1 as t^{th} subgradient.
8: **end for**

Theorem 4. *Suppose $\{g_t\}$ are heavy-tailed stochastic gradient such that $\mathbb{E}[g_t] \in \partial\ell_t(w_t)$, $|\mathbb{E}[g_t]| \le G$, $\mathbb{E}[|g_t - \mathbb{E}[g_t]|^{\mathfrak{p}}] \le \sigma^{\mathfrak{p}}$ for some $\mathfrak{p} \in (1, 2]$. If we set $\tau = T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}$ then with probability at least $1 - \delta$, Algorithm 2 guarantees:*

$$R_T(u) \le \tilde{O}\left[\epsilon \log\frac{1}{\delta} + |u|T^{1/\mathfrak{p}}(\sigma + G)\log\frac{T}{\delta}\log\frac{|u|T}{\epsilon}\right]$$

Theorem 4 suggests regret with heavy-tailed gradients g_t has a $\mathfrak{p}$ dependence of $\tilde{O}(T^{1/\mathfrak{p}})$, which is optimal [Bubeck et al., 2013, Vural et al., 2022].

6 Dimension-free Extension

So far, we have only considered 1-dimensional problems. In this section, we demonstrate the extension to dimension-free, which is achieved by using a reduction from Cutkosky and Orabona [2018]. The original reduction extends a 1-dimensional algorithm to a dimension-free one by dissecting the problem into a "magnitude" and a "direction" learner. The direction learner is a constrained OLO algorithm $\mathcal{A}^{nd}$ which outputs a vector $\mathbf{v}_t$ with $\|\mathbf{v}_t\| \leq 1$ in response to $\mathbf{g}_1, \ldots, \mathbf{g}_{t-1}$, while the magnitude learner is an unconstrained OLO algorithm $\mathcal{A}^{1d}$ which outputs $x_t \in \mathbb{R}$ in response to $\langle \mathbf{g}_1, \mathbf{v}_1 \rangle, \ldots \langle \mathbf{g}_{t-1}, \mathbf{v}_{t-1} \rangle$. The output of the entire algorithm is $\mathbf{w}_t = x_t \mathbf{v}_t$. Suppose $\mathcal{A}_{1d}$ and $\mathcal{A}_{nd}$ have regret guarantee of $R_T^{1d}(u)$ and $R_T^{nd}(\mathbf{u})$, respectively. Then regret of the dimension-free reduction is bounded by $R_T(\mathbf{u}) \leq \|\mathbf{u}\| R_T^{nd}(\mathbf{u}/\|\mathbf{u}\|) + R_T^{1d}(\|\mathbf{u}\|)$. Thus, in order to apply this reduction we need to exhibit a $\mathcal{A}^{1d}$ and $\mathcal{A}^{nd}$ that achieves low regret on heavy-tailed losses. For the magnitude learner $\mathcal{A}^{1d}$, we use can use the 1d Algorithm 2 that we just developed. The remaining question is how to develop a direction learner that can handle heavy-tailed subgradients. Fortunately, this is much easier since the direction learner is constrained to the unit ball.

To build this direction learner, we again apply subgradient clipping, and feed the clipped subgradients to the standard FTRL algorithm with quadratic regularizer (i.e. "lazy" online gradient descent). This procedure is described in Algorithm 3. Note there is no regularization implemented in Algorithm 3 although $\hat{\mathbf{g}}_t$ induces bias. This is because of $\mathcal{A}^{nd}$ runs on the unit ball, careful tuning of τ is sufficient to control the bias - a concrete demonstration of how much more intricate the unconstrained case is! Finally, the full dimension-free reduction is displayed in Algorithm 4 with its high probability guarantee stated in Theorem 5. The details are presented in Appendix D.

Algorithm 3 Unit Ball Gradient clipping with FTRL

Require: time horizon T, gradient clipping parameter τ, regularizer weight η

Set $\eta = 1/\tau$
for $t = 1$ to T **do**
    Compute $\mathbf{v}_t \in \operatorname{argmin}_{\mathbf{v}:\|\mathbf{v}\| \leq 1} \sum_{i=1}^{t-1} \langle \hat{\mathbf{g}}_t, \mathbf{v} \rangle + \frac{1}{2\eta} \|\mathbf{v}\|^2$
    Output $\mathbf{v}_t$, receive gradient $\mathbf{g}_t$
    Set $\hat{\mathbf{g}}_t = \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|} \min(\tau, \|\mathbf{g}_t\|)$
end for

Algorithm 4 Dimension-free Gradient clipping for (σ, G) Heavy-tailed gradients

---

Require: Subgradients $\mathfrak{p}^{th}$ moment bound $\sigma^{\mathfrak{p}}$, time horizon T, Set Algorithm 2, 3 as $\mathcal{A}^{1d}$, $\mathcal{A}^{nd}$.

Set $\sigma_{1d} = (\sigma^{\mathfrak{p}} + 2G^{\mathfrak{p}})^{1/\mathfrak{p}}$ and $\tau_{1d} = T^{1/\mathfrak{p}}(\sigma_{1d}^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}} = T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + 3G^{\mathfrak{p}})^{1/\mathfrak{p}}$
Initialize $\mathcal{A}^{1d}$ with parameters $\sigma \leftarrow \sigma_{1d}$ and $\tau \leftarrow \tau_{1d}$
Initialize $\mathcal{A}^{nd}$ with parameters $\sigma \leftarrow \sigma$ and $\tau \leftarrow T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}$.
for $t = 1$ to T **do**
    Receive $x_t \in \mathbb{R}$ from $\mathcal{A}^{1d}$,
    Receive $\mathbf{v}_t \in \mathbb{R}^d$, $\|\mathbf{v}_t\| \leq 1$ from $\mathcal{A}^{nd}$
    Play output $\mathbf{w}_t = x_t \mathbf{v}_t$
    Suffer loss $\ell_t(\mathbf{w}_t)$, receive gradients $\mathbf{g}_t$
    Send $g_t = \langle \mathbf{g}_t, \mathbf{v}_t \rangle$ as the t^{th} gradient to $\mathcal{A}^{1d}$
    Send $\mathbf{g}_t$ as the t^{th} gradient to $\mathcal{A}^{nd}$
end for

Theorem 5. *Suppose that for all t, $\{\mathbf{g}_t\}$ are heavy-tailed stochastic subgradients satisfying $\mathbb{E}[\mathbf{g}_t] \in \partial \ell_t(\mathbf{w}_t)$, $\|\mathbb{E}[\mathbf{g}_t]\| \leq G$ and $\mathbb{E}[\|\mathbf{g}_t - \mathbb{E}[\mathbf{g}_t]\|^{\mathfrak{p}}] \leq \sigma^{\mathfrak{p}}$ for some $\mathfrak{p} \in (1, 2]$. Then, with probability at least $1 - \delta$, Algorithm 4 guarantees*

$$R_T(\mathbf{u}) = \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u}) \leq \tilde{O}\left[\epsilon \log \frac{1}{\delta} + \|\mathbf{u}\| T^{1/\mathfrak{p}}(\sigma + G) \log \frac{T}{\delta} \log \frac{\|\mathbf{u}\| T}{\epsilon}\right]$$

Complexity Analysis: Algorithm 4 requires $O(d)$ space. It also requires $O(d)$ time for all operations except solving the fixed-point equation in Algorithm 1 (line 5). This can be solved via binary search to arbitrary precision ϵ_0 for an overall complexity of $O(d + \log(1/\epsilon_0))$. This is essentially $O(d)$ in practice for any $d > 64$.

7 Conclusion

We have presented a framework for building parameter-free algorithms that achieve high probability regret bounds for heavy-tailed subgradient estimates. This improves upon prior work in several ways: high probability bounds were previously unavailable even for the restricted setting of *bounded* subgradient estimates, while even in-expectation bounds were previously unavailable for heavy-tailed subgradients. Our development required two new techniques: first, we described a regularization scheme that effectively "cancels" potentially problematic iterate-dependent variance terms arising in standard martingale concentration arguments. This allows for high probability bounds with bounded sub-exponential estimates, and we hope may be of use in other scenarios where the iterates appear in variance calculations. The second combines clipping with another new regularization scheme that "cancels" another problematic iterate-dependent *bias* term. On its own, this technique actually can be used to recover in-expectation bounds for heavy-tailed estimates.

Limitations: Our algorithm has several limitations that suggest open questions: first, our two regularization schemes each introduce potentially suboptimal logarithmic factors. The first one introduces a higher logarithmic dependence on T, while the second introduces a higher logarithmic dependence on $\|\mathbf{u}\|$ because the optimal clipping parameter τ depends on $\log(\|\mathbf{u}\|)$. Beyond this, our algorithms require knowledge of the parameters σ and τ. Adapting to an unknown value of even one of these parameters remains a challenging problem.

References

N. Agarwal, B. Bullins, E. Hazan, S. Kakade, and K. Singh. Online control with adversarial disturbances. In *International Conference on Machine Learning*, pages 111–119. PMLR, 2019.

A. Balsubramani. Sharp finite-time iterated-logarithm martingale concentration. *arXiv preprint arXiv:1405.2639*, 2014.

S. Bubeck, N. Cesa-Bianchi, and G. Lugosi. Bandits with heavy tail. *IEEE Transactions on Information Theory*, 59(11):7711–7717, 2013.

N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge University Press, 2006.

L. Chen, H. Luo, and C.-Y. Wei. Impossible tuning made possible: A new expert algorithm and its applications. In *Conference on Learning Theory*, pages 1216–1259. PMLR, 2021.

A. Cutkosky. *Algorithms and Lower Bounds for Parameter-free Online Learning*. PhD thesis, Stanford University, 2018.

A. Cutkosky. Combining online learning guarantees. In *Proceedings of the Thirty-Second Conference on Learning Theory*, pages 895–913, 2019.

A. Cutkosky and K. Boahen. Online learning without prior information. In *Conference on Learning Theory*, pages 643–677, 2017.

A. Cutkosky and H. Mehta. High-probability bounds for non-convex stochastic optimization with heavy tails. In *Proceedings of the 35st International Conference on Neural Information Processing Systems*, pages 4883–4895, 2021.

A. Cutkosky and F. Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pages 1493–1529, 2018.

J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. In *Conference on Learning Theory (COLT)*, pages 257–269, 2010a.

J. C. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *COLT*, volume 10, pages 14–26. Citeseer, 2010b.

D. J. Foster, S. Kale, M. Mohri, and K. Sridharan. Parameter-free online learning via model selection. In *Advances in Neural Information Processing Systems*, pages 6020–6030, 2017.

E. Gorbunov, M. Danilova, and A. Gasnikov. Stochastic optimization with heavy-tailed noise via accelerated gradient clipping. *Advances in Neural Information Processing Systems*, 33:15042–15053, 2020.

N. J. Harvey, C. Liaw, and S. Randhawa. Simple and optimal high-probability bounds for strongly-convex stochastic gradient descent. *arXiv preprint arXiv:1909.00843*, 2019.

E. Hazan. Introduction to online convex optimization. *arXiv preprint arXiv:1909.05207*, 2019.

S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021.

K.-S. Jun and F. Orabona. Parameter-free online convex optimization with sub-exponential noise. In *Conference on Learning Theory*, pages 1802–1823. PMLR, 2019.

A. Kavis, K. Y. Levy, and V. Cevher. High probability bounds for a class of nonconvex algorithms with adagrad stepsize. *arXiv preprint arXiv:2204.02833*, 2022.

M. Kempka, W. Kotlowski, and M. K. Warmuth. Adaptive scale-invariant online algorithms for learning linear models. In *International Conference on Machine Learning*, pages 3321–3330, 2019.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

X. Li and F. Orabona. A high probability analysis of adaptive sgd with momentum. In *Workshop on Beyond First Order Methods in ML Systems at ICML'20*, 2020.

L. Madden, E. Dall'Anese, and S. Becker. High probability convergence bounds for stochastic gradient descent assuming the polyak-lojasiewicz inequality. *arXiv preprint arXiv:2006.05610*, 2020.

Z. Mhammedi and W. M. Koolen. Lipschitz and comparator-norm adaptivity in online learning. *Conference on Learning Theory*, pages 2858–2887, 2020.

F. Orabona. A modern introduction to online learning. *arXiv preprint arXiv:1912.13213*, 2019.

F. Orabona and K.-S. Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *arXiv preprint arXiv:2110.14099*, 2021.

F. Orabona and D. Pál. Coin betting and parameter-free online learning. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29*, pages 577–585. Curran Associates, Inc., 2016.

S. J. Reddi, S. Kale, and S. Kumar. On the convergence of adam and beyond. In *International Conference on Learning Representations*, 2018.

S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

J. Tropp. Freedman's inequality for matrix martingales. *Electronic Communications in Probability*, 16:262–270, 2011.

D. van der Hoeven. User-specified local differential privacy in unconstrained adaptive online learning. In *Advances in Neural Information Processing Systems*, pages 14103–14112, 2019.

J. Ville. Etude critique de la notion de collectif, gauthier-villars, paris, 1939. *Monographies des Probabilités. Calcul des Probabilités et ses Applications*, 1939.

V. Vovk. Hoeffding's inequality in game-theoretic probability. *arXiv preprint arXiv:0708.2502*, 2007.

N. M. Vural, L. Yu, K. Balasubramanian, S. Volgushev, and M. A. Erdogdu. Mirror descent strikes again: Optimal stochastic convex optimization under infinite noise variance. *arXiv preprint arXiv:2202.11632*, 2022.

I. Waudby-Smith and A. Ramdas. Estimating means of bounded random variables by betting. *arXiv preprint arXiv:2010.09686*, 2020.

J. Zhang, S. P. Karimireddy, A. Veit, S. Kim, S. Reddi, S. Kumar, and S. Sra. Why are adaptive methods good for attention models? *Advances in Neural Information Processing Systems*, 33: 15383–15393, 2020.

P. Zhou, J. Feng, C. Ma, C. Xiong, S. C. H. Hoi, et al. Towards theoretically understanding why sgd generalizes better than adam in deep learning. *Advances in Neural Information Processing Systems*, 33:21285–21296, 2020.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 928–936, 2003.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section **??**.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] This work considers a purely mathematical investigation of concentration of regret, and we do not envision any negative social impact.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]

(c) Did you include any new assets either in the supplemental material or as a URL? [N/A]

(d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...

   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]

   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]

   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Optimistic Online Learning for Predictable Regularizer

Algorithm 1 provides output w_t by solving $w_t = x_t - y_t \nabla \psi_t(w_t)$, where $x_t \in \mathbb{R}, y_t \geq 0$ are output from sub-algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$, $\psi_t(w)$ is defined in equation (6). Under the constants for $\psi_t(w)$ defined in Theorem 3, the following Lemma shows the existence of solution.

Lemma 6 (Existence of Solution)**.** *for $x_t \in \mathbb{R}, y_t \geq 0$,*

$$w_t = x_t - y_t \nabla \psi_t(w_t)$$

where

$$\nabla \psi_t(w) = \operatorname{sign}(w) \sum_{j=1}^{2} k_j p_j \frac{|w|^{p_j - 1}}{(|w|^{p_j} + X_j)^{1 - 1/p_j}}$$

for some $k_j, X_j > 0, p_j > 1$ and $j = 1, 2$. Then w_t lies in the interval of $\left(x_t - y_t \sum_{j=1}^{2} k_j p_j, x_t\right]$ when $x_t \geq 0$, and in the interval of $\left[x_t, x_t + y_t \sum_{j=1}^{2} k_j p_j\right)$. Further,

$$h(w) = w - x_t + y_t \nabla \psi_t(w)$$

is monotonic in w.

Proof. We suppose that $x_t \geq 0$. The case $x_t < 0$ is entirely identical.

case (a): consider $y_t \neq 0$,

$$w_t = x_t - y_t \operatorname{sign}(w_t) \sum_{j=1}^{2} k_j p_j \frac{|w_t|^{p_j - 1}}{(|w_t|^{p_j} + X_j)^{(p_j - 1)/p_j}}$$

rearrange

$$\frac{x_t - w_t}{y_t} = \operatorname{sign}(w_t) \sum_{j=1}^{2} k_j p_j \frac{|w_t|^{p_j - 1}}{(|w_t|^{p_j} + X_j)^{(p_j - 1)/p_j}}$$

Let $f(w_t), g(w_t)$ to be the left and right handside of the last expression. Both functions are continuous in w_t for under assumption of x_t, y_t, k_j, p_j, X_j for $j = 1$ and 2. When $w_t^* = x_t$:

$$f(w_t^*) - g(w_t^*) = 0 - \sum_{j=1}^{2} k_j p_j \frac{|w_t^*|^{p_j - 1}}{(|w_t^*|^{p_j} + X_j)^{(p_j - 1)/p_j}} \leq 0$$

When $w_t^* = x_t - y_t \sum_{j=1}^{2} k_j p_j$:

$$f(w_t^*) - g(w_t^*) = \sum_{j=1}^{2} k_j p_j - \operatorname{sign}\left(x_t - y_t \sum_{j=1}^{2} k_j p_j\right) \sum_{j=1}^{2} k_j p_j \frac{|w_t^*|^{p_j - 1}}{(|w_t^*|^{p_j} + X_j)^{(p_j - 1)/p_j}} > 0$$

By intermediate value Theorem $f(w_t) = g(w_t)$ at w_t in between x_t and $x_t - y_t \sum_{j=1}^{2} k_j p_j$.

case (b): when $y_t = 0$, $w_t = x_t$.

