Second Order Path Variationals in Non-Stationary Online Learning

Dheeraj Baby
dheeraj@cs.ucsb.edu

Yu-Xiang Wang
yuxiangw@cs.ucsb.edu

Department of Computer Science
University of California Santa Barbara

Abstract

We consider the problem of *universal* dynamic regret minimization under exp-concave and smooth losses. We show that appropriately designed Strongly Adaptive algorithms achieve a dynamic regret of $\tilde{O}(d^2 n^{1/5}[\mathcal{TV}_1(\boldsymbol{w}_{1:n})]^{2/5} \vee d^2)$, where n is the time horizon and $\mathcal{TV}_1(\boldsymbol{w}_{1:n})$ a path variational based on *second order differences* of the comparator sequence. Such a path variational naturally encodes comparator sequences that are piece-wise linear – a powerful family that tracks a variety of non-stationarity patterns in practice (Kim et al., 2009). The aforementioned dynamic regret is shown to be optimal modulo dimension dependencies and poly-logarithmic factors of n. To the best of our knowledge, this path variational has not been studied in the non-stochastic online learning literature before. Our proof techniques rely on analysing the KKT conditions of the offline oracle and requires several non-trivial generalizations of the ideas in Baby and Wang (2021) where the latter work only implies an $\tilde{O}(n^{1/3})$ regret for the current problem.

1 INTRODUCTION

Online Convex Optimization (OCO) (Zinkevich, 2003; Hazan, 2016) is a widely studied setup in machine learning that has witnessed a myriad of influential applications such as time series forecasting, building recommendation engines etc. In this setting, a learner plays an iterative game with an adversary that last for n rounds. In each round $t \in [n] := \{1, \ldots, n\}$, the learner makes a decision $\boldsymbol{p}_t$ that belongs to a *decision space* $\mathcal{D} \subset \mathbb{R}^d$. Then a convex loss *loss function* $f_t : \mathbb{R}^d \to \mathbb{R}$ is revealed by the adversary. The learner suffers a cost of $f(\boldsymbol{p}_t)$ at round t for making its decision. Now, given a *benchmark space* of decisions $\mathcal{W} \subseteq \mathcal{D}$, we aim to study learners that can control its *dynamic regret* against *any* sequence of comparators from the benchmark:

$$R_n(\boldsymbol{w}_{1:n}) := \sum_{t=1}^{n} f_t(\boldsymbol{p}_t) - f_t(\boldsymbol{w}_t), \qquad (1)$$

where we abbreviate the *comparator sequence* $\boldsymbol{w}_{1:n} := \boldsymbol{w}_1, \ldots, \boldsymbol{w}_n$ where each $\boldsymbol{w}_t \in \mathcal{W}$. This is known to be a good metric for characterizing the performance of a learner in non-stationary environments (Zinkevich, 2003; Zhang et al., 2018a; Cutkosky, 2020). The quantity in Eq.(1) is also sometimes referred as *universal dynamic regret* (Zhang et al., 2018a) because we do not impose any constraints on the comparator sequence $\boldsymbol{w}_{1:n}$ except that each sequence member must belong to the benchmark set $\mathcal{W}$. This is a different and more powerful way of tackling distribution-shifts than other methods that model the environment explicitly (e.g., Besbes et al., 2015; Baby and Wang, 2020).

Let us illustrate the point in a weather forecasting application in which $f_t(\boldsymbol{w}_t) = \ell(y_t, x_t^T \boldsymbol{w}_t)$ where x_t is a feature vector (e.g., humidity and temperature at Day t), y_t is the actual precipitation of the next day and ℓ is a loss function. The underlying distribution of $y_t|x_t$ is determined by nature and could drift over time due to unobserved variables such as climate change. The approach of Besbes et al. (2015); Baby and Wang (2020) would be to assume a model, e.g., $y_t = x_t^T w_t^* +$ noise and control the regret against $w_{1:n}^*$ in terms of the variation of the *true* regression coefficients over time. In contrast, a *universal dynamic regret* approach will not make any assumption about the world, but instead will compete with the best time-varying sequence of comparators that can be chosen *in hindsight*. In the case when the model is correct, we can choose the comparators to be $w_{1:n}^*$; otherwise, we can compete with the best sequence of linear predictors that optimally balances the bias and variance.

A bound on $R_n(\boldsymbol{w}_{1:n})$ is usually expressed in terms of the time horizon n and a *path variation* that captures

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.

the smoothness of the comparator sequence $\boldsymbol{w}_{1:n}$. Some examples of such path variationals include $P(\boldsymbol{w}_{1:n}) = \sum_{t=1}^{n-1} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_2$ (Zinkevich, 2003) and more recently $\mathcal{TV}(\boldsymbol{w}_{1:n}) = \sum_{t=1}^{n-1} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_1$ (Baby and Wang, 2021).

Comparator as a discretized function. When we view the sequence of comparators as a function of time, it is natural to describe them as a discretization of (continuous-time) functions residing in some non-parametric function classes. We now proceed to expand upon this idea. For a function $f : [0,1] \to \mathbb{R}$ that is k times (weakly) differentiable, define the Total Variation (TV) of its k^{th} derivative $f^{(k)}$ to be:

$$TV(f^{(k)}) := \sup_{0=z_1<\ldots<z_{N+1}=1} \sum_{i=1}^{N} |f^{(k)}(z_{i+1}) - f^{(k)}(z_i)|. \tag{2}$$

If the function has $k+1$ continuous derivatives, then $TV(f^{(k)})$ is equivalent to $\int_0^1 |f^{(k+1)}(x)|dx$. Given $n, C_n > 0$ one may define the function space:

$$\mathcal{F}_k(C_n) := \{f : [0,1] \to \mathbb{R} | TV(f^{(k)}) \le C_n\}.$$

This space is known to contain functions that have a piece-wise degree k polynomial structure (Tibshirani, 2014). We can generate interesting comparator sequence families by discretizing such function spaces. First we fix some notations. For a sequence of vectors $\boldsymbol{v}_{1:\ell} := \boldsymbol{v}_1, \ldots, \boldsymbol{v}_\ell$, define the first order discrete difference operation $D\boldsymbol{v} := \boldsymbol{v}_2 - \boldsymbol{v}_1, \ldots, \boldsymbol{v}_\ell - \boldsymbol{v}_{\ell-1}$. For any positive integer k, the k^{th} order discrete difference – D^k – of a sequence is obtained via applying the operation D for k times. For a sequence $\boldsymbol{v}_{1:\ell}$, we define $\|\boldsymbol{v}_{1:\ell}\|_1 = \sum_{j=1}^{\ell} \|\boldsymbol{v}_j\|_1$

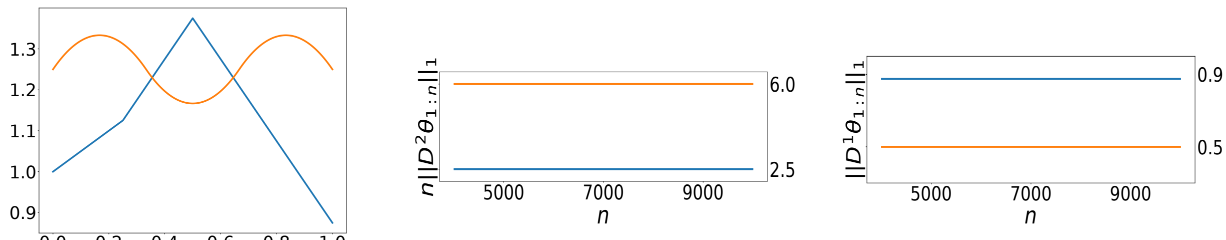
The higher order TV distance. Next, we define a path length which is the discrete analogue of $TV(f^{(k)})$ in Eq.(2) as follows:

$$\mathcal{TV}_k(\boldsymbol{w}_{1:n}) := n^k \|D^{k+1} \boldsymbol{w}_{1:n}\|_1. \tag{3}$$

It is a **common misconception** that constraining $\mathcal{TV}_k$ can be alarmingly restrictive due to the presence of the multiplicative factor of n^k in its definition. To clarify that this is not the case, we observe that the multiplicative factor of n^k arises naturally as a consequence of the Riemann approximation of the continuous Total Variation displayed in Eq.(2) at a resolution $1/n$. This observation leads to the following scheme of generating sequences with $\mathcal{TV}_1(\boldsymbol{w}_{1:n}) = O(C_n)$ for any given number C_n: Along any coordinate $j \in [d]$, generate the sequence $\boldsymbol{w}_1[j], \ldots, \boldsymbol{w}_n[j]$ via sampling a function $f_j(x) \in \mathcal{F}_k(C_{n,j})$ at points $x = i/n$ for $i \in [n]$ with the property that $\sum_{j=1}^{d} C_{n,j} = C_n$. For example, considering the case of $k=1$ and $d=1$, if $TV(f^{(1)})$ is $O(n^\alpha)$ for some $\alpha \ge 0$, then $\mathcal{TV}_1(\boldsymbol{w}_{1:n}) := n\|D^2 \boldsymbol{w}_{1:n}\|_1$ is also $O(n^\alpha)$ despite the multiplicative factor of n appearing in the quantity $\mathcal{TV}_1(\boldsymbol{w}_{1:n})$. A demonstration of this phenomenon for $\alpha = 0$ is displayed in Fig.1.

Why is this useful? In this paper, we focus on comparators with bounded $\mathcal{TV}_1$ distance (i.e Eq.(3) with $k=1$). Our goal will be to bound the dynamic regret Eq.(1) against $\boldsymbol{w}_{1:n}$ as a function of n and $\mathcal{TV}_1(\boldsymbol{w}_{1:n})$. Due to the presence of second order differencing operation in the definition of $\mathcal{TV}_1$, this path length is ideal to capture the variation incurred by comparators with piece-wise linear structure across each coordinate (see Definition 1). The points where the sequence transition from one linear structure to other can be interpreted as abrupt changes or events in the underlying comparator dynamics. The value of $\mathcal{TV}_1(\boldsymbol{w}_{1:n})$ simultaneously captures the sparsity (due to the presence of L_1 norm in Eq.(3)) and intensity of such changes. Many real world time series data are known to contain piece-wise linear trends. See for example Fig.2 or Kim et al. (2009) for more examples. Hence controlling the dynamic regret in terms of $\mathcal{TV}_1(\boldsymbol{w}_{1:n})$ has *significant practical value*.