Finally, by inspection the derivative of $h(w)$ with respective to w is always positive, hence is monotonic in w so that we can numerically solve for $h(w_t^*) = 0$ via binary search. □

Algorithm 1 requires the base algorithms $\mathcal{A}_1$ and $\mathcal{A}_2$ to satisfy a "second-order" regret bound, such as provided by Algorithm 1 of Cutkosky and Orabona [2018]. We assume the base algorithms are designed to handle only $1-$Lipschitz losses, so the following Lemma provide a simple linear transformation that allows the base algorithm to cope with any Lipschitz constant.

Lemma 7 (Algorithm Transformation)**.** *Suppose an algorithm $\mathcal{A}$ obtains regret $\sum_{t=1}^T \langle g_t, w_t - u\rangle \le \epsilon + R_T(u)$ for some function R_T for any sequence $\{g_t\}$ such that $|g_t| \le G$. Then, given some $\bar{\epsilon} > 0$, consider the algorithm that plays $\bar{w}_t = \frac{\bar{\epsilon}G}{\epsilon\bar{G}} w_t$ in response to subgradients $\{\bar{g}_t\}$ with $|\bar{g}_t| \le \bar{G}$, where w_t is the output of $\mathcal{A}$ on the sequence $\{g_t\}$ with $g_t = \frac{G}{\bar{G}}\bar{g}_t$. This procedure ensures regret:*

$$\sum_{t=1}^T \langle \bar{g}_t, \bar{w}_t - u\rangle \le \bar{\epsilon} + \frac{\bar{\epsilon}}{\epsilon} R_T\left(\frac{\epsilon\bar{G}}{\bar{\epsilon}G} u\right)$$

Proof. Since $|g_t| \le G$ by construction, we have:

$$\begin{aligned}\sum_{t=1}^T \langle \bar{g}_t, \bar{w}_t - u\rangle &= \frac{\bar{G}}{G}\sum_{t=1}^T \left\langle g_t, \frac{\bar{\epsilon}G}{\epsilon\bar{G}} w_t - u\right\rangle \\ &= \frac{\bar{\epsilon}}{\epsilon}\sum_{t=1}^T \left\langle g_t, w_t - \frac{\epsilon\bar{G}}{\bar{\epsilon}G} u\right\rangle \\ &\le \bar{\epsilon} + \frac{\bar{\epsilon}}{\epsilon} R_T\left(\frac{\epsilon\bar{G}}{\bar{\epsilon}G} u\right)\end{aligned}$$

□

Intuitively, if we instantiate this Lemma with an algorithm obtaining $R_T(u) = \epsilon + |u|\sqrt{T\log(|u|T/\epsilon)} + |u|\log(|u|T/\epsilon)$ for 1-Lipschitz losses, we can obtain for any ϵ, an algorithm for G-Lipschitz losses with regret $\epsilon + |u|G\sqrt{T\log(|u|GT/\epsilon)} + |u|G\log(|u|GT/\epsilon)$.

We are now at the stage to prove Theorem 2. We restate the Theorem for reference, followed by its proof.

Theorem 2. *Suppose $\mathcal{A}_1$ ensure that given some $\epsilon > 0$ and a sequence c_t with $|c_t| \le 1$:*

$$\sum_{t=1}^T \langle c_t, w_t - u\rangle \le \epsilon + A|u|\sqrt{\sum_{t=1}^T |c_t|^2\left(1 + \log\left(\frac{|u|^2 T^C}{\epsilon^2} + 1\right)\right)} + B|u|\log\left(\frac{|u|T^C}{\epsilon} + 1\right)$$

for all u for some positive constants A, B, C, and that $\mathcal{A}_2$ obtains the same guarantee for all $u \ge 0$, then for $|g_t| \le b$, $|\nabla\psi_t(w_t)| \le H$, we have the following guarantee from Algorithm 1,

$$R_T^{\mathcal{A}}(u) \le O\left[\epsilon + |u|\left(\sqrt{\max\left(0, \sum_{t=1}^T |g_t|^2 - |\nabla\psi_t(w_t)|^2\right)} + (b+H)\log T\right)\right]$$

Proof. The proof is similar to optimistic reduction in Cutkosky [2019], which combines regret guarantees from two online learning algorithms. First, we observe that $\mathcal{A}_1$ outputs x'_t from line 3 and receives gradients $(g_t + \nabla\psi_t(w_t))/(b+H) \le 1$ from line 9 in Algorithm 1. Hence we apply Lemma 7 by choosing $\epsilon = \bar{\epsilon}$, and set $G = 1$, $\bar{G} = b + H$, we have the following holds for any u,

$$\begin{aligned}R_T^1(u) &= \sum_{t=1}^T \langle g_t + \nabla\psi_t(w_t), x_t - u\rangle \\ &\le \epsilon + A|u|\sqrt{\sum_{t=1}^T |g_t + \nabla\psi_t(w_t)|^2\left[1 + \log\left(\frac{(b+H)^2|u|^2 T^C}{\epsilon^2} + 1\right)\right]} \\ &\quad + B(b+H)|u|\log\left(\frac{(b+H)|u|T^C}{\epsilon} + 1\right)\end{aligned}$$

Similarly for $\mathcal{A}_2$ outputs y'_t and receives $\frac{-\langle g_t + \nabla\psi_t(w_t), \nabla\psi_t(w_t)\rangle}{H(b+H)}$, Hence use Lemma 7 by setting $\epsilon = \bar{\epsilon}$, $G = 1$ and $\bar{G} = H(b+H)$, we have the following for all $y_\star$:

$$R_T^2(y_\star) = \sum_{t=1}^T -\langle g_t + \nabla\psi_t(w_t), \nabla\psi_t(w_t)\rangle(y_t - y_\star)$$

$$\leq \epsilon + A|y_\star|\sqrt{\sum_{t=1}^{T}\langle g_t + \nabla\psi_t(w_t), \nabla\psi_t(w_t)\rangle^2\left[1+\log\left(\frac{(b+H)^2H^2|y_\star|^2T^C}{\epsilon^2}+1\right)\right]}$$
$$+ B(b+H)H|y_\star|\log\left(\frac{(b+H)H|y_\star|T^C}{\epsilon}+1\right)$$

The relationship between the $R_T^{\mathcal{A}}(u)$ bounded by linearized loss and $R_T^1(u), R_T^2(y_\star)$ is revealed:

$$R_T^{\mathcal{A}}(u) \leq \sum_{t=1}^{T}\langle g_t + \nabla\psi_t(w_t), w_t - u\rangle$$
$$= \sum_{t=1}^{T}\langle g_t + \nabla\psi_t(w_t), x_t - u\rangle - y_t\langle g_t + \nabla\psi_t(w_t), \nabla\psi_t(w_t)\rangle$$
$$\leq \inf_{y_\star\geq 0} R_T^1(u) + R_T^2(y_\star) - y_\star\sum_{t=1}^{T}\langle g_t + \nabla\psi_t(w_t), \nabla\psi_t(w_t)\rangle$$

use identity $-2\langle a, b\rangle = \|a-b\|^2 - \|a\|^2 - \|b\|^2$

$$= \inf_{y_\star\geq 0} R_T^1(u) + R_T^2(y_\star) + \frac{y_\star}{2}\sum_{t=1}^{T}|g_t|^2 - |g_t + \nabla\psi_t(w_t)|^2 - |\nabla\psi_t(w_t)|^2$$
$$\leq \inf_{y_\star\geq 0} 2\epsilon + A|u|\sqrt{\sum_{t=1}^{T}|g_t + \nabla\psi_t(w_t)|^2\left[1+\log\left(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1\right)\right]}$$
$$+ B(b+H)|u|\log\left[\frac{(b+H)|u|T^C}{\epsilon}+1\right] + B(b+H)H|y_\star|\log\left[\frac{(b+H)H|y_\star|T^C}{\epsilon}+1\right]$$
$$+ A|y_\star|\sqrt{\sum_{t=1}^{T}\langle g_t + \nabla\psi_t(w_t), \nabla\psi_t(w_t)\rangle^2\left[1+\log\left(\frac{(b+H)^2H^2|y_\star|^2T^C}{\epsilon^2}+1\right)\right]}$$
$$+ \frac{y_\star}{2}\sum_{t=1}^{T}|g_t|^2 - |g_t + \nabla\psi_t(w_t)|^2 - |\nabla\psi_t(w_t)|^2$$

let $X = \sum_{t=1}^{T}|g_t + \nabla\psi_t(w_t)|^2$

$$\leq \inf_{y_\star\geq 0}\sup_{X\geq 0} 2\epsilon + A|u|\sqrt{X\left[1+\log\left(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1\right)\right]}$$
$$+ B(b+H)|u|\log\left[\frac{(b+H)|u|T^C}{\epsilon}+1\right] + B(b+H)H|y_\star|\log\left[\frac{(b+H)H|y_\star|T^C}{\epsilon}+1\right]$$
$$+ A|y_\star|\sqrt{XH^2\left[1+\log\left(\frac{(b+H)^2H^2|y_\star|^2T^C}{\epsilon^2}+1\right)\right]}$$
$$+ \frac{y_\star}{2}\sum_{t=1}^{T}\left(|g_t|^2 - |\nabla\psi_t(w_t)|^2\right) - \frac{y_\star}{2}X$$
$$\leq \inf_{y_\star\geq 0}\sup_{X\geq 0}\sup_{Z\geq 0} 2\epsilon + A|u|\sqrt{X\left[1+\log\left(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1\right)\right]}$$
$$+ B(b+H)|u|\log\left[\frac{(b+H)|u|T^C}{\epsilon}+1\right] + B(b+H)H|y_\star|\log\left[\frac{(b+H)H|y_\star|T^C}{\epsilon}+1\right]$$
$$+ A|y_\star|\sqrt{ZH^2\left[1+\log\left(\frac{(b+H)^2H^2|y_\star|^2T^C}{\epsilon^2}+1\right)\right]}$$

$$+\frac{y_\star}{2}\sum_{t=1}^{T}\left(|g_t|^2-|\nabla\psi_t(w_t)|^2\right)-\frac{y_\star}{4}(X+Z)$$

set

$$y_\star=\min\left(\frac{2A|u|\sqrt{1+\log((b+H)^2|u|^2T^C/\epsilon^2+1)}}{\sqrt{\max(0,\sum_{t=1}^{T}\left(|g_t|^2-|\nabla\psi_t(w_t)|^2\right))}},\frac{|u|}{H}\right)$$

$$\begin{aligned}
&\le\sup_{X\ge0}\sup_{Z\ge0}2\epsilon+A|u|\sqrt{X\left[1+\log\left(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1\right)\right]}\\
&+B(b+H)|u|\log\left[\frac{(b+H)|u|T^C}{\epsilon}+1\right]-\frac{y_\star}{4}(X+Z)\\
&+B(b+H)|u|\log\left[\frac{(b+H)|u|T^C}{\epsilon}+1\right]\\
&+A|y_\star|\sqrt{ZH^2\left[1+\log\left(\frac{(b+H)^2H^2|y_\star|^2T^C}{\epsilon^2}+1\right)\right]}\\
&+A|u|\sqrt{1+\log(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1)}\sqrt{\max(0,\sum_{t=1}^{T}\left(|g_t|^2-|\nabla\psi_t(w_t)|^2\right))}
\end{aligned}$$

For $a,b>0$, $\sup_x a\sqrt{x}-bx=a^2/4b$, apply the identity to both $\sup_{X>0},\sup_{Z>0}$

$$\begin{aligned}
&\le 2\epsilon+A^2|u|^2\left[1+\log\left(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1\right)\right]/y_\star\\
&+A^2|y_\star|H^2\left[1+\log\left(\frac{(b+H)^2H^2|y_\star|^2T^C}{\epsilon^2}+1\right)\right]\\
&+2B(b+H)|u|\log\left[\frac{(b+H)|u|T^C}{\epsilon}+1\right]\\
&+A|u|\sqrt{1+\log(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1)}\sqrt{\max(0,\sum_{t=1}^{T}\left(|g_t|^2-|\nabla\psi_t(w_t)|^2\right))}
\end{aligned}$$

substitute $y_\star$

$$\begin{aligned}
&\le 2\epsilon+\frac{A}{2}|u|\sqrt{\left[1+\log\left(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1\right)\right]}\sqrt{\max(0,\sum_{t=1}^{T}\left(|g_t|^2-|\nabla\psi_t(w_t)|^2\right))}\\
&+A^2|u|H\left[1+\log\left(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1\right)\right]+2B(b+H)|u|\log\left[\frac{(b+H)|u|T^C}{\epsilon}+1\right]\\
&+A|u|\sqrt{1+\log(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1)}\sqrt{\max(0,\sum_{t=1}^{T}\left(|g_t|^2-|\nabla\psi_t(w_t)|^2\right))}\\
&\le 2\epsilon+\frac{3A}{2}|u|\sqrt{\left[1+\log\left(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1\right)\right]}\sqrt{\max(0,\sum_{t=1}^{T}\left(|g_t|^2-|\nabla\psi_t(w_t)|^2\right))}\\
&+|u|\left(A^2H+2B(b+H)\right)\left[1+\log\left(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1\right)\right]
\end{aligned}$$

Define a constant N

$$N=1+\log\left(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1\right)$$

Then, $R_T^{\mathcal{A}}(u)$ can be written as,

$$R_T^{\mathcal{A}}(u) \le \sum_{t=1}^{T} \langle g_t + \nabla\psi_t(w_t), w_t - u\rangle$$
$$\le 2\epsilon + |u| \left[\frac{3A}{2} \sqrt{N \max\left(0, \sum_{t=1}^{T} |g_t|^2 - |\nabla\psi_t(w_t)|^2\right)} + \left(A^2 H + 2B(b+H)\right) N \right]$$

□

The following Lemma shows the magnitude of w_t as a function of t, where $\{w_t\}$ is a sequence of output from algorithm 1. We shall see later on that a coarse bound for w_t helps to proof Theorem 3.

Lemma 8 (Exponential Growing Output)**.** *Suppose $\mathcal{A}$ is an arbitrary OLO algorithm that guarantees regret $R_T^{\mathcal{A}}(0) = \sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{w}_t\rangle \le \epsilon$ for all sequence $\{\mathbf{g}_t\}$ with $\|\mathbf{g}_t\| \le G$. Then it must hold that $\|\mathbf{w}_t\| \le \frac{\epsilon}{2G} 2^t$ for all t.*

Proof. We will first prove by contradiction that $\|\mathbf{w}_t\| \le \frac{\epsilon - \sum_{i=1}^{t-1}\langle \mathbf{g}_i, \mathbf{w}_i\rangle}{G}$ for all t for all sequences $\{\mathbf{g}_t\}$. Suppose that there is some t and sequence $\mathbf{g}_1, \dots, \mathbf{g}_{t-1}$ such that $\|\mathbf{w}_t\| > \frac{\epsilon - \sum_{i=1}^{t-1}\langle \mathbf{g}_i, \mathbf{w}_i\rangle}{G}$. Then, consider $\mathbf{g}_t = G \frac{\mathbf{w}_t}{\|\mathbf{w}_t\|}$. Then we have:

$$R_t(0) = \sum_{i=1}^{t-1} \langle \mathbf{g}_i, \mathbf{w}_i\rangle + \langle \mathbf{g}_t, \mathbf{w}_t\rangle > \epsilon$$

which is a contradiction, and so $\|\mathbf{w}_t\| \le \frac{\epsilon - \sum_{i=1}^{t-1}\langle \mathbf{g}_i, \mathbf{w}_i\rangle}{G}$.

Now, if we define $H_t = \epsilon - \sum_{i=1}^{t} \langle \mathbf{g}_i, \mathbf{w}_i\rangle$, we have $\|\mathbf{w}_t\| \le \frac{H_{t-1}}{G}$. Therefore:

$$\begin{aligned} H_t &= H_{t-1} - \langle \mathbf{g}_t, \mathbf{w}_t\rangle \\ &\le 2H_{t-1} \\ &\le 2^t H_0 \\ &= \epsilon 2^t \end{aligned}$$

Thus, we have $\|\mathbf{w}_t\| \le \frac{H_{t-1}}{G} = \frac{\epsilon}{2G} 2^t$ as desired. □

B Cancellation for Gradients with Sub-exponential Noise

In this Section, we ultimately provide the proof for Theorem 3. We first show a few algebraic lemma followed by the property of the regularizer, Then we show the proof for Theorem 3 by combining different lemma with the outlines listed in Section 4.