Fast rate phenomenon. Path lengths of the form $\mathcal{TV}_1$ (or more generally $\mathcal{TV}_k$) have gained significant attention and have been the subject of extensive study in the stochastic non-parametric regression community for over two decades (van de Geer, 1990; Donoho and Johnstone, 1998; Kim et al., 2009; Tibshirani, 2014; Wang et al., 2016). These works aim to estimate an unknown scalar (i.e $d=1$) sequence $w_{1:n}$ from n noisy observations $y_t = w_t + \mathcal{N}(0, \sigma^2)$ in an offline setting. They propose algorithms that produce estimates $\hat{\boldsymbol{w}}_{1:n}$ such that the expected total squared error $\sum_{t=1}^{n} E[(\hat{w}_t - w_t)^2]$ is controlled. In particular, an estimation rate of $\tilde{O}(n^{1/5}[\mathcal{TV}_1(\boldsymbol{w}_{1:n})]^{2/5})$ is shown to be attainable for the squared loss ($\tilde{O}(\cdot)$ hides poly-logarithmic factors of n). On the other hand, squared error losses are also exp-concave in a compact domain. Baby and Wang (2021) proposes algorithms for controlling dynamic regret under exp-concave losses. Applying their algorithm will lead to an estimation error of $\tilde{O}(n^{1/3}[\mathcal{TV}_0(\boldsymbol{w}_{1:n})]^{2/3})$. However, there can be scenarios where the rate of $\tilde{O}(n^{1/5}[\mathcal{TV}_1(\boldsymbol{w}_{1:n})]^{2/5})$ can be faster than the rate of $\tilde{O}(n^{1/3}[\mathcal{TV}_0(\boldsymbol{w}_{1:n})]^{2/3})$. For instance, consider the canonical (and practically relevant) example in Fig.1. Let $\boldsymbol{w}_{1:n}$ be generated by discretizing the function in the left panel at a resolution $1/n$. We can see that $\mathcal{TV}_0(\boldsymbol{w}_{1:n}) \le 1 = O(1)$ and $\mathcal{TV}_1(\boldsymbol{w}_{1:n}) \approx 2.5 = O(1)$ (for $n \ge 10$). Here, the aforementioned results from stochastic non-parametric regression can yield a rate of $\tilde{O}(n^{1/5})$ while existing state-of-the-art results from adversarial online learning can only lead to $O(n^{1/3})$ rate of estimation. We refer the reader to Remark 7 for a discussion about more such examples.



(a) Two weakly differentiable functions

(b) discrete TV^1 distance

(c) discrete TV^0 distance

Figure 1: *A TV^1 bounded comparator sequence $\boldsymbol{w}_{1:n}$ can be obtained by discretizing the weakly differentiable functions displayed in Fig(a) at points i/n, $i \in [n]$. In Fig(b), we plot the TV^1 distance (which is equal to $n\|D^2\boldsymbol{w}_{1:n}\|_1$ by definition) of the generated sequence for various sequence lengths n. Blue (orange) curve in Fig(b) corresponds to the statistics of the discretization of the blue (orange) curve in Fig(a). As n increases the discrete TV^1 distance converges to a constant value given by the continuous TV^1 distance of the functions in Fig(a). In Fig(c) we plot the TV^0 distance of the discretizations. Thus in this example, we see that both $\|D^1\boldsymbol{w}_{1:n}\|_1$ and $n\|D^2\boldsymbol{w}_{1:n}\|_1$ are $O(1)$ as n grows. Since TV^1 distance of the sequences is $O(1)$, the algorithm that we propose in Section 3 is able to obtain the faster dynamic regret rate of $\tilde{O}(n^{1/5})$ as opposed to the rate of $\tilde{O}(n^{1/3})$ obtainable from Baby and Wang (2021) for sequences with bounded TV^0 distance. Furthermore, the functions in Fig(a) are reminiscent to the real-life trends observed in Fig.2.*

Central question and summary of results. A natural question that we ask here is:

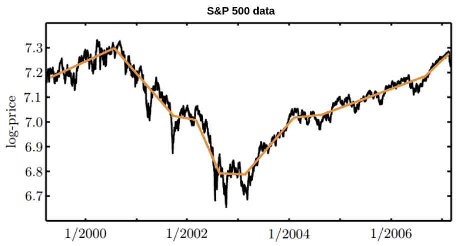
> Can we attain a universal dynamic regret (Eq.(1)) of $\tilde{O}^*(n^{1/5}[\mathcal{TV}_1(\boldsymbol{w}_{1:n})]^{2/5})$ when the losses are exp-concave without imposing any stochastic assumptions?

Here O^* hides dimension dependencies. We remark that the rate of $\tilde{O}(n^{1/5}[\mathcal{TV}_1(\boldsymbol{w}_{1:n})]^{2/5})$ is faster than $\tilde{O}(n^{1/3}[\mathcal{TV}_0(\boldsymbol{w}_{1:n})]^{2/3})$ iff $\mathcal{TV}_1(\boldsymbol{w}_{1:n}) = O\left(n^{1/3}[\mathcal{TV}_0(\boldsymbol{w}_{1:n})]^{5/3}\right)$. In what follows, we refer to this regime as the **low $\boldsymbol{TV}_1$ regime**. We emphasize that this regime is not too restrictive as many different examples can be encompassed by it (see for eg. Fig.1 and Remark 7). A starting point in answering our central question is to observe that a sequence will have low $\mathcal{TV}_1$ distance if it exhibits a piece-wise linear structure across each coordinate and the number of linear sections (or kinks) is sparse. A sequence that is linear across each coordinate within some interval can be perfectly described using a *fixed* vector $\boldsymbol{u} \in \mathbb{R}^{2d}$ where $\boldsymbol{u}[2k-1:2k] \in \mathbb{R}^2$ specifies the slope and intercept along coordinate $k \in [d]$. We will call such $\boldsymbol{u}$ to be a *linear predictor*. If an algorithm guarantees that its *static regret* against fixed linear predictors within *any* interval is controlled, one can hope to perform nearly as well as the comparator sequence with low enough $\mathcal{TV}_1$. This is precisely an application of Strongly Adaptive algorithms (Hazan and Seshadhri, 2007; Daniely et al., 2015; Adamskiy et al., 2016; Cutkosky, 2020) which aim to control their static regret in any interval and hence we can use them off-the-shelf to achieve our goal. We refer the reader to Section 3 for more details. Below, we briefly summarize our contributions:

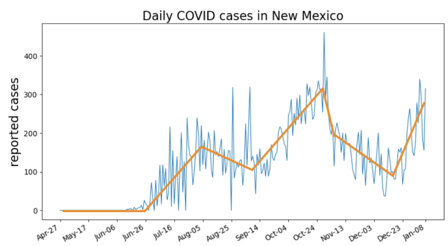
- We show that by using appropriate Strongly Adaptive algorithms, one can attain the (near) *optimal* universal dynamic regret rate of $\tilde{O}^*\left(\min\{n^{1/5}[\mathcal{TV}_1(\boldsymbol{w}_{1:n})]^{2/5}, n^{1/3}[\mathcal{TV}_0(\boldsymbol{w}_{1:n})]^{2/3}\}\right)$ (Theorem 3 and Proposition 5; $a \vee b = \max\{a,b\}$) whenever the comparators $\boldsymbol{w}_{1:n} \in \mathcal{TV}^{(1)}(C_n)$ and the losses are exp-concave and gradient smooth (see Section 4 for the list of Assumptions and associated definitions). Further this rate is attained *without prior knowledge* of the path lengths $\mathcal{TV}_1(\boldsymbol{w}_{1:n})$ and $\mathcal{TV}_0(\boldsymbol{w}_{1:n})$.
- To the best of our knowledge, we are the *first* to introduce path variationals based on second order differences to the setting of *adversarial* online learning. We show how to import the *fast rate* phenomenon observed in stochastic non-parametric regression problem under squared loss into the problem of controlling *universal* dynamic regret under general exp-concave losses with no stochastic assumptions.

Even though in our proofs, we analyse the KKT conditions of an offline optimization problem akin to the spirit of Baby and Wang (2021), this similarity is only superficial. The offline optimization problem analysed in this work is different from what is considered in Baby and Wang (2021). So the KKT conditions, regret decomposition and the proof strategies we use are also different. Further, we introduce several new non-trivial ideas and generalizations (see Section 4.1 and Appendix B) while exploiting the smoothness of sequences with low $\mathcal{TV}_1$ distance to attain the challenging goal of deriving *faster* (in comparison to Baby and Wang (2021)) *universal* dynamic regret rates.

Before we end this section, we briefly describe how the



(a) S&P500 stock prices



(b) Daily COVID cases

Figure 2: *Fig.(a) displays S&P500 data and Fig.(b) displays Daily COVID cases reported in the state of New Mexico, USA. In both scenarios the underlying trend (obtained via an L1 Trend Filter (Kim et al., 2009)) exhibits a weakly differentiable piece-wise linear structure (orange).*

present work provide a new direction in the research thread of dynamic regret minimization.

Notes on general outlook and potential impact. Any meaningful dynamic regret bound has to be parameterized by the particular comparator sequence to avoid a trivial linear regret, i.e.,

$$\sum_{t=1}^{n} f_t(\boldsymbol{p}_t) - f_t(\boldsymbol{w}_t) \leq \mathrm{DynamicRegretBound}(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n).$$

Almost all existing dynamic regret bounds are parameterized by the movement costs — a functional of consecutive differences of the comparator sequence. In particular, we can define $V_{p,q} := \sum_{t=2}^{n} \|\boldsymbol{w}_t - \boldsymbol{w}_{t-1}\|_q^p$. This includes almost all existing variations, e.g., pathlength ($p = 1, q = 2$), square pathlength ($p = 2, q = 2$), total variation ($p = 1, q = 1$), number of changes ($p > 0, q = 0$) and so on. The optimal universal dynamic regret for each functional under different loss classes are now well-known.

While it appears to be a complete story if we roll back to the general problem, there should be many other ways we can parameterize the $\mathrm{DynamicRegretBound}(\boldsymbol{w}_1, ..., \boldsymbol{w}_n)$ by exploiting other structures of the comparator sequence than via the movement costs vector. This work can be thought of as the first one to resort to this idea. We reveal that one can attain faster rates by exploiting the smoothness / regularity (in terms of piece-wise polynomial structures) in the comparator sequence. As an example, when the first derivative (aka the $\mathcal{TV}_1$ distance) of the sequence has bounded variation, we show that the dynamic regret improves to $O(n^{1/5})$. This idea traces back to the nonparametric regression literature, where the higher order smoothness of functions are often used. It is our hope that this work can inspire further collaborations between researchers in these two (mostly disparate) communities of online learning and offline non-parametric regression.

We remark that our work only touches the surface of the idea of characterizing dynamic regret via interesting smoothness metrics of the comparator sequence. Indeed, there are other interesting functionals of the comparator sequence that we can exploit, e.g., periodicity, smoothness in an appropriately transformed domain and so on. We believe this is an interesting future direction for researchers working in dynamic regret minimisation. The present work takes only the first steps towards realising this bigger goal.

2 RELATED WORK

Here we recall the most relevant works. The work of Baby and Wang (2020) aims at controlling Eq.(1) under squared error when noisy realizations of a $\mathcal{TV}_1$ bounded sequence is revealed sequentially. For this setting, they propose an algorithm namely AdaVAW that combines Vovk-Azoury-Warmuth forecaster with wavelet denoising which relies strongly on the iid noise assumption and losses being squared error. The absence of such stochastic assumptions and handling general exp-concave losses in our setting poses a significant challenge in controlling the dynamic regret. Overall, we can conclude that results in the current paper dominates that of AdaVAW for TV order $k = 1$. As mentioned in Section 1, the work of Baby and Wang (2021) fails to attain optimal regret rate for the current problem. We refer the readers to Appendix F for a detailed description on why the analysis of Baby and Wang (2021) fails to attain optimal regret rate in our setting where the comparators has low $\mathcal{TV}_1$ distance. Baby et al. (2021) reported experiments where they use a Strongly Adaptive algorithm for competing against best linear predictor in each time window for the task of forecasting COVID-19 cases. This method was shown to empirically out-perform state-of-the-art trend forecasting strategies. However, they didn't provide analysis for this strategy while our work supplements it with necessary theoretical grounding albeit with a slightly different Strongly Adaptive algorithm. Apart from these works, there is a rich body of literature on dynamic regret minimization such as (Jadbabaie et al., 2015; Yang et al., 2016; Mokhtari et al., 2016; Chen et al., 2018; Zhang et al., 2018a,b; Yuan and Lamperski, 2020; Goel and Wierman, 2019; Baby and Wang, 2019; Zhao et al., 2020; Zhao and Zhang, 2021; Zhao et al., 2022; Baby and Wang, 2022a; Jacobsen and Cutkosky, 2022; Baby and Wang, 2022b; Zhang et al., 2023). However, to the best of our knowledge none of these works are known to attain the optimal dynamic regret rate for our setting. An elaborate literature survey is deferred to Appendix A.