Lemma 9. *For $x \ge 0, a > 0$, $p \ge 1$*

$$\frac{a}{(x+a)^{1-\frac{1}{p}}} \ge (x+a)^{\frac{1}{p}} - x^{\frac{1}{p}}$$

Proof.

$$x^{\frac{1}{p}} (x+a)^{1-\frac{1}{p}} \ge x^{\frac{1}{p}} x^{1-\frac{1}{p}} = x$$

rearrange

$$\begin{aligned} 0 &\ge x - x^{\frac{1}{p}}(x+a)^{1-\frac{1}{p}} \\ a &\ge (x+a) - x^{\frac{1}{p}}(x+a)^{1-\frac{1}{p}} \end{aligned}$$

divide both side by $(x+a)^{1-\frac{1}{p}}$, we complete the proof □

Lemma 10. *For $x, a \geq 0, p \geq 1$:*

$$(x+a)^{1/p} \leq x^{1/p} + a^{1/p}$$

Proof.

$$x + a = (x^{1/p})^p + (a^{1/p})^p \leq (x^{1/p} + a^{1/p})^p$$

raise to the power of $1/p$ to complete the proof □

Lemma 11. *For $\mathbf{x} \in \mathbb{R}^d$, if $\|\cdot\|_p$ is the $p-$norm:*

$$\frac{1}{d^{1/p}}\|\mathbf{x}\|_p \leq \|\mathbf{x}\|_\infty \leq \|\mathbf{x}\|_p$$

Proof. Clearly, it suffices to consider $\mathbf{x} = (x_1, \ldots, x_d)$ with $x_i \geq 0$ for all i. let $i_* = \text{argmax}_i\, x_i$, then for all $a \geq 0$

$$\|\mathbf{x}\|_\infty = x_i^* = (x_i^{*p})^{1/p} \leq (x_i^{*p} + a)^{1/p}$$

setting $a = \sum_{i \neq i^*} x_i^p$, demonstrates the upper bound.

For the lower bound:

$$\frac{1}{d^{1/p}}\|\mathbf{x}\|_p = \frac{1}{d^{1/p}}\left(\sum_{i=1}^d x_i^p\right)^{1/p} \leq \frac{1}{d^{1/p}}\left(d x_i^{*p}\right)^{1/p} = x_i^*$$

□

Lemma 12. *For $x > 0$,*

$$\log(x + \exp(1)) \leq \max(0, \log(x)) + \exp(1)$$

Proof. For $x \in (0,1]$, $\log(x) \leq 0$. Thus, the inequality holds since $\log(x + \exp(1)) \leq \log(1 + \exp(1)) \leq \exp(1)$.

For $x > 1$, we have $\log(x) > 0$. Let

$$h(x) = \log(x + \exp(1))$$
$$f(x) = \log(x) + \exp(1)$$

Taking derivatives,

$$h'(x) = 1/(x + \exp(1))$$
$$f'(x) = 1/x$$

Thus, $f'(x) > h'(x)$ for $x > 1$. Now, since $f(1) \geq h(1)$, we have $f(x) > h(x)$ for $x > 1$.

Combining both case we complete the proof. □

Lemma 13 (Cumulative Huber Loss)**.** *Consider r_t as in equation (5) (copied below for convience):*

$$r_t(w; c, p, \alpha_0) = \begin{cases} c\left(p|w| - (p-1)|w_t|\right)\frac{|w_t|^{p-1}}{(\sum_{i=1}^t |w_i|^p + \alpha_0^p)^{1-1/p}}, & |w| > |w_t| \\ c|w|^p \frac{1}{(\sum_{i=1}^t |w_i|^p + \alpha_0^p)^{1-1/p}}, & |w| \leq |w_t| \end{cases}$$

Then, with fixed parameter $c, \alpha_0 > 0$, $p \geq 1$,

$$\sum_{t=1}^T r_t(w_t) \geq c\left(\left(\sum_{t=1}^T |w_t|^p + \alpha_0^p\right)^{1/p} - \alpha_0\right) \tag{10}$$

$$\sum_{t=1}^T r_t(u) \leq cp|u|T^{1/p}\left[\left(\log\frac{T|u|^p + \alpha_0^p}{\alpha_0^p}\right)^{(p-1)/p} + 1\right] \tag{11}$$

Proof. Define index set $A_T = \{t : |w_t| < |u|, t = 1, \cdots, T\}$, and let $n(A_T)$ be the cardinality of A_T. Let $S_t = \alpha_0^p + \sum_{i=1}^t |w_i|^p$. First, we show lower bound for $\sum_{t=1}^T r_t(w)$. Since $p \geq 1$,

$$\sum_{t=1}^T r_t(w_t) = c\sum_{t=1}^T \frac{|w_t|^p}{(\sum_{i=1}^t |w_i|^p + \alpha_0^p)^{1-1/p}}$$
$$= c\sum_{t=1}^T \frac{|w_t|^p}{(|w_t|^p + S_{t-1})^{1-1/p}}$$

use Lemma 9, set $a = |w_t|^p, x = S_{t-1}, a + x = S_t$

$$\geq c\sum_{t=1}^T (S_t^{1/p} - S_{t-1}^{1/p})$$
$$= c(S_T^{1/p} - \alpha_0)$$
$$= c\left((\sum_{t=1}^T |w_t|^p + \alpha_0^p)^{1/p} - \alpha_0 \right)$$

Now, we upper bound of $\sum_t r_t(u)$. We partition the sum into two terms, and bound them individually:

$$\sum_{t=1}^T r_t(u) \leq cp|u| \underbrace{\sum_{\substack{t\leq T\\ |w_t|\leq|u|}} \frac{|w_t|^{p-1}}{(\sum_{i=1}^t |w_i|^p + \alpha_0^p)^{1-1/p}}}_{A} + c|u|^p \underbrace{\sum_{\substack{t\leq T\\ |w_t|>|u|}} \frac{1}{(\sum_{i=1}^t |w_i|^p + \alpha_0^p)^{1-1/p}}}_{B}$$

First, we bound A:

$$A \leq \sum_{\substack{t\leq T\\ |w_t|\leq|u|}} \frac{|w_t|^{p-1}}{\left(\sum_{\substack{i\leq t,\\ |w_i|\leq|u|}} |w_i|^p + \alpha_0^p\right)^{1-1/p}}$$

by Holder's inequality $\langle a, b\rangle \leq \|a\|_m \|b\|_n$, where $\frac{1}{m} + \frac{1}{n} = 1$. Set $m = p, n = \frac{p}{p-1}$.

$$\leq n(A_T)^{1/p} \left(\sum_{\substack{t\leq T\\ |w_t|\leq|u|}} \frac{|w_t|^p}{\sum_{\substack{i\leq t,\\ |w_i|\leq|u|}} |w_i|^p + \alpha_0^p} \right)^{(p-1)/p}$$
$$\leq n(A_T)^{1/p} \left(\int_{\alpha_0^p}^{\alpha_0^p + \sum_{t\in A_T} |w_t|^p} \frac{1}{x} dx \right)^{(p-1)/p}$$
$$= n(A_T)^{1/p} \left(\log \frac{\sum_{t\in A_T} |w_t|^p + \alpha_0^p}{\alpha_0^p} \right)^{(p-1)/p}$$
$$\leq n(A_T)^{1/p} \left(\log \frac{n(A_T)|u|^p + \alpha_0^p}{\alpha_0^p} \right)^{(p-1)/p}$$
$$\leq T^{1/p} \left(\log \frac{|u|^p T + \alpha_0^p}{\alpha_0^p} \right)^{(p-1)/p}$$

Now, we bound B:

$$B \leq \sum_{\substack{t\leq T\\ |w_t|>|u|}} \frac{1}{(\sum_{\substack{i\leq t,\\ |w_i|>|u|}} |w_i|^p + \alpha_0^p)^{1-1/p}}$$

$$
\begin{aligned}
&\le \sum_{\substack{t\le T\\ |w_t|>|u|}} \frac{1}{[n(\{i\le t: |w_i|>|u|\})|u|^p+\alpha_0^p]^{1-1/p}}\\
&= \frac{1}{|u|^{p-1}} \sum_{\substack{t\le T\\ |w_t|>|u|}} \frac{1}{\left[n(\{i\le t: |w_i|>|u|\})+(\frac{\alpha_0}{|u|})^p\right]^{1-1/p}}\\
&\le \frac{1}{|u|^{p-1}} \int_{(\alpha_0/|u|)^p}^{(\alpha_0/|u|)^p+(T-n(A_T))} \frac{1}{x^{1-1/p}}dx\\
&= \frac{p}{|u|^{p-1}}\left[(T-n(A_T))+(\alpha_0/|u|)^p)^{1/p}-(\alpha_0/|u|)\right]
\end{aligned}
$$

by Lemma 10, set $x=T-n(A_T), a=(\alpha_0/|u|)^p$

$$
\begin{aligned}
&\le \frac{p}{|u|^{p-1}}(T-n(A_T))^{1/p}\\
&\le \frac{p}{|u|^{p-1}}T^{1/p}
\end{aligned}
$$

Combining A and B:

$$
\sum_{t=1}^T r_t(u) \le cp|u|T^{1/p}\left[\left(\log\frac{|u|^pT+\alpha_0^p}{\alpha_0^p}\right)^{(p-1)/p}+1\right]
$$

□

Lemma 14. *Consider the regularization function $\psi_t(w)$ defined in equation (6) with parameter $c_1, p_1, \alpha_1, c_2, p_2, \alpha_2$, displayed below for reference:*

$$
\psi_t(w) = r_t(w; c_1, p_1, \alpha_1) + r_t(w; c_2, p_2, \alpha_2)
$$

When we set

$$
\begin{aligned}
c_1 &= 2\sigma\sqrt{\log\left(\tfrac{32}{\delta}\left[\log\left(2^{T+1}\right)+2\right]^2\right)}, \quad & c_2 &= 32b\log\left(\tfrac{224}{\delta}\left[\log\left(1+\tfrac{b}{\sigma}2^{T+2}\right)+2\right]^2\right),\\
p_1 &= 2, \qquad p_2=\log T, & \alpha_1 &= \epsilon/c_1, \qquad \alpha_2=\epsilon\sigma/(4b(b+H))
\end{aligned}
$$

as defined Theorem 3, Then

$$
\sum_{t=1}^T \psi_t(w_t) \ge c_1\sqrt{\sum_{t=1}^T |w_t|^2+\alpha_1^2} + c_2\max\left(\alpha_2, \max_{t\in\{1,\cdots,T\}}|w_t|\right) - \epsilon\left(1+\frac{c_2\sigma}{4b(b+H)}\right)
$$

$$
\sum_{t=1}^T \psi_t(u) \le 2c_1|u|\sqrt{T}\left[\sqrt{\log\left(\frac{T|u|^2c_1^2}{\epsilon^2}+1\right)}+1\right] + 3c_2p_2|u|\left(\max\left(0, p_2\left(\log\frac{|u|}{\alpha_2}+1\right)\right)+4\right)
$$

Proof. The proof builds on Lemma 13. We show the algebra for r_t with fixed tuple of parameters (c_i, p_i, α_i) for $i=1,2$, respectively. The main difference is due to the value of p_i.

For $i=1$:

$$
\begin{aligned}
\sum_{t=1}^T r_t(w_t; c_1, p_1, \alpha_1) &\ge c_1\left(\sqrt{\sum_{t=1}^T |w_t|^2+\alpha_1^2}-\alpha_1\right)\\
&= c_1\sqrt{\sum_{t=1}^T |w_t|^2+\alpha_1^2}-\epsilon
\end{aligned}
$$

By equation (11)

$$
\sum_{t=1}^T r_t(u; c_1, p_1, \alpha_1) \le 2c_1|u|\sqrt{T}\left[\sqrt{\log\frac{T|u|^2+\alpha_1^2}{\alpha_1^2}}+1\right]
$$

$$= 2c_1|u|\sqrt{T}\left[\sqrt{\log\left(\frac{T|u|^2c_1^2}{\epsilon^2}+1\right)}+1\right]$$

For $i = 2$: by equation (10) and Lemma 11:

$$\sum_{t=1}^{T} r_t(w_t; c_2, p_2, \alpha_2) \geq c_2\left(\max\left(\alpha_2, \max_{t\in\{1,\cdots,T\}}|w_t|\right) - \alpha_2\right)$$
$$= c_2 \max\left(\alpha_2, \max_{t\in\{1,\cdots,T\}}|w_t|\right) - \frac{\epsilon c_2 \sigma}{4b(b+H)}$$

By equation (11)

$$\sum_{t=1}^{T} r_t(w_t; c_2, p_2, \alpha_2) \leq c_2p_2|u|\exp(1)\left[\left(\log\frac{T|u|^{p_2}+\alpha_2^{p_2}}{\alpha_2^{p_2}}\right)^{(p_2-1)/p_2}+1\right]$$
$$\leq c_2p_2|u|\exp(1)\left[\left(\log\frac{T|u|^{p_2}+\exp(1)\alpha_2^{p_2}}{\alpha_2^{p_2}}\right)^{(p_2-1)/p_2}+1\right]$$
$$\leq c_2p_2|u|\exp(1)\left[\left(\log\frac{T|u|^{p_2}+\exp(1)\alpha_2^{p_2}}{\alpha_2^{p_2}}\right)+1\right]$$
$$= c_2p_2|u|\exp(1)\left[\log\left(T\frac{|u|^{p_2}}{\alpha_2^{p_2}}+\exp(1)\right)+1\right]$$

Since $T\frac{|u|^{p_2}}{\alpha_2^{p_2}} > 0$, invoke Lemma 12 by substituting $x = T\frac{|u|^{p_2}}{\alpha_2^{p_2}}$

$$\leq c_2p_2|u|\exp(1)\left(\max\left(0, \log\left(\frac{T|u|^{p_2}}{\alpha_{2_{p_2}}}\right)\right)+\exp(1)+1\right)$$
$$= c_2p_2|u|\exp(1)\left(\max\left(0, p_2\log\left(\frac{\exp(1)|u|}{\alpha_2}\right)\right)+\exp(1)+1\right)$$
$$= c_2p_2|u|\exp(1)\left(\max\left(0, p_2\left(\log\frac{|u|}{\alpha_2}+1\right)\right)+\exp(1)+1\right)$$
$$\leq 3c_2p_2|u|\left(\max\left(0, p_2\left(\log\frac{|u|}{\alpha_2}+1\right)\right)+4\right)$$

Combining both cases of $i = 1, 2$, we complete the proof. □

Now we are at the stage to prove Theorem 3. We restate the Theorem for reference, followed by the proof.

Theorem 3. *Suppose $\{g_t\}$ are stochastic subgradients such that $\mathbb{E}[g_t] \in \partial\ell_t(w_t)$, $|g_t| \leq b$ and $\mathbb{E}[g_t^2|w_t] \leq \sigma^2$ almost surely for all t. Set the following constants for $\psi_t(w)$ shown in equation (6) for any $0 < \delta \leq 1$, $\epsilon > 0$,*

$$c_1 = 2\sigma\sqrt{\log\left(\tfrac{32}{\delta}\left[\log\left(2^{T+1}\right)+2\right]^2\right)}, \quad c_2 = 32b\log\left(\tfrac{224}{\delta}\left[\log\left(1+\tfrac{b}{\sigma}2^{T+2}\right)+2\right]^2\right),$$
$$p_1 = 2, \qquad p_2 = \log T, \qquad \alpha_1 = \epsilon/c_1, \qquad \alpha_2 = \epsilon\sigma/(4b(b+H))$$

where $H = c_1p_1 + c_2p_2$, $|\nabla\psi_t(w_t)| \leq H$. Then, with probability at least $1-\delta$, algorithm 1 guarantees

$$R_T(u) \leq \tilde{O}\left[\epsilon\log\frac{1}{\delta} + |u|b\log\frac{1}{\delta} + |u|\sigma\sqrt{T\log\frac{1}{\delta}}\right]$$

Proof. The proof is a composition of concentration bounds and our Lemmas for the regularizers, following the outline in Section 4. Previously, we defined $\epsilon_t = \nabla\ell_t(\mathbf{w}_t) - g_t$. $|\epsilon_t| \leq 2b$ and $\mathbb{E}[\epsilon_t^2] \leq \sigma^2$.

Step 1 : We first derive a concentration bound for the NOISE term defined in equation (2). Notice that $\{|u|\epsilon_i\}$ is a martingale difference sequence. Then by Lemma 23, with probability at least $1-\frac{\delta}{4}$,

$$\left|\sum_{t=1}^{T} u\epsilon_t\right| \leq 4|u|b\log\frac{8}{\delta} + |u|\sigma\sqrt{2T\log\frac{8}{\delta}} \tag{12}$$

Now, we coarsely bound the output w_t from Algorithm 1. At each round t, w_t is updated by solving

$$\begin{aligned} w_t &= x_t - y_t\nabla\psi_t(w_t) \\ &= \frac{x_t'}{b+H} - \frac{y_t'}{H(b+H)}\nabla\psi_t(w_t) \end{aligned}$$

where x_t', y_t' are outputs from some algorithm in which the regret at the origin is bounded by some positive ϵ with Lipschitz constant 1. By Lemma 8,

$$|x_t| \leq \frac{\epsilon}{2(b+H)}2^t \qquad\qquad |y_t| \leq \frac{\epsilon}{2H(b+H)}2^t$$

Now, define $k = b+H$. Then, by triangle inequality and $|\nabla\psi_t(w_t)| \leq H$

$$|w_t| \leq |x_t| + |y_t||\nabla\psi_t(w_t)| \leq \frac{\epsilon}{b+H}2^t = \frac{\epsilon}{k}2^t \tag{13}$$

Finally, $\{w_t\epsilon_t\}$ is a martingale difference sequence that satisfies:

$$\begin{aligned} &\mathbb{E}[w_t^2\epsilon_t^2 \mid g_1,\cdots,g_t] \leq |w_t|^2\sigma^2 \\ &|w_t\epsilon_t| \leq 2|w_t|b \end{aligned}$$

where w_t depends on $g_1,\cdots,g_{t-1}$ only. Hence by Proposition 17 $\{w_t\epsilon_t\}$ is $(|w_t|\sigma, 4|w_t|b)$sub-exponential. Then we apply Theorem 18 by setting $\nu = \epsilon\sigma/k$ to obtain that with probability at least $1-\frac{\delta}{4}$

$$\begin{aligned} \left|\sum_{t=1}^{T} w_t\epsilon_t\right| \leq 2&\sqrt{\sigma^2\sum_{t=1}^{T} w_t^2\log\left(\frac{32}{\delta}\left[\log\left(\left[\frac{k}{\epsilon}\sqrt{\sum_{t=1}^{T} w_t^2}\right]_1\right)+2\right]^2\right)} \\ &+ 8\max(\epsilon\sigma/k, 4b\max_{t\leq T}|w_t|)\log\left(\frac{224}{\delta}\left[\log\left(\frac{\max(\epsilon\sigma/k, 4b\max_{t\leq T}|w_t|)}{\epsilon\sigma/k}\right)+2\right]^2\right) \end{aligned} \tag{14}$$

We now simplify the $\log\log$ term with a worst case upper bound of $|w_t|$. From equation (13), we have

$$\max i \leq t|w_i| \leq \tfrac{\epsilon}{k}2^t, \qquad \textstyle\sum_{i=1}^{t} w_i^2 \leq \frac{\epsilon^2}{k^2}\sum_{i=1}^{t} 4^i = \frac{\epsilon^2}{k^2}\frac{4(4^t-1)}{3}$$

Hence

$$\log\left(\left[\frac{k}{\epsilon}\sqrt{\sum_{i=1}^{t} w_i^2}\right]_1\right) \leq \log\left(\left[\sqrt{2\cdot 4^T}\right]_1\right) \leq \log\left(2^{T+1}\right)$$

$$\begin{aligned} \log\left(\frac{\max(\epsilon\sigma/k, 4b\max_{t\leq T}|w_t|)}{\epsilon\sigma/k}\right) &\leq \log\left(1+\frac{4bk\max_{t\leq T}|w_t|)}{\epsilon\sigma}\right) \\ &\leq \log\left(1+\frac{b}{\sigma}2^{T+2}\right) \end{aligned}$$

Notice that the double-logarithm in (14) is critical to ameliorate this exponential bound on $|w_t|$!

Substitute the above inequalities into equation (14), and combining with equation (12) by union bound, with probability at least $1-\frac{\delta}{2}$:

$$\text{NOISE} \leq 4|u|b\log\frac{8}{\delta} + |u|\sigma\sqrt{2T\log\frac{8}{\delta}} + 2\sqrt{\sigma^2\sum_{t=1}^{T} w_t^2\log\left(\frac{32}{\delta}\left[\log\left(2^{T+1}\right)+2\right]^2\right)}$$

$$+ 8\max(\epsilon\sigma/k, 4b\max_{t\le T}|w_t|)\log\left(\frac{224}{\delta}\left[\log\left(1+\frac{b}{\sigma}2^{T+2}\right)+2\right]^2\right)$$

$$= 4|u|b\log\frac{8}{\delta} + |u|\sigma\sqrt{2T\log\frac{8}{\delta}} + 2\sqrt{\sigma^2\sum_{t=1}^T w_t^2\log\left(\frac{32}{\delta}\left[\log\left(2^{T+1}\right)+2\right]^2\right)}$$

$$+ 32b\max(\epsilon\sigma/4kb, \max_{t\le T}|w_t|)\log\left(\frac{224}{\delta}\left[\log\left(1+\frac{b}{\sigma}2^{T+2}\right)+2\right]^2\right)$$

$$= 4|u|b\log\frac{8}{\delta} + |u|\sigma\sqrt{2T\log\frac{8}{\delta}} + c_1\sqrt{\sum_{t=1}^T w_t^2} + c_2\max(\epsilon\sigma/4kb, \max_{t\le T}|w_t|) \tag{15}$$

Step 2 : Next, we derive a bound on $\sum_{t=1}^T\langle g_t, w_t - u\rangle$. Our approach builds upon the motivation sketched in equation (3). We define $R_T^{\mathcal{A}}(u) = \sum_{t=1}^T \hat{\ell}_t(w_t) - \hat{\ell}_t(u)$. Notice that $R_T^{\mathcal{A}}(u)$ can then be bounded by Theorem 2. Thus, we copy over equation (3) below, and apply Theorem 2 and Lemma 14 to bound the regret

$$\sum_{t=1}^T\langle g_t, w_t - u\rangle \le R_T^{\mathcal{A}}(u) - \sum_{t=1}^T\psi_t(w_t) + \sum_{t=1}^T\psi_t(u)$$

$$\le 2\epsilon + |u|\left[\frac{3A}{2}\sqrt{N\max\left(0, \sum_{t=1}^T|g_t|^2 - |\nabla\psi_t(w_t)|^2\right)} + \left(A^2H + 2B(b+H)\right)N\right]$$

$$- c_1\sqrt{\sum_{t=1}^T|w_t|^2 + \alpha_1^2} - c_2\max\left(\alpha_2, \max_{t\in\{1,\cdots,T\}}|w_t|\right) + \epsilon\left(1 + \frac{c_2\sigma}{2b(b+H)}\right)$$

$$+ 2c_1|u|\sqrt{T}\left[\sqrt{\log\frac{T|u|^2+\alpha_1^2}{\alpha_1^2}} + 1\right] + 3c_2p_2|u|\left(\max\left(0, p_2\left(\log\frac{|u|}{\alpha_2}+1\right)\right)+4\right) \tag{16}$$

where $N = 1 + \log\left(\frac{(b+H)^2|u|^2T^C}{\epsilon^2}+1\right)$ and A, B, C are some positive constants.

Step 3 : As shown in equation (4), the regret is derived by combining equation (15) and (16). We observe that the martingale concentration from Step 1 will be cancelled by the negative regularization terms from Step 2 to complete the proof:

$$R_T(u) \le \epsilon\left(3 + \frac{8\sigma}{b+H}\log\left(\frac{224}{\delta}\left[\log\left(1+\frac{b}{\sigma}2^{T+2}\right)+2\right]^2\right)\right)$$

$$+ |u|\left[4c_1(A^2+B)N + \frac{3A}{2}\sqrt{N\max\left(0, \sum_{t=1}^T|g_t|^2 - |\nabla\psi_t(w_t)|^2\right)}\right]$$

$$+ |u|b\left[2BN + 4\log\frac{8}{\delta} + \frac{c_2\log T}{b}\left((A^2+2B)N + 3\left(\max\left(0, \log T\left(\log\frac{|u|}{\alpha_2}+1\right)\right)+4\right)\right)\right]$$

$$+ |u|\sqrt{T}\left[2c_1\left(\sqrt{\log\left(\frac{T|u|c_1^2}{\epsilon^2}+1\right)}+1\right) + \sigma\sqrt{2\log\frac{8}{\delta}}\right] \tag{17}$$

The above holds for probability at least $1 - \frac{\delta}{2}$.

Step 4: For the final statement, we must remove the random quantity $\sum_{t=1}^T g_t^2$ appearing in the bound. Fortunately, this is achievable via a relatively straightforward application of Bernstein-style bounds. In particular, by Lemma 24, with probability at least $1 - \delta/2$

$$\sum_{t=1}^{T} |g_t|^2 \le \frac{3}{2}T\sigma^2 + \frac{5}{3}b^2 \log\frac{2}{\delta}$$

Thus, further upper bound equation 17 by union bound, we have with probability at least $1 - \delta$,

$$\begin{aligned}
R_T(u) \le \epsilon &\left(3 + \frac{8\sigma}{b+H}\log\left(\frac{224}{\delta}\left[\log\left(1 + \frac{b}{\sigma}2^{T+2}\right) + 2\right]^2\right)\right) \\
&+ |u|\left[4c_1(A^2+B)N + \frac{3A}{2}\sqrt{N\left(\frac{3}{2}T\sigma^2 + \frac{5}{3}b^2\log\frac{2}{\delta}\right)}\right] \\
&+ |u|b\left[2BN + 4\log\frac{8}{\delta} + \frac{c_2\log T}{b}\left((A^2+2B)N + 3\left(\max\left(0, \log T\left(\log\frac{|u|}{\alpha_2} + 1\right)\right) + 4\right)\right)\right] \\
&+ |u|\sqrt{T}\left[2c_1\left(\sqrt{\log\left(\frac{T|u|c_1^2}{\epsilon^2} + 1\right)} + 1\right) + \sigma\sqrt{2\log\frac{8}{\delta}}\right] \qquad (18)
\end{aligned}$$

□

C Gradient Clipping for Heavy-tailed Gradients

First, we show the property of truncated heavy-tailed gradients followed by the proof of Theorem 4. These elementary facts can be found in Zhang et al. [2020], but we reproduce the proofs for completeness.

Lemma 15 (Clipped Gradient Properties)**.** *Suppose* $\mathbf{g}_t$ *is heavy-tailed random vector,* $\|\mathbb{E}[\mathbf{g}_t]\| \le G$*,* $\mathbb{E}[\|\mathbf{g}_t - \mathbb{E}[\mathbf{g}_t]\|^{\mathfrak{p}}] \le \sigma^{\mathfrak{p}}$ *for some* $\mathfrak{p} \in (1, 2]$ *and* $\sigma \le \infty$*. Define truncated gradient* $\hat{\mathbf{g}}_t$ *with a positive clipping parameter* τ*:*

$$\hat{\mathbf{g}}_t = \frac{\mathbf{g}_t}{\|\mathbf{g}_t\|}\min(\tau, \|\mathbf{g}_t\|)$$

Let $\boldsymbol{\mu} = \mathbb{E}[\mathbf{g}_t]$*. Then:*

$$\begin{aligned}
\|\mathbb{E}[\hat{\mathbf{g}}_t] - \boldsymbol{\mu}\| &\le \frac{2^{\mathfrak{p}-1}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})}{\tau^{\mathfrak{p}-1}} \\
\mathbb{E}[\|\hat{\mathbf{g}}_t\|^2] &\le 2^{\mathfrak{p}-1}\tau^{2-\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})
\end{aligned}$$

Proof. By Jensen's inequality

$$\begin{aligned}
\|\mathbb{E}[\hat{\mathbf{g}}_t] - \boldsymbol{\mu}\| &\le \mathbb{E}[\|\hat{\mathbf{g}}_t - \mathbf{g}_t\|] \\
&\le \mathbb{E}[\|\mathbf{g}_t\|\mathbf{1}[\|\mathbf{g}_t\| \ge \tau]] \\
&\le \mathbb{E}[\|\mathbf{g}_t\|^{\mathfrak{p}}/\tau^{\mathfrak{p}-1}] \\
&\le \mathbb{E}[(\|\mathbf{g}_t - \boldsymbol{\mu}\| + \|\boldsymbol{\mu}\|)^{\mathfrak{p}}/\tau^{\mathfrak{p}-1}] \\
&= \frac{2^{\mathfrak{p}}}{\tau^{\mathfrak{p}-1}}\mathbb{E}[(\frac{1}{2}\|\mathbf{g}_t - \boldsymbol{\mu}\| + \frac{1}{2}\|\boldsymbol{\mu}\|)^2] \\
&\le \frac{2^{\mathfrak{p}-1}}{\tau^{\mathfrak{p}-1}}\left(\mathbb{E}[\|\mathbf{g}_t - \boldsymbol{\mu}\|^2] + \mathbb{E}[\|\boldsymbol{\mu}\|^2]\right) \\
&\le \frac{2^{\mathfrak{p}-1}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})}{\tau^{\mathfrak{p}-1}}
\end{aligned}$$

The second last inequality was due to convexity and linearity of expectation. In term of the variance, the algebra is similar:

$$\begin{aligned}
\mathbb{E}[\|\hat{\mathbf{g}}_t\|^2] &\le \mathbb{E}[\|\mathbf{g}_t\|^{\mathfrak{p}}\tau^{2-\mathfrak{p}}] \\
&\le \mathbb{E}[(\|\mathbf{g}_t - \boldsymbol{\mu}\| + \|\boldsymbol{\mu}\|)^{\mathfrak{p}}\tau^{2-\mathfrak{p}}]
\end{aligned}$$

$$= 2^{\mathfrak{p}}\tau^{2-\mathfrak{p}}\,\mathbb{E}[(\frac{1}{2}\|\mathbf{g}_t - \boldsymbol{\mu}\| + \frac{1}{2}\|\boldsymbol{\mu}\|)^{\mathfrak{p}}]$$
$$\leq 2\left(\mathbb{E}[\|\mathbf{g}_t - \boldsymbol{\mu}\|^{\mathfrak{p}}] + \mathbb{E}[\|\boldsymbol{\mu}\|^{\mathfrak{p}}]\right)$$
$$\leq 2^{\mathfrak{p}-1}\tau^{2-\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})$$

□

We now restate Theorem 4 followed by its proof.

Theorem 4. *Suppose* $\{g_t\}$ *are heavy-tailed stochastic gradient such that* $\mathbb{E}[g_t] \in \partial \ell_t(w_t)$, $|\mathbb{E}[g_t]| \leq G$, $\mathbb{E}[|g_t - \mathbb{E}[g_t]|^{\mathfrak{p}}] \leq \sigma^{\mathfrak{p}}$ *for some* $\mathfrak{p} \in (1, 2]$. *If we set* $\tau = T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}$ *then with probability at least* $1 - \delta$, *Algorithm 2 guarantees:*

$$R_T(u) \leq \tilde{O}\left[\epsilon \log \frac{1}{\delta} + |u| T^{1/\mathfrak{p}}(\sigma + G) \log \frac{T}{\delta} \log \frac{|u|T}{\epsilon}\right]$$

Proof. We copy over $\phi(w)$ and regret formula in equation (8) and (9) from section 5 here, as the analysis will be following the cancellation-by-regularization strategy described in section 5.

$$\phi(w) = 2^{\mathfrak{p}-1}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})|w|/\tau^{\mathfrak{p}-1} \qquad\qquad \hat{\ell}_t(w) = \langle \mathbb{E}[\hat{g}_t], w - u\rangle + \phi(w)$$

$$R_T(u) \leq \underbrace{\sum_{t=1}^{T} \left(\langle \nabla \ell_t(w_t) - \mathbb{E}[\hat{g}_t], w_t - u\rangle - \phi(w_t) + \phi(u)\right)}_{D} + \underbrace{\sum_{t=1}^{T} \hat{\ell}_t(w_t) - \hat{\ell}_t(u)}_{E}$$

We split the regret into two parts. The term D is controlled by cancellation-by-regularization through careful choice of $\phi_t(w)$ and clipping parameter τ. Term E is controlled in high probability through Algorithm 1 (which uses a *different* cancellation-by-regularization strategy) by sending $\hat{g}_t + \nabla\phi(w_t)$ as the t^{th} subgradient. Specifically, we can view $\hat{g}_t + \nabla\phi(w_t)$ as a sub-exponential and bounded noisy gradient and $\mathbb{E}[\hat{g}_t + \nabla\phi(w_t)] = \nabla\hat{\ell}_t(w_t)$, so that Theorem 3 provides a high probability bound for term E.