3 THE ALGORITHM

FLH-SIONS: inputs: exp-concavity factor σ and n SIONS base learners $E^1, \ldots, E^n$ initialized with parameters $\epsilon = 2$, $\eta = \sigma$ and $C = 20$. (see Fig. 4)

1. For each t, $v_t = (v_t^{(1)}, \ldots, v_t^{(t)})$ is a probability vector in $\mathbb{R}^t$. Initialize $v_1^{(1)} = 1$.
2. For any SIONS expert E_j with $j \leq t$, define $\boldsymbol{x}_j^{(t)} = [1, t-j+1]^T$ to be given to E_j at time t before making its prediction $E_j(t) \in \mathbb{R}^d$.
3. In round t, set $\forall j \leq t$, $\boldsymbol{y}_t^j \leftarrow E_j(t)$ (the prediction of the j^{th} base learner at time t). Play $\boldsymbol{p}_t = \sum_{j=1}^t v_t^{(j)} \boldsymbol{y}_t^{(j)}$.
4. After receiving f_t, set $\hat{v}_{t+1}^{(t+1)} = 0$ and perform update for $1 \leq i \leq t$:
$$\hat{v}_{t+1}^{(i)} = \frac{v_t^{(i)} e^{-\sigma f_t(\boldsymbol{x}_t^{(i)})}}{\sum_{j=1}^t v_t^{(j)} e^{-\sigma f_t(\boldsymbol{x}_t^{(j)})}}$$
5. Addition step - Set $v_{t+1}^{(t+1)}$ to $1/(t+1)$ and for $i \neq t+1$:
$$v_{t+1}^{(i)} = (1-(t+1)^{-1})\hat{v}_{t+1}^{(i)}$$

Figure 3: FLH algorithm of Hazan and Seshadhri (2007) with SIONS (see Fig.4) base experts

SIONS: inputs: exp-concavity factor η, $\epsilon > 0$ and $C > 0$.

1. For any round t, we define $\tilde{f}_t(\boldsymbol{v}) = f_j(\boldsymbol{x}_t^T \boldsymbol{v}[1:2], \boldsymbol{x}_t^T \boldsymbol{v}[3:4], \ldots, \boldsymbol{x}_t^T \boldsymbol{v}[2k-1:2k])$ for any vector $\boldsymbol{v} \in \mathbb{R}^{2d}$.
2. At round $t+1$:
   (a) Receive co-variate $\boldsymbol{x}_{t+1} \in \mathbb{R}^2$.
   (b) Let $\mathcal{K}_{t+1} = \{\boldsymbol{w} \in \mathbb{R}^{2d} : |\boldsymbol{x}_{t+1}^T \boldsymbol{w}[2k-1:2k]| \leq C$ for all $k \in [d]\}$.
   (c) Let $\boldsymbol{A}_t = \epsilon \boldsymbol{I}_{2d} + \eta \sum_{j=1}^t \nabla \tilde{f}_j(\boldsymbol{v}_j) \nabla \tilde{f}_j(\boldsymbol{v}_j)^T$.
   (d) Let $\boldsymbol{u}_{t+1} = \boldsymbol{v}_t - \boldsymbol{A}_t^{-1} \nabla \tilde{f}_t(\boldsymbol{v}_t)$.
   (e) Let $\boldsymbol{v}_{t+1} = \text{argmin}_{\boldsymbol{w} \in \mathcal{K}_{t+1}} \|\boldsymbol{w} - \boldsymbol{u}_{t+1}\|_{\boldsymbol{A}_t}$.
   (f) Play $\boldsymbol{w}_{t+1} \in \mathbb{R}^d$ such that $\boldsymbol{w}_{t+1}[k] = \boldsymbol{x}_{t+1}^T \boldsymbol{v}_{t+1}[2k-1:2k]$ for all $k \in [d]$.

Figure 4: An instance of SIONS algorithm from Luo et al. (2016).

In this section, we formally describe the main algorithm FLH-SIONS (Follow the Leading History-Scale Invariant Online Newton Step) in Fig.3 and provide intuition on why it can favorably control the dynamic regret against $\mathcal{TV}_1$ bounded comparators. For the sake of simplicity, we capture the intuition in a uni-variate setting where the comparators $w_t \in \mathcal{W} \subset \mathbb{R}$ for all $t \in [n]$.