First, we bound D**,** and show it's independent of $|w_t|$

$$D \leq \sum_{t=1}^{T} |\nabla \ell_t(w_t) - \mathbb{E}[\hat{g}_t]|(|w_t| + |u|) - 2^{\mathfrak{p}-1}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})/\tau^{\mathfrak{p}-1} \sum_{t=1}^{T} |w_t| + 2^{\mathfrak{p}-1}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})T|u|/\tau^{\mathfrak{p}-1}$$

since $\nabla \ell_t(w_t) = \mathbb{E}[g_t]$, by Lemma 15, $|\nabla \ell_t(w_t) - \mathbb{E}[\hat{g}_t]| \leq 2^{\mathfrak{p}-1}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})/\tau^{\mathfrak{p}-1}$.

$$\leq 2^{\mathfrak{p}} T|u|(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})/\tau^{\mathfrak{p}-1}$$

set $\tau = T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}$

$$= 2^{\mathfrak{p}}|u| T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}$$
$$\leq 4|u| T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}$$

Now we bound E in high probability with $\tau = T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}$. We sometimes will substitute the value of τ and sometimes leave it as it is during the derivation for convenience. Define the noise as ϵ_t,

$$\epsilon_t = \nabla\hat{\ell}_t(w_t) - (\hat{g}_t + \nabla\phi(w_t)) = \mathbb{E}[\hat{g}_t] - \hat{g}_t$$

From the definition of gradient clipping, $|\hat{g}_t| \leq \tau$. Also by Lemma 15,

$$\mathbb{E}[\hat{g}_t^2 | w_t] \leq 2\tau^{2-\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}) = 2\tau^2 T^{-1}$$

Hence term E can be bounded by Theorem 3, where we set the following constants for Algorithm 1,

$$c_1 = 2\tau\sqrt{\tfrac{2}{T}\log\left(\tfrac{32}{\delta}\left[\log\left(2^{T+1}\right) + 2\right]^2\right)}, \quad c_2 = 32\tau\log\left(\tfrac{224}{\delta}\left[\log\left(1 + \sqrt{T}2^{T+5/2}\right) + 2\right]^2\right),$$
$$p_1 = 2, \qquad p_2 = \log T, \qquad \alpha_1 = \epsilon/c_1, \qquad \alpha_2 = (\sqrt{2}\epsilon)/(4\sqrt{T}(\tau + H))$$

where $H = c_1 p_1 + c_2 p_2$. Let $N = 1 + \log\left(\frac{(\tau+H)^2|u|^2 T^C}{\epsilon^2} + 1\right)$. Then by equation (17), with probability at least $1 - \frac{\delta}{2}$ for some positive A, B, C,

$$E \leq \epsilon\left(3 + \frac{8\tau\sqrt{2/T}}{\tau + H}\log\left(\frac{224}{\delta}\left[\log\left(1 + \sqrt{T}2^{T+5/2}\right) + 2\right]^2\right)\right)$$
$$+ |u|\left[4c_1(A^2 + B)N + \frac{3A}{2}\sqrt{N\max\left(0, \sum_{t=1}^{T}|g_t|^2 - |\nabla\psi_t(w_t)|^2\right)}\right]$$
$$+ |u|\tau\left[2BN + 4\log\frac{8}{\delta} + \frac{c_2\log T}{\tau}\left((A^2 + 2B)N + 3\left(\max\left(0, \log T\left(\log\frac{|u|}{\alpha_2} + 1\right)\right) + 4\right)\right)\right]$$
$$+ |u|\sqrt{T}\left[2c_1\left(\sqrt{\log\left(\frac{T|u|c_1^2}{\epsilon^2} + 1\right)} + 1\right) + 2\tau\sqrt{\frac{1}{T}\log\frac{8}{\delta}}\right]$$

Combining D, E and substitute τ when convenient and group in terms of the product of $|u|$ with T to some power of $\mathfrak{p}$

$$R_T(\mathbf{u}) \leq \epsilon\left(3 + \frac{8\tau\sqrt{2/T}}{\tau + H}\log\left(\frac{224}{\delta}\left[\log\left(1 + \sqrt{T}2^{T+5/2}\right) + 2\right]^2\right)\right)$$
$$+ |u|\frac{3A}{2}\sqrt{N\max\left(0, \sum_{t=1}^{T}|\hat{g}_t + \nabla\phi(w_t)|^2 - |\nabla\psi_t(w_t)|^2\right)}$$
$$+ |u|T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}\left[2BN + 2\sqrt{\log\frac{8}{\delta}} + 4\log\frac{8}{\delta}\right.$$
$$+ 4\sqrt{2\log\left(\frac{32}{\tau}[\log(2^{T+1}) + 2]^2\right)}\left(\frac{1}{\sqrt{T}}(A^2 + 2B)N + \sqrt{\log\left(\frac{T|u|c_1^2}{\epsilon^2} + 1\right)} + 1\right)$$
$$\left. + \frac{c_2\log T}{\tau}\left(2(A^2 + B)N + 3\max\left(\max\left(0, \log T\left(\log\frac{|u|}{\alpha_2} + 1\right)\right) + 4\right) + 4\right)\right] \tag{19}$$

For the final statement, notice that although $\hat{g}_t$ is a random quantity, we can bound $\sum_{t=1}^{T}|\hat{g}_t|^2$ with high probability. By Lemma 15, $\mathbb{E}[\hat{g}_t^2] \leq 2^{\mathfrak{p}-1}\tau^{2-\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})$ and note $|\hat{g}_t| \leq \tau$. Thus by Lemma 24, with probability at least $1 - \delta/2$

$$\sum_{t=1}^{T}|\hat{g}_t|^2 \leq 3T2^{\mathfrak{p}-2}\tau^{2-\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}) + 2\tau^2\log\frac{2}{\delta}$$
$$\leq \tau^2(3 + 2\log\frac{2}{\delta})$$

Finally, since equation (19) holds for probability as least $1 - \frac{\delta}{2}$, we further upperbound $|\hat{g}_t|^2$ by union bound for our final regret guarantee with probability at least $1 - \delta$

$$R_T(u) \leq \epsilon\left(3 + \frac{8\tau\sqrt{2/T}}{\tau + H}\log\left(\frac{224}{\delta}\left[\log\left(1 + \sqrt{T}2^{T+5/2}\right) + 2\right]^2\right)\right)$$
$$+ |u|\frac{3A}{2}\sqrt{N\max\left(0, \tau^2(3 + 2\log\frac{2}{\delta}) + \sum_{t=1}^{T}|\nabla\phi(w_t)|^2 - |\nabla\psi_t(w_t)|^2\right)}$$
$$+ |u|T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}\left[2BN + 2\sqrt{\log\frac{8}{\delta}} + 4\log\frac{8}{\delta}\right.$$

$$+ 4\sqrt{2\log\left(\frac{32}{\delta}\left[\log\left(2^{T+1}\right)+2\right]^2\right)}\left(\frac{1}{\sqrt{T}}(A^2+2B)N+\sqrt{\log\left(\frac{T|u|c_1^2}{\epsilon^2}+1\right)}+1\right)$$
$$+\frac{c_2\log T}{\tau}\left(2(A^2+B)N+3\max\left(0,\log T\log\left(\frac{3|u|}{\alpha_2}\right)\right)+4\right)\Bigg]$$

where $|\nabla\phi(w_t)|\leq 2^{\mathfrak{p}-1}T^{\frac{1}{\mathfrak{p}}-1}(\sigma^{\mathfrak{p}}+G^{\mathfrak{p}})^{\frac{1}{\mathfrak{p}}}$ □

D Dimension-free Gradient Clipping for Heavy-tailed Gradients

Lemma 16. *(Unit Ball Domain Algorithm High Probability) Suppose $\{\mathbf{g}_t\}$ is a sequence of heavy-tailed stochastic gradient vectors such that $\mathbb{E}[|\mathbf{g}_t|]\leq G$, $\mathbb{E}[\|\mathbf{g}_t-\mathbb{E}[\mathbf{g}_t]\|^{\mathfrak{p}}]\leq\sigma^{\mathfrak{p}}$ for some $\mathfrak{p}\in(1,2]$. Let $\hat{\mathbf{g}}_t$ be the clipped gradient $\hat{\mathbf{g}}_t=\mathbf{g}_t/\|\mathbf{g}_t\|\min(\tau,\|\mathbf{g}_t\|)$, where τ is set as $T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}}+G^{\mathfrak{p}})^{1/\mathfrak{p}}$. The constrained domain on unit ball ensures $\|\mathbf{w}_t\|,\|\mathbf{u}\|\leq 1$. Then with probability at least $1-\delta$, algorithm 3 guarantees*

$$\begin{aligned}R_T^{nd}(\mathbf{u})&\leq\sum_{t=1}^T\langle\mathbb{E}[\mathbf{g}_t],\mathbf{w}_t-\mathbf{u}\rangle\\&\leq T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}}+G^{\mathfrak{p}})^{1/\mathfrak{p}}\left(\frac{1}{2}\|\mathbf{u}\|^2+\left(\frac{3}{2}+\frac{20}{3}\log\frac{2}{\delta}\right)+15\sqrt{\log\frac{160}{\delta}}\right.\\&\qquad\left.+184\log\left(\frac{448}{\delta}\left[\log\left(2+\frac{16}{\tau}\right)+1\right]^2\right)+4\right)\end{aligned}$$

Proof. The analysis follows similar to the $1d$ analysis as seen in Theorem 4. We run a standard Follow-the-regularized-leader (FTRL) algorithm with L_2 regularization on clipped gradient $\hat{\mathbf{g}}_t$ instead of the true gradient $\nabla\ell_t(\mathbf{w}_t)$. A 'bias' term was introduced by $\mathbb{E}[\mathbf{g}_t]-\mathbb{E}[\hat{\mathbf{g}}_t]$ which can be regulated to be sublinear by the clipping parameter τ. In addition, a 'noise' term due to $\mathbb{E}[\hat{\mathbf{g}}_t]-\hat{\mathbf{g}}_t$ can be bounded with high probability. We decompose the regret and label the corresponding parts below,

$$\begin{aligned}R_T^{nd}(\mathbf{u})&=\sum_{t=1}^T\langle\nabla\ell_t(\mathbf{w}_t),\mathbf{w}_t-\mathbf{u}\rangle\\&=\sum_{t=1}^T\langle\hat{\mathbf{g}}_t,\mathbf{w}_t-\mathbf{u}\rangle+\sum_{t=1}^T\langle\nabla\ell_t(\mathbf{w}_t)-\mathbb{E}[\hat{\mathbf{g}}_t],\mathbf{w}_t-\mathbf{u}\rangle+\sum_{t=1}^T\langle\mathbb{E}[\hat{\mathbf{g}}_t]-\hat{\mathbf{g}}_t,\mathbf{w}_t-\mathbf{u}\rangle\\&\leq\underbrace{\|\sum_{t=1}^T\langle\hat{\mathbf{g}}_t,\mathbf{w}_t-\mathbf{u}\rangle\|}_{\text{FTRL}}+\underbrace{\sum_{t=1}^T\langle\nabla\ell_t(\mathbf{w}_t)-\mathbb{E}[\hat{\mathbf{g}}_t],\mathbf{w}_t-\mathbf{u}\rangle}_{\text{'bias'}}+\underbrace{\|\sum_{t=1}^T\langle\mathbb{E}[\hat{\mathbf{g}}_t]-\hat{\mathbf{g}}_t,\mathbf{w}_t-\mathbf{u}\rangle\|}_{\text{'noise'}}\end{aligned}$$

Now we will bound the regret in three step.

Step 1 : For the part controlled by a FTRL algorithm with fixed L_2 regularization weight η (see Corollary 7.8 in Orabona [2019])

$$\sum_{t=1}^T\langle\hat{\mathbf{g}}_t,\mathbf{w}_t-\mathbf{u}\rangle\leq\frac{1}{2\eta}\|\mathbf{u}\|^2+\frac{\eta}{2}\sum_{t=1}^T\|\hat{\mathbf{g}}_t\|^2$$

Since $\|\hat{\mathbf{g}}_t\|\leq\tau$, and by Lemma 15, $\mathbb{E}[\|\hat{\mathbf{g}}_t\|^2]\leq 2^{\mathfrak{p}-1}\tau^{2-\mathfrak{p}}(\sigma^{\mathfrak{p}}+G^{\mathfrak{p}})\leq 2\tau/T$. By Proposition 17, $\mathbf{g}_t$ is $(\tau\sqrt{2/T},2\tau)$ sub-exponential. Hence by Lemma 24, with probability at least $1-\delta$

$$\leq\frac{1}{2\eta}\|\mathbf{u}\|^2+\frac{\eta}{2}\left(3\tau^2+\frac{20}{3}\tau^2\log\frac{1}{\delta}\right)$$

set $\eta=1/\tau$

$$\leq\frac{\tau}{2}\|\mathbf{u}\|^2+\tau\left(\frac{3}{2}+\frac{20}{3}\log\frac{1}{\delta}\right)$$

Step 2 : For the 'bias' term, note $\nabla \ell_t(\mathbf{w}_t) = \mathbb{E}[\mathbf{g}_t]$, by Lemma 15, $\| \mathbb{E}[\hat{\mathbf{g}}_t] - \mathbb{E}[\mathbf{g}_t] \| \le \frac{2^{\mathfrak{p}-1}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})}{\tau^{\mathfrak{p}-1}}$, and the constrained domain suggests $\|\mathbf{w}_t - \mathbf{u}\| \le 2$

$$\sum_{t=1}^{T} \langle \nabla \ell_t(\mathbf{w}_t) - \mathbb{E}[\hat{\mathbf{g}}_t], \mathbf{w}_t - \mathbf{u} \rangle \le \frac{2^{\mathfrak{p}} T}{\tau^{\mathfrak{p}-1}} (\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})$$

Step 3 : For a high probability bound for the 'noise' term, let $X_t = \langle \mathbb{E}[\hat{\mathbf{g}_t}] - \hat{\mathbf{g}}_t, \mathbf{w}_t - \mathbf{u} \rangle$, hence $\{X_t\}$ is a vector valued MDS adapted to filtration $\mathcal{F}_t$ with the following bound almost surely,

$$\begin{aligned} X_t &\le (\mathbb{E}[\hat{\mathbf{g}_t}] + \|\hat{\mathbf{g}}_t\| \|)\mathbf{w}_t - \mathbf{u}\| \\ &\le 2\tau \|\mathbf{w}_t - \mathbf{u}\| \end{aligned}$$

$$\mathbb{E}[\|X_t\|^2 \mid \mathcal{F}_t] \le \|\mathbf{w}_t - \mathbf{u}\|^2 \, \mathbb{E}[\|\hat{\mathbf{g}}_t\|^2 \mid \mathcal{F}_t]$$

by Lemma 15

$$\begin{aligned} &\le \|\mathbf{w}_t - \mathbf{u}\|^2 2^{\mathfrak{p}-1} \tau^{2-\mathfrak{p}} (\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}) \\ &\le 2 \|\mathbf{w}_t - \mathbf{u}\|^2 T^{2/\mathfrak{p}-1} (\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{2/\mathfrak{p}} \\ &= 2\|\mathbf{w}_t - \mathbf{u}\|^2 \tau^2 T^{-1} \end{aligned}$$

and both bounds are $\mathcal{F}_{t-1}$ measurable. By proposition 17 $\{\langle \mathbb{E}[\hat{\mathbf{g}}_t] - \hat{\mathbf{g}}_t, \mathbf{w}_t - \mathbf{u} \rangle\}$ is $\{\sqrt{2}\|\mathbf{w}_t - \mathbf{u}\| \tau T^{-1/2}, 4\tau \|\mathbf{w}_t - \mathbf{u}\|\}$ sub-exponential noise. Use Theorem 19 and set $\nu = \tau$, with probability at least $1 - \delta$,

$$\begin{aligned} \left\| \sum_{t=1}^{T} X_t \right\| &\le 5 \sqrt{2 \frac{\tau^2}{T} \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{u}\|^2 \log \left(\frac{16}{\delta} \left[\log \left(\left[\sqrt{\frac{2}{T} \sum_{t=1}^{T} \|\mathbf{w}_t - \mathbf{u}\|^2} \right]_1 \right) + 1 \right]^2 \right)} \\ &\quad + 23 \max(\tau, 4\tau \max_{t \le T} \|\mathbf{w}_t - \mathbf{u}\|) \log \left(\frac{224}{\delta} \left[\log \left(2 \max(1, \frac{4}{\tau} \max_{t \le T} \|\mathbf{w}_t - \mathbf{u}\|) \right) + 1 \right]^2 \right) \\ &\le 15\tau \sqrt{\log \frac{80}{\delta}} + 184\tau \log \left(\frac{224}{\delta} \left[\log \left(2 + \frac{16}{\tau} \right) + 1 \right]^2 \right) \\ &= T^{1/\mathfrak{p}} (\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}} \left(15 \sqrt{\log \frac{80}{\delta}} + 184 \log \left(\frac{224}{\delta} \left[\log \left(2 + \frac{16}{\tau} \right) + 1 \right]^2 \right) \right) \end{aligned}$$

Composition : Combining the high probability bound from step 1 and 3 by union bound and a deterministic bound from step 2, with probability at least $1 - \delta$, we have the following regret guarantee,

$$\begin{aligned} R_T^{nd}(\mathbf{u}) &\le \frac{\tau}{2} \|\mathbf{u}\|^2 + \tau \left(\frac{3}{2} + \frac{20}{3} \log \frac{2}{\delta} \right) + \frac{2^{\mathfrak{p}} T}{\tau^{\mathfrak{p}-1}} (\sigma^{\mathfrak{p}} + G^{\mathfrak{p}}) \\ &\quad + T^{1/\mathfrak{p}} (\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}} \left(15 \sqrt{\log \frac{160}{\delta}} + 184 \log \left(\frac{448}{\delta} \left[\log \left(2 + \frac{16}{\tau} \right) + 1 \right]^2 \right) \right) \end{aligned}$$

substitute $\tau = T^{1/\mathfrak{p}} (\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}$ and group terms by factorizing some power of T

$$\begin{aligned} &\le T^{1/\mathfrak{p}} (\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}} \left(\frac{1}{2} \|\mathbf{u}\|^2 + \left(\frac{3}{2} + \frac{20}{3} \log \frac{2}{\delta} \right) + 15 \sqrt{\log \frac{160}{\delta}} \right. \\ &\quad \left. + 184 \log \left(\frac{448}{\delta} \left[\log \left(2 + \frac{16}{\tau} \right) + 1 \right]^2 \right) + 4 \right) \end{aligned}$$