Definition 1. *Within an interval $[a, b]$, we say that the comparator $w_{a:b}$ is a* linear signal *or assumes a* linear structure *if the slope $w_{t+1} - w_t$ is constant for all $t \in [a, b-1]$.*

As described in Section 1, we are interested in competing against comparator sequences $w_{1:n}$ that have a piecewise linear structure (across each coordinate in multi-dimensions). The durations / intervals of $[n]$ where the comparator is a fixed linear signal is unknown to the learner. Suppose that an ideal oracle provides us with the exact locations of these intervals of $[n]$. Consider an interval $[a, b]$ provided by the oracle where the comparator has a fixed linear structure given by $w_t = \boldsymbol{\mu}^T \boldsymbol{x}_a^{(t)}$ for the *co-variates* $\boldsymbol{x}_a^{(t)} := [1, t-a+1]^T$ and $\boldsymbol{\mu}$ such that $|w_t|$ is $O(1)$ bounded for all $t \in [a, b]$. An effective strategy for the learner is to deploy an online algorithm E_a that starts from time a such that within the interval $[a, b]$ its regret:

$$R_{[a,b]}(\boldsymbol{\mu}) := \sum_{t=a}^{b} f_t\left(E_a(t)\right) - f_t(\boldsymbol{\mu}^T \boldsymbol{x}_a^{(t)})$$

is controlled. Here $E_a(t)$ is the predictions of the algorithm E_a at time t. Under exp-concave losses, an $O(\log n)$ bound on the above regret can be achieved by the SIONS algorithm (Fig.4) from (Luo et al. (2016), Theorem 2) run with co-variates $\boldsymbol{x}_a^{(t)}$.

In practice, the locations of such ideal intervals are unknown to us. So we maintain a pool of n base SIONS experts in Fig.3 where the expert E_τ starts at time τ with the monomial co-variate $\boldsymbol{x}_\tau^{(t)} = [1, t-\tau+1]^T$ for all $t \geq \tau$. The adaptive regret guarantee of FLH with exp-concave losses (due to Hazan and Seshadhri (2007), Theorem 3.2) keeps the regret wrt *any* base expert to be small. In particular, FLH-SIONS satisfies that

$$\sum_{t=\tau}^{j} f_t(p_t) - f_t\left(E_\tau(t)\right) = O(\log n),$$

where p_t are the predictions of FLH-SIONS and $j \geq \tau$ for *any* $\tau \in [n]$. Hence for the interval $[a, b]$ given by the ideal oracle, it follows that

$$\begin{aligned}\sum_{t=a}^{b} f_t(p_t) - f_t(\boldsymbol{\mu}^T \boldsymbol{x}_a^{(t)}) &\leq \sum_{t=a}^{b} f_t\left(E_a(t)\right) - f_t(\boldsymbol{\mu}^T \boldsymbol{x}_a^{(t)}) \\ &\quad + O(\log n) \\ &= R_{[a,b]}(\boldsymbol{\mu}) + O(\log n) \\ &= O(\log n), \end{aligned} \tag{4}$$

where in the last equation, we appealed to the logarithmic static regret of SIONS from Luo et al. (2016). As a minor technical remark, we note that the original results of Luo et al. (2016) assume that the losses are of the form $\tilde{f}_j(\boldsymbol{w}) = f_j(\boldsymbol{x}_j^T \boldsymbol{w})$ for a uni-variate function f_j. However, we show in Lemma 32 (in Appendix) that their regret bounds can be straightforwardly extended to handle multi-variate losses f_j as in Line 1 of Fig.4 which is useful in our multi-dimensional setup.

Thus ultimately, the regret of the FLH-SIONS procedure is well controlled within each interval provided by the ideal oracle, thus allowing us to be competent against the piece-wise linear comparator. We remark that while both FLH and SIONS are well-known existing algorithms, our use of them with monomial co-variates is new. Our dynamic regret analysis is new too, which uncovers previously unknown properties of a particular combination of these existing algorithmic components using novel proof techniques.

4 MAIN RESULTS

In this section, we explain the assumptions used and the main results of this paper. Then we provide a brief proof sketch for Theorem 3 in a uni-variate setting highlighting the technical challenges overcome along the way. The case of multiple dimensions is handled by constructing suitable reductions that will allow us to re-use much of the analytical machinery developed for the case of uni-variate setting. For the sake of clarity, we present a detailed overview of our proof strategy in Appendix B. The following are the assumptions made.

A1. For all $t \in [n]$, the comparators $\boldsymbol{w}_t$ belongs to a given benchmark space $\mathcal{W} \subset \mathbb{R}^d$. Further we have $\mathcal{W} \subseteq [-1, 1]^d$.

A2. The loss function $f_t : \mathbb{R}^d \to \mathbb{R}$ revealed at time t is 1-Lipschitz in $\|\cdot\|_2$ norm over the interval $[-20, 20]^d$.

A3. The losses f_t are 1-gradient Lipschitz over the interval $[-20, 20]^d$. This implies that $f_t(\boldsymbol{y}) \le f_t(\boldsymbol{x}) + \nabla f_t(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{1}{2}\|(\boldsymbol{y} - \boldsymbol{x})\|_2^2$ for all $\boldsymbol{x}, \boldsymbol{y} \in [-20, 20]^d$.

A4. The losses f_t are σ exp-concave over $[-20, 20]^d$. This implies that $f_t(\boldsymbol{y}) \ge f_t(\boldsymbol{x}) + \nabla f_t(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x}) + \frac{\sigma}{2}\left(\nabla f_t(\boldsymbol{x})^T(\boldsymbol{y} - \boldsymbol{x})\right)^2$ for all $\boldsymbol{x}, \boldsymbol{y} \in [-20, 20]^d$.

Assumptions A3 and A4 ensure the smoothness and curvature of the losses which we crucially rely to derive fast regret rates. Assumptions about Lipschitzness as in A2 are usually standard in online learning. In assumption A1 we consider comparators that belong to an interval that is smaller than the intervals in other assumptions. This is due to the fact that we allow our algorithms to be improper in the sense that the decisions of the algorithm may lie outside the benchmark space $\mathcal{W}$.

We start with a lower bound on the dynamic regret (Eq.(1)) which is obtained by adapting the arguments in Donoho and Johnstone (1998) to the case of bounded sequences as in Assumption A1. See Appendix E for a proof.