□

We restate Theorem 5 for reference followed by its proof

Theorem 5. *Suppose that for all t, $\{\mathbf{g}_t\}$ are heavy-tailed stochastic subgradients satisfying $\mathbb{E}[\mathbf{g}_t] \in \partial \ell_t(\mathbf{w}_t)$, $\|\mathbb{E}[\mathbf{g}_t]\| \le G$ and $\mathbb{E}[\|\mathbf{g}_t - \mathbb{E}[\mathbf{g}_t]\|^{\mathfrak{p}}] \le \sigma^{\mathfrak{p}}$ for some $\mathfrak{p} \in (1, 2]$. Then, with probability at least $1 - \delta$, Algorithm 4 guarantees*

$$R_T(\mathbf{u}) = \sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u}) \le \tilde{O}\left[\epsilon \log \frac{1}{\delta} + \|\mathbf{u}\| T^{1/\mathfrak{p}}(\sigma + G) \log \frac{T}{\delta} \log \frac{\|\mathbf{u}\| T}{\epsilon}\right]$$

Proof. This result relies on the reduction from dimension-free learning to 1d learning presented in Theorem 2 of Cutkosky and Orabona [2018]. This result implies that the regret 4 can be be bounded as:

$$\sum_{t=1}^{T} \ell_t(\mathbf{w}_t) - \ell_t(\mathbf{u}) \le R_T^{1d}(\|\mathbf{u}\|) + \|\mathbf{u}\| R_T^{nd}(\mathbf{u}/\|\mathbf{u}\|)$$

where

$$R_T^{1d}(\|u\|) = \sum_{t=1}^{T} \langle g_t^{1d}, x_t - \|u\| \rangle$$
$$R_T^{nd}(\mathbf{u}/\|\mathbf{u}\|) = \sum_{t=1}^{T} \langle \mathbf{g}_t, \mathbf{v}_t - \mathbf{u}/\|\mathbf{u}\| \rangle$$

First, by setting Algorithm 2 as $\mathcal{A}^{1d}$ is , Theorem 4 provides a bound for $R_T^{1d}(\|\mathbf{u}\|)$ with appropriate parameters set in the theorem. We run $\mathcal{A}^{1d}$ on subgradients $g_t = \langle \mathbf{g}_t, \mathbf{v}_t \rangle$, and $\|\mathbf{v}_t\| \le 1$. We show $|\mathbb{E}[g_t]|$ and $\mathbb{E}[\|g_t - \mathbb{E}[g_t]\|^{\mathfrak{p}}]$ are bounded. Notice that g_t only depends on $\mathbf{g}_1, \cdots, \mathbf{g}_{t-1}$. Thus, by tower rule, $\mathbb{E}[g_t] = \mathbb{E}[\langle \mathbf{g}_t, \mathbf{v}_t \rangle] = \mathbb{E}[\langle \mathbb{E}[\mathbf{g}_t], \mathbf{v}_t \rangle \mid \mathbf{g}_1, \cdots, \mathbf{g}_{t-1}]$. Then, since $\|\mathbf{v}_t\| \le 1$ we have $|\langle \mathbb{E}[\mathbf{g}_t], \mathbf{v}_t \rangle| \le G$. Now we are left to show a bounded central moment:

$$\begin{aligned}
\mathbb{E}[|g_t - \mathbb{E}[g_t]|^{\mathfrak{p}}] &= \mathbb{E}[|\langle \mathbf{g}_t - \mathbb{E}[\mathbf{g}_t], \mathbf{v}_t \rangle + \langle \mathbb{E}[\mathbf{g}_t], \mathbf{v}_t \rangle - \mathbb{E}[\langle \mathbf{g}_t, \mathbf{v}_t \rangle]|^{\mathfrak{p}}] \\
&\le \mathbb{E}[|\langle \mathbf{g}_t - \mathbb{E}[\mathbf{g}_t], \mathbf{v}_t \rangle|^{\mathfrak{p}}] + |\langle \mathbb{E}[\mathbf{g}_t], \mathbf{v}_t \rangle|^{\mathfrak{p}} + |\mathbb{E}[\langle \mathbf{g}_t, \mathbf{v}_t \rangle]|^{\mathfrak{p}} \\
&\le \mathbb{E}[\|\mathbf{g}_t - \mathbb{E}[\mathbf{g}_t]\|^{\mathfrak{p}}] + 2\|\mathbb{E}[\mathbf{g}_t]\|^{\mathfrak{p}} \\
&\le \hat{\sigma}^{\mathfrak{p}} + 2G^{\mathfrak{p}} = \sigma^{\mathfrak{p}}
\end{aligned}$$

Then by Theorem 4, when Algorithm 2 is run on $g_t = \langle \mathbf{g}_t, \mathbf{v}_t \rangle$, we obtain with probability at least $1 - \delta$

$$\begin{aligned}
R_T^{1d}(\|\mathbf{u}\|) \le \epsilon &\left(3 + \frac{8\tau\sqrt{2/T}}{\tau + H} \log\left(\frac{224}{\delta}\left[\log\left(1 + \sqrt{T} 2^{T+5/2}\right) + 2\right]^2\right)\right) \\
&+ \|\mathbf{u}\| \frac{3A}{2} \sqrt{N \max\left(0, \tau^2(3 + 2\log\frac{2}{\delta}) + \sum_{t=1}^{T} |\nabla\phi(w_t)|^2 - |\nabla\psi_t(w_t)|^2\right)} \\
&+ \|\mathbf{u}\| T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}} \Bigg[2BN + 2\sqrt{\log\frac{8}{\delta}} + 4\log\frac{8}{\delta} \\
&\quad + 4\sqrt{2\log\left(\frac{32}{\delta}\left[\log\left(2^{T+1}\right) + 2\right]^2\right)} \left(\frac{1}{\sqrt{T}}(A^2 + 2B)N + \sqrt{\log\left(\frac{T\|\mathbf{u}\| c_1^2}{\epsilon^2} + 1\right)} + 1\right) \\
&\quad + \frac{c_2 \log T}{\tau}\left(2(A^2 + B)N + 3\max\left(0, \log T \log\left(\frac{3\|\mathbf{u}\|}{\alpha_2}\right)\right) + 4\right)\Bigg] \qquad (20)
\end{aligned}$$

In terms of $R_T^{nd}(\mathbf{u}/\|\mathbf{u}\|)$, Lemma 16 implies that with probability at least $1 - \delta/2$,

$$R_T^{nd}(\mathbf{u}/\|\mathbf{u}\|) \le T^{1/\mathfrak{p}}(\hat{\sigma}^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}\left(\frac{9}{2} + \left(\frac{3}{2} + \frac{20}{3}\log\frac{4}{\delta}\right) + 15\sqrt{\log\frac{320}{\delta}}\right.$$

$$+ 184 \log\left(\frac{896}{\delta}\left[\log\left(2+\frac{16}{\tau}\right)+1\right]^2\right)\Bigg) \tag{21}$$

Finally, replace δ as $\delta/2$ in equation (20), then combining with equation (21), we have the regret guarantee with probability at least $1-\delta$

$$\begin{aligned}
R_T(\mathbf{u}) &\le R_T^{1d}(\|\mathbf{u}\|) + \|\mathbf{u}\| R_T^{nd}(\mathbf{u}/\|\mathbf{u}\|) \\
&\le \epsilon\left(3 + \frac{8\tau\sqrt{2/T}}{\tau+H}\log\left(\frac{448}{\delta}\left[\log\left(1+\sqrt{T}2^{T+5/2}\right)+2\right]^2\right)\right) \\
&\quad + \|\mathbf{u}\|\frac{3A}{2}\sqrt{N\max\left(0, \tau^2(3+2\log\frac{4}{\delta}) + \sum_{t=1}^T |\nabla\phi(w_t)|^2 - |\nabla\psi_t(w_t)|^2\right)} \\
&\quad + \|\mathbf{u}\| T^{1/\mathfrak{p}}(\sigma^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}\Bigg[2BN + 2\sqrt{\log\frac{16}{\delta}} + 4\log\frac{16}{\delta} \\
&\quad + 4\sqrt{2\log\left(\frac{64}{\delta}\left[\log\left(2^{T+1}\right)+2\right]^2\right)}\left(\frac{1}{\sqrt{T}}(A^2+2B)N + \sqrt{\log\left(\frac{T\|\mathbf{u}\|c_1^2}{\epsilon^2}+1\right)}+1\right) \\
&\quad + \frac{c_2\log T}{\tau}\left(2(A^2+B)N + 3\max\left(0, \log T\log\left(\frac{3\|\mathbf{u}\|}{\alpha_2}\right)\right)+4\right)\Bigg] \\
&\quad + \|\mathbf{u}\| T^{1/\mathfrak{p}}(\hat{\sigma}^{\mathfrak{p}} + G^{\mathfrak{p}})^{1/\mathfrak{p}}\Bigg[\frac{9}{2} + \left(\frac{3}{2} + \frac{20}{3}\log\frac{4}{\delta}\right) \\
&\quad + 15\sqrt{\log\frac{320}{\delta}} + 184\log\left(\frac{896}{\delta}\left[\log\left(2+\frac{16}{\tau}\right)+1\right]^2\right)\Bigg]
\end{aligned}$$

□

E Technical concentration bounds

In this section we collect some technical results on concentration martingale difference sequences. These results are not new, and the proofs are for the most part exercises in known techniques (e.g. Howard et al. [2021], Balsubramani [2014]), but we cannot find simple explicit statements of exactly the forms we require in the literature, so we provide them here and include proofs for completeness. Our approach follows the martingale mixture method used by Balsubramani [2014]. We make no effort to achieve optimal constants, in several cases explicitly choosing weaker bounds to make the final numbers less complicated.

To start, we recall the notion of a sub-exponential martingale difference sequence (MDS) as follows:

Definition 1. *Suppose $\{X_t\}$ is a sequence of random variables adapted to a filtration $\mathcal{F}_t$ such that $\{X_t, \mathcal{F}_t\}$ is a martingale difference sequence. Further, suppose $\{\sigma_t, b_t\}$ are random variables such that σ_t, b_t are both $\mathcal{F}_{t-1}$-measurable for all t. Then, $\{X_t, \mathcal{F}_t\}$ is $\{\sigma_t, b_t\}$ sub-exponential if*

$$\mathbb{E}[\exp(\lambda X_t)|\mathcal{F}_{t-1}] \le \exp(\lambda^2\sigma_t^2/2)$$

almost everywhere for all $\mathcal{F}_{t-1}$-measurable λ satisfying $\lambda < 1/b_t$.

We have the following useful way to obtain sub-exponential tails:

Proposition 17. *Suppose $\{X_t, \mathcal{F}_t\}$ is a MDS such that $\mathbb{E}[X_t^2|\mathcal{F}_t] \le \sigma_t^2$ and $|X_t| \le b_t$ almost everywhere for all t for some sequence of random variable $\{\sigma_t, b_t\}$ such that σ_t, b_t is $\mathcal{F}_{t-1}$-measurable. Then X_t is $(\sigma_t, 2b_t)$ sub-exponential.*

The main results of this section are the following two Theorems. First for scalar random variables we have the following one-sided concentration bound:

Theorem 18. *Suppose $\{X_t, \mathcal{F}_t\}$ is a (σ_t, b_t) sub-exponential martingale difference sequence. Let ν be an arbitrary constant. Then with probability at least $1-\delta$, for all t it holds that:*

$$\sum_{i=1}^t X_i \le 2\sqrt{\sum_{i=1}^t \sigma_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/2\nu^2}\right]_1\right)+2\right]^2\right)} + 8\max(\nu, \max_{i\le t} b_i)\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu, \max_{i\le t} b_i)}{\nu}\right)+2\right]^2\right)$$

where $[x]_1 = \max(1, x)$.

Next, for vector valued random variables, we have the following bound:

Theorem 19. *Suppose that $\{X_t, \mathcal{F}_t\}$ is a vector-valued martingale difference sequence such that $\mathbb{E}[\|X_t\|^2|\mathcal{F}_{t-1}] \le \sigma_t^2$ and $\|X_t\| \le b_t$ almost everywhere for some sequence $\{\sigma_t, b_t\}$ such that σ_t, b_t is $\mathcal{F}_{t-1}$-measurable. Let $\nu \ge 0$ be an arbitrary constant. Then with probability at least $1-\delta$, for all t we have:*

$$\left\|\sum_{i=1}^t X_i\right\| \le 5\sqrt{\sum_{i=1}^t \sigma_i^2 \log\left(\frac{16}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right)+2\right]^2\right)} + 23\max(\nu, \max_{i\le t} b_i)\log\left(\frac{224}{\delta}\left[\log\left(\frac{2\max(\nu, \max_{i\le t} b_i)}{\nu}\right)+2\right]^2\right)$$

where $[x]_1 = \max(1, x)$.

E.1 Time-Uniform Concentration of sums of sub-exponential MDS

In this section, we will prove Theorem 18.

Theorem 18. *Suppose $\{X_t, \mathcal{F}_t\}$ is a (σ_t, b_t) sub-exponential martingale difference sequence. Let ν be an arbitrary constant. Then with probability at least $1-\delta$, for all t it holds that:*

$$\sum_{i=1}^t X_i \le 2\sqrt{\sum_{i=1}^t \sigma_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/2\nu^2}\right]_1\right)+2\right]^2\right)} + 8\max(\nu, \max_{i\le t} b_i)\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu, \max_{i\le t} b_i)}{\nu}\right)+2\right]^2\right)$$

where $[x]_1 = \max(1, x)$.

Proof. First, observe that by replacing b_t with $\max_{i\le t} b_i$, we may assume $b_t \ge b_{t-1}$ for all t with probability 1. Notice that after this operation, X_t is still (σ_t, b_t) sub-exponential. Under this assumption, it suffices to prove the result with b_t in place of $\max_{i\le t} b_i$.

Define $M_t = \sum_{i=1}^t X_i$ and let $\pi(\eta)$ be a to-be-specified probability density function on $\mathbb{R}$. Define:

$$Z_t = \int_0^{1/\max(\nu, b_t)} \pi(\eta)\exp\left[\eta M_t - \frac{\eta^2\sum_{i=1}^t \sigma_i^2}{2}\right]\, d\eta$$

Notice that $Z_0 \le \int_0^\infty \pi(\eta)\, d\eta = 1$. We claim that Z_t is itself a supermartingale adapted to the same filtration $\mathcal{F}_t$:

$$\mathbb{E}\left[Z_t|\mathcal{F}_{t-1}\right] = \mathbb{E}\left[\int_0^{1/\max(\nu, b_t)} \pi(\eta)\exp\left[\eta M_t - \frac{\eta^2\sum_{i=1}^t \sigma_i^2}{2}\right]\, d\eta|\, \mathcal{F}_{t-1}\right]$$

$$= \mathbb{E}\left[\int_0^{1/\max(\nu,b_t)} \pi(\eta)\exp\left[\eta M_{t-1} - \frac{\eta^2 \sum_{i=1}^{t-1}\sigma_i^2}{2}\right]\exp\left[\eta X_t - \frac{\eta^2\sigma_t^2}{2}\right]\,d\eta|\,\mathcal{F}_{t-1}\right]$$

$$= \int_0^{1/\max(\nu,b_t)} \pi(\eta)\exp\left[\eta M_{t-1} - \frac{\eta^2 \sum_{i=1}^{t-1}\sigma_i^2}{2}\right]\mathbb{E}\left[\exp\left[\eta X_t - \frac{\eta^2\sigma_t^2}{2}\right]|\,\mathcal{F}_{t-1}\right]\,d\eta$$

Use the sub-exponentiality of X_t:

$$\leq \int_0^{1/\max(\nu,b_t)} \pi(\eta)\exp\left[\eta M_{t-1} - \frac{\eta^2 \sum_{i=1}^{t-1}\sigma_i^2}{2}\right]\,d\eta$$

Use $b_t \geq b_{t-1}$:

$$\leq \int_0^{1/\max(\nu,b_{t-1})} \pi(\eta)\exp\left[\eta M_{t-1} - \frac{\eta^2 \sum_{i=1}^{t-1}\sigma_i^2}{2}\right]\,d\eta$$

$$= Z_{t-1}$$

Therefore, by Ville's maximal inequality [Ville, 1939], we have that for all $\delta > 0$:

$$P\left[\sup_t Z_t \geq 1/\delta\right] \leq \delta Z_0 = \delta$$

Put another way, with probability at least $1-\delta$, $Z_t \leq 1/\delta$ for all t.

Now, let us define $\pi(\eta)$. With the benefit of foresight, we choose a density on $[0, 1/\nu]$:

$$\pi(\eta) = \frac{1}{\eta(\log(1/\eta\nu)+2)^2} \tag{22}$$

We have

$$\pi(\eta) = \frac{d}{d\eta}\frac{1}{\log(1/\eta\nu)+2} \tag{23}$$

$$\frac{d\pi(\eta)}{d\eta} = -\frac{\log(1/\eta\nu)}{\eta^2\left(\log(1/\eta\nu)+2\right)^3} \tag{24}$$

so that $\int_0^{1/\nu}\pi(\eta)\,d\eta = 1$ and $\pi(\eta)$ is decreasing on $[0, 1/\nu]$.

Next, for any given $\eta_\star \in [0, 1/\nu]$, for all $K \geq 1$, for all $\eta \in [\eta_\star/K, \eta_\star]$, we have:

$$\int_{\eta_\star/K}^{\eta_\star}\pi(\eta)\,d\eta \geq \pi(\eta_\star)\frac{K-1}{K}\eta_\star$$

$$\geq \frac{(K-1)}{K(\log(1/\eta_\star\nu)+2)^2}$$

Further, for all $\eta \in [\eta_\star/K, \eta_\star]$, if $M_t \geq 0$ we also have:

$$\eta M_t - \frac{\eta^2\sum_{i=0}^{t-1}\sigma_i^2}{2} \geq \frac{\eta_\star M_t}{K} - \frac{\eta_\star^2\sum_{i=0}^{t-1}\sigma_i^2}{2}$$

and otherwise of course we have $M_t \leq 0$.

Combining these observations, we have that with probability at least $1-\delta$, for all $\mathcal{F}_{t-1}$-measurable $\eta_\star$ satisfying $\eta_\star \leq 1/\max(\nu, b_t)$ and all $K \geq 1$, either $M_t \leq 0$ or

$$\frac{(K-1)}{K(\log(1/\eta_\star\nu)+2)^2}\exp\left(\frac{\eta_\star M_t}{K} - \frac{\eta_\star^2\sum_{i=1}^{t}\sigma_i^2}{2}\right) \leq \int_0^{1/b_t}\pi(\eta)\exp\left[\eta M_t - \frac{\eta^2\sum_{i=1}^{t}\sigma_i^2}{2}\right]\,d\eta \leq \delta$$

Now, rearranging this identity implies:

$$M_t \le \log\left(\frac{K}{(K-1)\delta}\left[\log\left(\frac{1}{\eta_\star \nu}\right)+2\right]^2\right)\frac{K}{\eta_\star} + \frac{K\eta_\star \sum_{i=1}^t \sigma_i^2}{2} \tag{25}$$

So that overall we may discard the $M_t \le 0$ case as it is strictly weaker than the above.

Now, again with the benefit of foresight, let us select

$$\eta_\star = \min\left(\frac{1}{\max(\nu, b_t)}, \sqrt{\frac{2\log\left[\frac{K}{(K-1)\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/2\nu^2}\right]_1\right)+2\right]^2\right]}{\sum_{i=1}^t \sigma_i^2}}\right)$$

where $[x]_1 = \max(1, x)$. Notice that $\eta_\star$ is $\mathcal{F}_{t-1}$-measurable since b_t and $\sigma_1, \dots, \sigma_t$ are $\mathcal{F}_{t-1}$-measurable, and $\eta_\star \in [0, 1/\nu]$ with probability 1.

Now, to analyze the expression (25), we will consider both cases of the above minimum. First, let us assume

$$1/\max(\nu, b_t) \ge \sqrt{\frac{2\log\left[\frac{K}{(K-1)\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/2\nu^2}\right]_1\right)+2\right]^2\right]}{\sum_{i=1}^t \sigma_i^2}}$$

Then we can bound $\eta_\star$:

$$\begin{aligned}\eta_\star &\ge \sqrt{\frac{2\log\left[\frac{K}{(K-1)\delta}[2]^2\right]}{\sum_{i=1}^t \sigma_i^2}} \\ &\ge \sqrt{\frac{2}{\sum_{i=1}^t \sigma_i^2}}\end{aligned}$$

Therefore:

$$\log\left(\frac{K}{(K-1)\delta}\left[\log\left(\frac{1}{\eta_\star\nu}\right)+2\right]^2\right) \le \log\left(\frac{K}{(K-1)\delta}\left[\log\left(\left[\sqrt{\sum_{i=0}^{t-1}\sigma_i^2/2\nu^2}\right]_1\right)+2\right]^2\right)$$

from which we conclude:

$$M_t \le K\sqrt{2\sum_{i=1}^t \sigma_i^2 \log\left(\frac{K}{(K-1)\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/2\nu^2}\right]_1\right)+2\right]^2\right)}$$

Now, on the other hand let us suppose that

$$1/\max(\nu, b_t) < \sqrt{\frac{2\log\left[\frac{K}{(K-1)\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/2\nu^2}\right]_1\right)+2\right]^2\right]}{\sum_{i=1}^t \sigma_i^2}}$$

This implies:

$$\sum_{i=1}^t \sigma_i^2 < 2\min(\nu^2, b_t^2)\log\left[\frac{K}{(K-1)\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/2\nu^2}\right]_1\right)+2\right]^2\right]$$

Therefore, from Lemma 25 we have:

$$\sum_{i=1}^t \sigma_i^2 \le 8\max(\nu^2, b_t^2)\log\left[\frac{4\sqrt{K}}{\sqrt{(K-1)\delta}}\log\left(e+16\frac{\max(\nu^2,b_t^2)}{\nu^2}\right)\right]$$
$$\le 8\max(\nu^2, b_t^2)\log\left[\frac{4\sqrt{K}}{\sqrt{(K-1)\delta}}\log\left(20\frac{\max(\nu^2,b_t^2)}{\nu^2}\right)\right]$$

In this case, we will have $\eta_\star = 1/\max(\nu, b_t)$ and so obtain:

$$\begin{aligned}
M_t &\le K\max(\nu,b_t)\log\left(\frac{K}{(K-1)\delta}\left[\log\left(\frac{\max(\nu,b_t)}{\nu}\right)+2\right]^2\right)\\
&\quad + 4\max(\nu,b_t)\log\left[\frac{4\sqrt{K}}{\sqrt{(K-1)\delta}}\log\left(20\frac{\max(\nu^2,b_t^2)}{\nu^2}\right)\right]\\
&\le K\max(\nu,b_t)\log\left(\frac{K}{(K-1)\delta}\left[\log\left(\frac{\max(\nu,b_t)}{\nu}\right)+2\right]^2\right)\\
&\quad + 4\max(\nu,b_t)\log\left[\frac{8\sqrt{K}}{\sqrt{(K-1)\delta}}\log\left(\frac{5\max(\nu,b_{t-1})}{\nu}\right)\right]\\
&\le K\max(\nu,b_t)\log\left(\frac{K}{(K-1)\delta}\left[\log\left(\frac{\max(\nu,b_t)}{\nu}\right)+2\right]^2\right)\\
&\quad + 4\max(\nu,b_t)\log\left[\frac{8\sqrt{K}}{\sqrt{(K-1)\delta}}\left[\log\left(\frac{\max(\nu,b_t)}{\nu}\right)+2\right]\right]\\
&\le 5K\max(\nu,b_t)\log\left(\frac{8K}{(K-1)\delta}\left[\log\left(\frac{\max(\nu,b_t)}{\nu}\right)+2\right]^2\right)
\end{aligned}$$

Finally, let us set $K=\sqrt{2}$ to obtain:

$$\begin{aligned}
M_t &\le K\sqrt{2\sum_{i=1}^t \sigma_i^2 \log\left(\frac{K}{(K-1)\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/2\nu^2}\right]_1\right)+2\right]^2\right)}\\
&\quad + 5K\max(\nu,b_t)\log\left(\frac{8K}{(K-1)\delta}\left[\log\left(\frac{\max(\nu,b_{t-1})}{\nu}\right)+2\right]^2\right)\\
&\le 2\sqrt{\sum_{i=1}^t \sigma_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/2\nu^2}\right]_1\right)+2\right]^2\right)}\\
&\quad + 8\max(\nu,b_t)\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu,b_t)}{\nu}\right)+2\right]^2\right)
\end{aligned}$$

□

E.2 Bounds on Sums of Squares

It is also often useful to bound sums of the form $\sum_{i=1}^t Z_i^2$ for some sequence Z_i. Here we collect a useful bound:

Theorem 20. *Suppose $\{Z_i\}$ is a sequence of random variables adapted to a filtration $\{\mathcal{F}_t\}$. Further, suppose $\mathbb{E}[Z_i^2] \le \sigma_i^2$ and $|Z_i| \le b_i$ for all i with probability 1 for some σ_i and b_i for a sequence*

$\{\sigma_i, b_i\}$ *such that* σ_i *and* b_i *are* $\mathcal{F}_{i-1}$*-measurable. Then for any* $\nu > 0$*, with probability at least* $1-\delta$ *for all* t*:*

$$\sum_{i=1}^t Z_i^2 \leq 3\sum_{i=1}^t \sigma_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right)+2\right]^2\right)$$
$$+ 20\max(\nu^2, \max_{i\leq t} b_i^2)\log\left(\frac{112}{\delta}\left[\log\left(\frac{2\max(\nu, \max_{i\leq t} b_i)}{\nu}\right)+1\right]^2\right)$$

Proof. Define $X_t = Z_t^2 - \mathbb{E}[Z_t^2|\mathcal{F}_{t-1}]$ so that $\{X_t, \mathcal{F}_t\}$ is a MDS. Further, notice that $|X_t| \leq b_t^2$ for all t with probability 1 and

$$\begin{aligned}\mathbb{E}[X_t^2|\mathcal{F}_{t-1}] &\leq b_t^2\mathbb{E}[|Z_t^2 - \mathbb{E}[Z_t^2|\mathcal{F}_{t-1}]| \mid \mathcal{F}_{t-1}] \\ &\leq 2b_t^2\sigma_t^2\end{aligned}$$

Therefore, by Proposition 17, X_t is $(\sqrt{2b_t^2\sigma_t^2}, 2b_t^2)$ sub-exponential.

Thus, by our time-uniform concentration bound (Theorem 18), for any $\nu > 0$, with probability at least $1-\delta$ we have:

$$\begin{aligned}\sum_{i=1}^t X_i &\leq 2\sqrt{2\sum_{i=1}^t \sigma_i^2 b_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2 b_i^2/\nu^4}\right]_1\right)+2\right]^2\right)} \\ &\quad + 8\max(\nu^2, 2\max_{i\leq t} b_i^2)\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_t^2)}{\nu^2}\right)+2\right]^2\right) \\ &\leq 2\sqrt{2\max_{i\leq t} b_i^2\sum_{i=1}^t \sigma_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2 b_i^2/\nu^4}\right]_1\right)+2\right]^2\right)} \\ &\quad + 8\max(\nu^2, 2\max_{i\leq t} b_i^2)\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right)+2\right]^2\right)\end{aligned}$$

Now, by Young inequality:

$$\begin{aligned}&2\max_{i\leq t} b_i^2\sum_{i=1}^t \sigma_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2 b_i^2/\nu^4}\right]_1\right)+2\right]^2\right) \\ &\leq \left(\max_{i\leq t} b_i^2 \log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right)+2\right]^2\right)\right)^2 \\ &\quad + \left(\sum_{i=1}^t \sigma_i^2 \frac{\log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2 b_i^2/\nu^4}\right]_1\right)+2\right]^2\right)}{\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2,2\max_{i\leq t} b_i^2)}{\nu^2}\right)+2\right]^2\right)}\right)^2\end{aligned}$$

To simplify this expression, we consider the following identity:

$$\frac{\log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2 b_i^2/\nu^4}\right]_1\right)+2\right]^2\right)}{\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2,2\max_{i\leq t} b_i^2)}{\nu^2}\right)+2\right]^2\right)} \leq \frac{\log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\sqrt{\max_{i\leq t} b_i^2/\nu^2}\right]_1\right)+2\right]^2\right)}{\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2,2\max_{i\leq t} b_i^2)}{\nu^2}\right)+2\right]^2\right)}$$

$$\leq \frac{\log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right) + \frac{1}{2}\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right) + 2\right]^2\right)}{\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right) + 2\right]^2\right)}$$

Now, consider two cases, either

$$\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right) \leq \frac{1}{2}\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right)$$

or not. In the former case, we have:

$$\begin{aligned}
\frac{\log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2 b_i^2/\nu^4}\right]_1\right) + 2\right]^2\right)}{\log\left(\frac{28}{\delta}\left[\log\left(\frac{2\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right) + 2\right]^2\right)} &\leq \frac{\log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right) + \frac{1}{2}\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right) + 2\right]^2\right)}{\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right) + 2\right]^2\right)} \\
&\leq \frac{\log\left(\frac{4}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right) + 2\right]^2\right)}{\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right) + 2\right]^2\right)} \\
&\leq 1 \\
&\leq \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right) + 2\right]^2\right)
\end{aligned}$$

While in the latter case,

$$\begin{aligned}
\frac{\log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2 b_i^2/\nu^4}\right]_1\right) + 2\right]^2\right)}{\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right) + 2\right]^2\right)} &\leq \frac{\log\left(\frac{4}{\delta}\left[2\log\left(\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right) + 2\right]^2\right)}{\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right) + 2\right]^2\right)} \\
&\leq \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right) + 2\right]^2\right)
\end{aligned}$$

So in both cases, we have

$$\frac{\log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2 b_i^2/\nu^4}\right]_1\right) + 2\right]^2\right)}{\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right) + 2\right]^2\right)} \leq \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right) + 2\right]^2\right)$$

Therefore,

$$\begin{aligned}
2\max_{i\leq t} b_i^2 \sum_{i=1}^t \sigma_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2 b_i^2/\nu^4}\right]_1\right) + 2\right]^2\right) \\
\leq \left(\max_{i\leq t} b_i^2 \log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\leq t} b_i^2)}{\nu^2}\right) + 2\right]^2\right)\right)^2 \\
+ \left(\sum_{i=1}^t \sigma_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right) + 2\right]^2\right)\right)^2
\end{aligned}$$

Combining this with the identity $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, we have:

$$\sum_{i=1}^{t} X_i \le 2\sqrt{2\max_{i\le t} b_i^2 \sum_{i=1}^{t} \sigma_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^{t} \sigma_i^2 b_i^2/\nu^4}\right]_1\right)+2\right]^2\right)}$$
$$+ 8\max(\nu^2, 2\max_{i\le t} b_i^2)\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\le t} b_i^2)}{\nu^2}\right)+2\right]^2\right)$$
$$\le 2\sum_{i=1}^{t} \sigma_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^{t} \sigma_i^2/\nu^2}\right]_1\right)+2\right]^2\right)$$
$$+ 10\max(\nu^2, 2\max_{i\le t} b_i^2)\log\left(\frac{28}{\delta}\left[\log\left(\frac{\max(\nu^2, 2\max_{i\le t} b_i^2)}{\nu^2}\right)+2\right]^2\right)$$
$$\le 2\sum_{i=1}^{t} \sigma_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^{t} \sigma_i^2/\nu^2}\right]_1\right)+2\right]^2\right)$$
$$+ 20\max(\nu^2, \max_{i\le t} b_i^2)\log\left(\frac{112}{\delta}\left[\log\left(\frac{2\max(\nu, \max_{i\le t} b_i)}{\nu}\right)+1\right]^2\right)$$

Finally, observe that

$$\sum_{i=1}^{t} Z_i^2 = \sum_{i=1}^{t} X_i + \mathbb{E}[Z_i^2|\mathcal{F}_{t-1}]$$
$$\le \sum_{i=1}^{t} \sigma_i^2 + \sum_{i=1}^{t} X_i$$
$$\le 3\sum_{i=1}^{t} \sigma_i^2 \log\left(\frac{4}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^{t} \sigma_i^2/\nu^2}\right]_1\right)+1\right]^2\right)$$
$$+ 20\max(\nu^2, \max_{i\le t} b_i^2)\log\left(\frac{112}{\delta}\left[\log\left(\frac{2\max(\nu, \max_{i\le t} b_i)}{\nu}\right)+1\right]^2\right)$$

□

E.3 From scalar to vector (Hilbert space) concentration

In this section we extend our results to concentration of norm of vectors in Hilbert space. The technique follows that of Cutkosky and Mehta [2021], which makes use of a particular scalar sequence associated to any vector sequence described by Cutkosky [2018]. Given any sequence $X_1, X_2, \dots$ of vectors, define a sequence $s_1, s_2, \dots$ of scalars as follows:

1. $s_0 = 0$.
2. If $\sum_{i=1}^{t-1} X_i \ne 0$, set
$$s_t = \text{sign}\left(\sum_{i=1}^{t-1} s_i\right) \frac{\left\langle \sum_{i=1}^{t-1} X_i, X_t \right\rangle}{\left\| \sum_{i=1}^{t-1} X_i \right\|}$$
where we define $\text{sign}(z) = 1$ if $z \ge 0$ and $\text{sign}(z) = -1$ otherwise.
3. If $\sum_{i=1}^{t-1} X_i = 0$, set $s_t = 0$.