Proposition 2. *Under Assumptions A1-A4, any online algorithm necessarily suffers* $\sup_{\boldsymbol{w}_{1:n} \text{ with } \mathcal{TV}_1(\boldsymbol{w}_{1:n}) \le C_n} R_n(\boldsymbol{w}_{1:n}) = \Omega(d^{3/5} n^{1/5} C_n^{2/5} \vee d)$.

We have the following guarantee for FLH-SIONS.

Theorem 3. *Let $\boldsymbol{p}_t$ be the predictions of FLH-SIONS algorithm with parameters $\epsilon = 2$, $C = 20$ and exp-concavity factor σ. Under Assumptions A1-A4, we have that,*

$$\sum_{t=1}^{n} f_t(\boldsymbol{p}_t) - f_t(\boldsymbol{w}_t) = \tilde{O}(d^2 n^{1/5} [\mathcal{TV}_1(\boldsymbol{w}_{1:n})]^{2/5} \vee d^2),$$

where $\tilde{O}$ hides poly-logarithmic factors of n and $a \vee b = \max\{a, b\}$.

Remark 4. *Compared with the lower bound in Proposition 2, we conclude that the regret rate of the above theorem is optimal modulo factors of d and $\log n$.*

Proposition 5. *It can be shown that the same algorithm FLH-SIONS under the setting of Theorem 3 enjoys a regret rate of $\tilde{O}(n^{1/3}[\mathcal{TV}_0(\boldsymbol{w}_{1:n})]^{2/3})$ as well. This result is a straight-forward consequence of the arguments in Baby and Wang (2021) and summarized in Appendix D. When combined with Theorem 3 we conclude that under Assumptions A1-A4, FLH-SIONS attains an adaptive guarantee of*

$$\sum_{t=1}^{n} f_t(\boldsymbol{p}_t) - f_t(\boldsymbol{w}_t) = \tilde{O}\left(d^2 \left(\left(n^{1/5}[\mathcal{TV}_1(\boldsymbol{w}_{1:n})]^{2/5}\right) \wedge \left(n^{1/3}[\mathcal{TV}_0(\boldsymbol{w}_{1:n})]^{2/3}\right) \right) \vee d^2 \right),$$

for any comparator sequence $\boldsymbol{w}_{1:n}$. Here $\tilde{O}$ hides poly-logarithmic factors of n, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$.

From the above Proposition, we see that FLH-SIONS has the nice property of safe-guarding the regret in case the $\mathcal{TV}_1(\boldsymbol{w}_{1:n})$ distance of the comparator doesn't fall in the low $\mathcal{TV}_1$ regime defined in Section 1.

Remark 6. *We note that the upper bound in Proposition 5 does not contradict the lower bound in Proposition 2. The lower bound holds in a worst case sense while the upper bound in Proposition 5 is instance-dependant and can sometimes be faster than the worst case rate in Proposition 2. Indeed, for the hard comparator sequence $\boldsymbol{w}_{1:n}$ we construct in Appendix E, the rates $n^{1/5}[\mathcal{TV}_1(\boldsymbol{w}_{1:n})]^{2/5}$ and $n^{1/3}[\mathcal{TV}_0(\boldsymbol{w}_{1:n})]^{2/3}$ are of the same order.*

Remark 7. *We can construct many examples of sequences $\boldsymbol{w}_{1:n}$ that fall in the low $\mathcal{TV}_1$ regime besides the one in Fig.1, such that the rate of $n^{1/5}[\mathcal{TV}_1(\boldsymbol{w}_{1:n})]^{2/5} \vee 1$ is faster than the rate $n^{1/3}[\mathcal{TV}_0(\boldsymbol{w}_{1:n})]^{2/3} \vee 1$. We defer a non-exhaustive list of such examples to Appendix G.*

Remark 8. *One may ask if a simpler algorithm such as carefully tuned online gradient descend (OGD) can enjoy these fast rates too. However, Proposition 2 of Baby and Wang (2020) implies that properly tuned OGD algorithm which is optimal against set of all comparators $\boldsymbol{w}_{1:n}$ with $\mathcal{TV}_0(\boldsymbol{w}_{1:n}) \leq 1$ under convex losses, necessarily suffers a slower dynamic regret of $\Omega(n^{1/4})$ against set of all comparators with $\mathcal{TV}_1(\boldsymbol{w}_{1:n}) \leq 1$ under exp-concave losses (see Lemma 14 in Appendix A).*

Remark 9 (optimized implementation). *In the presented form of FLH-SIONS, we need to hedge over $O(n)$ SIONS base learners per round to make predictions. However, this can be reduced to hedging over $O(\log n)$ base learners per round by using AFLH from Hazan and Seshadhri (2007) or PAE from Zhang et al. (2021) as the aggregation algorithm. This route will help to achieve a near-linear overall complexity of $O(n \log n)$ calls to the base learners. The cost of doing this is that it enlarges the dynamic regret bound by a factor of $O(\log n)$. Moreover, the run-time of the SIONS base learners can be further ameliorated at the expense of slightly increasing the regret bound by using randomized sketchings as described in Luo et al. (2016). It is also possible to specify custom base learners in the case we know the form of losses ahead of time. If the base learners incur logarithmic static regret under the specified loss, then we can enjoy the regret rate specified in Proposition 5. For example, if the losses are linear regression type losses, one can use Vovk-Azoury-Warmuth forecaster (Cesa-Bianchi and Lugosi, 2006) as base learners. If the losses are logistic regression losses, one can use the algorithm in Jézéquel et al. (2020) as base learners. Such custom base learners can have much lower run-time than SIONS base learners which are designed to support the fully general exp-concave losses. We also note that losses such as linear and logistic regression losses are commonly used in practice as well.*

4.1 Proof Summary of Theorem 3 for one dimension

In what follows, we present several useful lemmas and provide a running sketch on how to chain them to arrive at Theorem 3 in a uni-variate setting (i.e $d = 1$). Detailed proofs are deferred to Appendix C.1.