Clearly if $\{X_t\}$ is a random sequence adapted to the filtration $\mathcal{F}_t$, then so is s_t. Now, these s_t have the following interesting property:

Lemma 21 (Cutkosky and Mehta [2021], Lemma 10)**.** *For all t, we have $|s_t| \le \|X_t\|$ and*

$$\left\|\sum_{i=1}^t X_i\right\| \le \left|\sum_{i=1}^t s_i\right| + \sqrt{\max_{i\le t}\|X_i\|^2 + \sum_{i=1}^t \|X_i\|^2}$$

Now, we need to use this result. The key is that if X_t is a MDS, then the s_t will be also. Thus, we can bound the sum of the X_t using the sum of the s_t, which can in turn be bounded by *scalar* martingale concentration bounds. Let us instantiate this using our previous bounds to obtain:

Theorem 19. *Suppose that $\{X_t, \mathcal{F}_t\}$ is a vector-valued martingale difference sequence such that $\mathbb{E}[\|X_t\|^2|\mathcal{F}_{t-1}] \le \sigma_t^2$ and $\|X_t\| \le b_t$ almost everywhere for some sequence $\{\sigma_t, b_t\}$ such that σ_t, b_t is $\mathcal{F}_{t-1}$-measurable. Let $\nu \ge 0$ be an arbitrary constant. Then with probability at least $1-\delta$, for all t we have:*

$$\begin{aligned}\left\|\sum_{i=1}^t X_i\right\| \le 5\sqrt{\sum_{i=1}^t \sigma_i^2 \log\left(\frac{16}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right)+2\right]^2\right)} \\ + 23\max(\nu, \max_{i\le t} b_i)\log\left(\frac{224}{\delta}\left[\log\left(\frac{2\max(\nu,\max_{i\le t} b_i)}{\nu}\right)+2\right]^2\right)\end{aligned}$$

where $[x]_1 = \max(1, x)$.

Proof. Observe from the construction of the s_t sequence that $\{s_t\}$ is a MDS adapted to $\mathcal{F}_t$, and that $\mathbb{E}[s_t^2 \mid \mathcal{F}_{t-1}] \le \sigma_{t-1}^2$ and $|s_t| \le b_{t-1}$. Therefore s_t is $\sigma_t, 2b_t$ subgaussian. Invoking Theorem 18 (with a union bound for a two-sided inequality), with probability at least $1-\delta/2$ we have:

$$\begin{aligned}\left|\sum_{i=1}^t s_i\right| \le 2\sqrt{\sum_{i=1}^t \sigma_i^2 \log\left(\frac{16}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/2\nu^2}\right]_1\right)+2\right]^2\right)} \\ + 16\max(\nu, \max_{i\le t} b_i)\log\left(\frac{112}{\delta}\left[\log\left(\frac{2\max(\nu,\max_{i\le t} b_i)}{\nu}\right)+2\right]^2\right)\end{aligned}$$

Next, observe that $\|X_t\|$ satisfies the conditions of Theorem 20 so that also with probability at least $1-\delta/2$:

$$\begin{aligned}\sum_{i=1}^t \|X_i\|^2 \le 3\sum_{i=1}^t \sigma_i^2 \log\left(\frac{8}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right)+2\right]^2\right) \\ + 20\max(\nu^2, \max_{i\le t} b_i^2)\log\left(\frac{224}{\delta}\left[\log\left(\frac{2\max(\nu,\max_{i\le t} b_i)}{\nu}\right)+1\right]^2\right)\end{aligned}$$

Putting this together with Lemma 21 we have

$$\begin{aligned}\left\|\sum_{i=1}^t X_i\right\| &\le \left|\sum_{i=1}^t s_i\right| + \sqrt{\max_{i\le t-1}\|X_i\|^2 + \sum_{i=1}^t \|X_i\|^2} \\ &\le \left|\sum_{i=1}^t s_i\right| + \sqrt{2\sum_{i=1}^t \|X_i\|^2}\end{aligned}$$

$$
\begin{aligned}
&\leq 2\sqrt{\sum_{i=1}^t \sigma_i^2 \log\left(\frac{16}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/2\nu^2}\right]_1\right)+2\right]^2\right)} \\
&\quad + 16\max(\nu, \max_{i\leq t} b_i)\log\left(\frac{112}{\delta}\left[\log\left(\frac{2\max(\nu,\max_{i\leq t} b_i)}{\nu}\right)+2\right]^2\right) \\
&\quad + \sqrt{6\sum_{i=1}^t \sigma_i^2 \log\left(\frac{8}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right)+2\right]^2\right)} \\
&\quad + \sqrt{40\max(\nu^2, \max_{i\leq t} b_i^2)\log\left(\frac{224}{\delta}\left[\log\left(\frac{2\max(\nu,\max_{i\leq t} b_i)}{\nu}\right)+1\right]^2\right)} \\
&\leq 5\sqrt{\sum_{i=1}^t \sigma_i^2 \log\left(\frac{16}{\delta}\left[\log\left(\left[\sqrt{\sum_{i=1}^t \sigma_i^2/\nu^2}\right]_1\right)+2\right]^2\right)} \\
&\quad + 23\max(\nu, \max_{i\leq t} b_i)\log\left(\frac{224}{\delta}\left[\log\left(\frac{2\max(\nu,\max_{i\leq t} b_i)}{\nu}\right)+2\right]^2\right)
\end{aligned}
$$

□

E.4 Proof of Proposition 17

Proposition 17. *Suppose $\{X_t, \mathcal{F}_t\}$ is a MDS such that $\mathbb{E}[X_t^2|\mathcal{F}_t] \leq \sigma_t^2$ and $|X_t| \leq b_t$ almost everywhere for all t for some sequence of random variable $\{\sigma_t, b_t\}$ such that σ_t, b_t is $\mathcal{F}_{t-1}$-measurable. Then X_t is $(\sigma_t, 2b_t)$ sub-exponential.*

Proof. Suppose $\lambda \leq 1/2b_t$. Then we compute for any $k \geq 2$:

$$
\begin{aligned}
\mathbb{E}\left[\frac{\lambda^k X_t^k}{k!} \mid \mathcal{F}_{t-1}\right] &\leq \frac{\lambda^k b_t^{k-2}}{k!}\mathbb{E}[X_t^2|\mathcal{F}_{t-1}] \\
&\leq \frac{\lambda^k b_t^{k-2}\sigma_t^2}{k!} \\
&\leq \frac{\lambda^2\sigma_t^2}{2^{k-2}k!}
\end{aligned}
$$

Further, since X_t is a MDS, we also have $\mathbb{E}\left[\lambda X_t \mid \mathcal{F}_{t-1}\right] = 0$. Therefore:

$$
\begin{aligned}
\mathbb{E}[\exp(\lambda X_t)|\mathcal{F}_{t-1}] &\leq 1 + \sum_{k=2}^\infty \frac{\lambda^2\sigma_t^2}{2^{k-2}k!} \\
&\leq 1 + \frac{\lambda^2\sigma_t^2}{2}\sum_{i=0}^\infty \frac{1}{2^i} \\
&= 1 + \frac{\lambda^2\sigma_t^2}{2} \\
&\leq \exp(\lambda^2\sigma_t^2/2)
\end{aligned}
$$

where the last line uses the identity $1 + x \leq \exp(x)$. □

E.5 Classical Concentration Bound for MDS

Lemma 22 (Scaled Sub-exponential)**.** *Suppose that $\{X_t, \mathcal{F}_t\}$ is a MDS such that $\mathbb{E}[X_t|\mathcal{F}_t] = 0$, $\mathbb{E}[X_t^2 \mid \mathcal{F}_t] \le \sigma$ and $|X_t| \le b$ almost surely for some fixed σ, b. Let ν be an arbitrary fixed number, then for all t with probability at least $1-\delta$:*

$$\nu X_t \le 2|\nu| b \log\frac{1}{\delta} + |\nu|\sigma\sqrt{2\log\frac{1}{\delta}}$$

Proof. First we have $\mathbb{E}[\nu^2 X_t^2] \le \nu^2\sigma^2, |\nu X_t| \le |\nu|b$ almost surely and $\{\nu X_t, \mathcal{F}_t\}$ is also a MDS. By Proposition 17, νX_t is $(|\nu|\sigma, 2|\nu|b)$ sub-exponential. Use definition 1 and tower rule,

$$\mathbb{E}[\exp(\lambda\nu X_t)] = \mathbb{E}[\mathbb{E}[\exp(\lambda\nu X_t) \mid \mathcal{F}_{t-1}]] = \exp(\lambda^2|\nu|\sigma^2/2)$$

for $\lambda \le 1/(2|\nu|b)$. The above inequality make us returns to the standard result of independent sub-exponential random variable.

$$\mathbb{E}[\exp(\lambda\nu X_t)] = \exp(\lambda^2|\nu|\sigma^2/2)$$

Hence, from the standard sub-exponential tails

$$P[\nu X_t \ge a] \le \begin{cases} \exp(-a^2/2\nu^2\sigma^2) & 0 \le a \le \sigma^2|\nu|/2b, \\ \exp(-a/2|\nu|b) & a > \sigma^2|\nu|/2b \end{cases}$$

Set above quantities as δ and rearrange for a:

$$a = \max\left(2|\nu|b\log\frac{1}{\delta}, \sqrt{2|\nu|^2\sigma^2\log\frac{1}{\delta}}\right)$$

Hence with probability at least $1-\delta$

$$\begin{aligned}\nu X_t &\le \max\left(2|\nu|b\log\frac{1}{\delta}, \sqrt{2|\nu|^2\sigma^2\log\frac{1}{\delta}}\right) \\ &\le 2|\nu|b\log\frac{1}{\delta} + |\nu|\sigma\sqrt{2\log\frac{1}{\delta}}\end{aligned}$$

□

Lemma 23 (Scaled Sub-exponential Sum)**.** *Suppose that $\{X_t, \mathcal{F}_t\}$ is a MDS such that $\mathbb{E}[X_t \mid \mathcal{F}_t] = 0$, $\mathbb{E}[X_t^2 \mid \mathcal{F}_t] \le \sigma$ and $|X_t| \le b$ almost surely. Let ν be an arbitrary fixed number, then for all t with probability at least $1-\delta$:*

$$\sum_{i=1}^{t} \nu X_i \le 2|\nu|b\log\frac{1}{\delta} + |\nu|\sigma\sqrt{2t\log\frac{1}{\delta}}, \quad \forall t$$

Proof.

$$\mathbb{E}[e^{\lambda\sum_{i=1}^{t}\nu X_i}] = \prod_{i=1}^{t}\mathbb{E}[e^{\lambda\nu X_i}]$$

for $|\lambda| \le \frac{1}{2|\nu|b}$, invoke Lemma 22

$$\le \exp(t\lambda^2|\nu|\sigma^2/2)$$

Let $\sigma' = \sigma\sqrt{t}, \sigma' \in \mathbb{R}$, hence we can directly use the concentration bound derived in Lemma 22 to complete the proof. □

Lemma 24 (Sub-exponential Squared)**.** *Suppose $\{X_t, \mathcal{F}_t\}$ is a MDS with $|X_t| \le b$ and $\mathbb{E}[X_t^2 \mid \mathcal{F}_t] \le \sigma^2$ almost surely for some fixed σ, b. Then with probability at least $1-\delta$*

$$\sum_{t=1}^{T} X_t^2 \le \frac{3\sigma^2}{2}T + \frac{5}{3}b^2\log\frac{1}{\delta}$$

Proof. Let $Z_t = X_t^2 - \mathbb{E}[X_t^2]$, $Z_0, \cdots, Z_T$ is a martingale difference sequence adapted to $\mathcal{F}_t$. Also $|Z_t| < X_t^2 \leq b^2$. Also

$$\mathbb{E}[Z_t^2] = \mathbb{E}[|Z_t| \cdot |Z_t|] \leq b^2\,\mathbb{E}[|Z_t|] \leq b^2\,\mathbb{E}[x_t^2] \leq b^2\sigma^2$$

From Freedman's inequality for martingale sequences (see e.g. Tropp [2011], or Lemma 11 in Cutkosky and Mehta [2021] for the form we use here), with probability at least $1-\delta$,

$$\sum_{t=1}^{T} Z_t \leq \frac{2}{3} b^2 \log\frac{1}{\delta} + \sigma b \sqrt{2T \log\frac{1}{\delta}}$$

Rearranging the definition of Z_t, with probability at least $1-\delta$:

$$\begin{aligned}\sum_{t=1}^{T} X_t^2 &= \sum_{t=1}^{T} Z_t + \sum_{t=1}^{T} \mathbb{E}[X_t^2] \\ &\leq \frac{2}{3} b^2 \log\frac{1}{\delta} + \sigma b \sqrt{2T \log\frac{1}{\delta}} + T\sigma^2 \end{aligned} \tag{26}$$

by young's inequality $\sigma b \leq \sigma^2/(2\lambda) + \lambda b^2/2$, set $\lambda = \sqrt{2\log\frac{1}{\delta}/T}$, we complete the proof ☐

E.6 Another Technical Lemma

Lemma 25. *Suppose Z is such that*

$$Z \leq A \log\left(B\left[\log\left(\left[C\sqrt{Z}\right]_1\right) + 2\right]^2\right)$$

for some constants $A, B, C \geq 0$, where $[x]_1 = \max(1, x)$ Then

$$Z \leq 4A \log\left(4\sqrt{B}\log(e + 16C^2 A)\right)$$

Proof. Expanding the logarithms in the given bound on Z, we have:

$$Z \leq A\log(B) + 2A\log(\log([C\sqrt{Z}]_1) + 2)$$

Now, since $a + b \leq \max(2a, 2b)$, we have that either $Z \leq 2A\log(B)$, or

$$Z \leq 4A\log(\log([C\sqrt{Z}]_1) + 2)$$

In the first case, we are done since $2A\log(B) \leq 4A\log\left(4\sqrt{B}\log(e + 16C^2A)\right)$, so let us consider only the second case $Z \leq 4A\log(\log([C\sqrt{Z}]_1) + 2)$. Now, we define the function

$$f(x) = 4A\log(\log([C\sqrt{x}]_1)) + 2) \tag{27}$$

Notice that for all $x \geq A$, we have either $f'(x) = 0$ or $C\sqrt{x} \geq 1$ and

$$f'(x) = \frac{2A}{x\log(C\sqrt{x}) + 2x} < 1$$

so that if $Z_\star$ is any value satisfying $Z_\star \geq A$ and $Z_\star \geq f(Z_\star)$, then we must have $Z \leq Z_\star$: otherwise $f(Z) = f(Z_\star) + \int_{Z_\star}^{Z} f'(z)dz < f(Z_\star) + Z - Z_\star \leq Z$, a contradiction. Let us consider:

$$Z_\star = 4A\log\left[4\log(e + QA)\right]$$

for some to-be-specified $Q \geq 0$. Notice that this $Z_\star$ clearly satisfies $Z_\star \geq 8A\log(2)$. Let us show that $Z_\star \geq f(Z_\star)$.

Again using $x + y \leq 2\max(x, y)$, we have:

$$f(Z_\star) \leq 4A\log\left(\max(4,\ 2\max(C\sqrt{Z_\star}, 1))\right)$$

$$= \max\left(4A\log(4),\ 4A\log(2\log(\max(C\sqrt{Z_\star},1)))\right)$$

Now, if $f(Z_\star) \leq 4A\log(4)$ then we clearly have $f(Z_\star) \leq Z_\star$ as desired. So, let us focus on the case $f(Z_\star) \leq 4A\log(2\log(\max(C\sqrt{Z_\star},1)))$. Next, if $C\sqrt{Z_\star} \leq e$, then we have $f(Z_\star) \leq 4A\log(2) \leq Z_\star$ again as desired. Thus we may further restrict to the case $C\sqrt{Z_\star} \geq e$ so that $\max(C\sqrt{Z_\star},1) = C\sqrt{Z_\star}$ and the bound on $f(Z_\star)$ is $4A\log(2\log(C\sqrt{Z_\star}))$. Plugging in our expression for $Z_\star$:

$$f(Z_\star) \leq 4A\log\left[2\log\left[2C\sqrt{A\log\left[4\log(e+QA)\right]}\right]\right]$$

Comparing with the expression for $Z_\star$, we see that to establish $Z_\star \geq f(Z_\star)$, it suffices to show:

$$(e+QA)^2 \geq 2C\sqrt{A\log\left[4\log(e+QA)\right]}$$

Now, using $\log(x) \leq x$ twice:

$$2C\sqrt{A\log(4\log(e+QA))} \leq 4C\sqrt{A}\sqrt{(e+QA)} \tag{28}$$
$$= \sqrt{16C^2A}\sqrt{e+QA} \tag{29}$$
$$\leq \sqrt{e+16C^2A}\sqrt{e+QA} \tag{30}$$

if we set $Q = 16C^2$, we will have:

$$= e+QA \tag{31}$$
$$\leq (e+QA)^2 \tag{32}$$

Thus, by setting $Q = 16C^2$, we will have $Z_\star \geq f(Z_\star)$ and so we have $f(Z_\star) \leq Z_\star$, which implies

$$\begin{aligned}Z &\leq Z_\star\\ &= 4A\log\left[4\log(e+16C^2A)\right]\\ &\leq 4A\log\left(4\sqrt{B}\log(e+16C^2A)\right)\end{aligned}$$

as desired since $B \geq 1$. □