Suppose that we need to compete against comparators whose $\mathcal{TV}_1$ distance (i.e $n\|D^2 w_{1:n}\|_1$) is bounded by some number C_n. This quantity could be unknown to the algorithm. Consider the *offline oracle* who has access to the entire sequence of loss functions $f_1, \ldots, f_n$ and the $\mathcal{TV}_1$ bound C_n. It may then solve for the strongest possible comparator respecting the $\mathcal{TV}_1$ bound through the following convex optimization problem.

$$\min_{\tilde{u}_1, \ldots, \tilde{u}_n} \quad \sum_{t=1}^{n} f_t(\tilde{u}_t)$$

$$\text{s.t.} \quad \|D^2 \tilde{u}_{1:n}\|_1 \leq C_n/n, \tag{5a}$$

$$-1 \leq \tilde{u}_t \ \forall t \in [n], \tag{5b}$$

$$\tilde{u}_t \leq 1 \ \forall t \in [n], \tag{5c}$$

Let $u_1, \ldots, u_n$ be the optimal solution of the above problem. This sequence will be referred as **offline optimal** henceforth. Clearly we have that the regret against any comparator sequence $\boldsymbol{w}_{1:n}$ with $\mathcal{TV}_1(\boldsymbol{w}_{1:n}) \leq C_n$ obeys

$$\sum_{t=1}^{n} f_t(p_t) - f_t(w_t) \leq \sum_{t=1}^{n} f_t(p_t) - f_t(u_t),$$

and hence it suffices to bound the right side of the above inequality.

Next, we provide a partition of the horizon with certain useful properties.

Lemma 10. ***(key partition)*** *For some interval $[a, b] \in [n]$, define $\ell_{a \to b} := b - a + 1$. There exists a partitioning of the time horizon $\mathcal{P} := \{[1_s, 1_t], \ldots, [i_s, i_t], \ldots [M_s, M_t]\}$ where $M = |\mathcal{P}|$ such that for any bin $[i_s, i_t] \in \mathcal{P}$ we have: 1) $\|D^2 u_{i_s:i_t}\|_1 \leq 1/\ell_{i_s \to i_t}^{3/2}$; 2) $\|D^2 u_{i_s:i_t+1}\|_1 > 1/\ell_{i_s \to i_t+1}^{3/2}$ and 3) $M = O\left(n^{1/5} C_n^{2/5} \vee 1\right)$.*

Going forward, the idea is to bound the dynamic regret within each bin in $\mathcal{P}$ by an $\tilde{O}(1)$ quantity. Then we can add them up across all bins to arrive at the guarantee of Theorem 3 (with $d = 1$). We pause to remark that eventhough this high-level idea resembles to that of (Baby and Wang, 2021), the underlying details of our analysis to materialize this idea requires highly non-trivial deviations from the path followed by (Baby and Wang, 2021).

First, we need some definitions. Consider a bin $[i_s, i_t] \in \mathcal{P}$ with length at-least 2. Let's define a co-variate $\boldsymbol{x}_j := [1, j - i_s + 1]^T$. Let $\boldsymbol{X}^T := [\boldsymbol{x}_{i_s}, \ldots, \boldsymbol{x}_{i_t}]$ be the matrix of co-variates and $u_{i_s:i_t} := [u_{i_s}, \ldots, u_{i_t}]^T$. Let $\boldsymbol{\beta} = \left(\boldsymbol{X}^T \boldsymbol{X}\right)^{-1} \boldsymbol{X}^T u_{i_s:i_t}$ be the least square fit coefficient computed with co-variates $\boldsymbol{x}_j$ and labels u_j. Define a second moment matrix $\boldsymbol{A} = \sum_{j=i_s}^{i_t} \boldsymbol{x}_j \boldsymbol{x}_j^T$. Let $\boldsymbol{\alpha} := \boldsymbol{\beta} - \boldsymbol{A}^{-1} \sum_{j=i_s}^{i_t} \nabla f_j(\boldsymbol{\beta}^T \boldsymbol{x}_j) \boldsymbol{x}_j$. ($\boldsymbol{A}^{-1}$ is guaranteed to exist when length of the bin is at-least 2). We remind the reader that $\nabla f_j(\boldsymbol{\beta}^T \boldsymbol{x}_j)$ is a scalar as we consider uni-variate f_j in this section.

We connect these quantities via a *key regret decomposition* as follows:

$$\sum_{j=i_s}^{i_t} f_j(p_j) - f_j(u_j) = \underbrace{\sum_{j=i_s}^{i_t} f_j(p_j) - f_j(\boldsymbol{\alpha}^T \boldsymbol{x}_j)}_{T_1} + \underbrace{\sum_{j=i_s}^{i_t} f_j(\boldsymbol{\alpha}^T \boldsymbol{x}_j) - f_j(\boldsymbol{\beta}^T \boldsymbol{x}_j)}_{T_2} + \underbrace{\sum_{j=i_s}^{i_t} f_j(\boldsymbol{\beta}^T \boldsymbol{x}_j) - f_j(u_j)}_{T_3} \tag{6}$$

It can be shown that $|\boldsymbol{\alpha}^T \boldsymbol{x}_j| \leq 20 = O(1)$. Hence the term T_1 can be controlled by an $O(\log n)$ bound due to Strong Adaptivity of FLH-SIONS as described in Section 3, Eq.(4). The quantity $\boldsymbol{\alpha}$ is obtained via moving in a direction reminiscent to that of Newton method. This is in sharp contrast to the one step gradient descent update used in Baby and Wang (2021). More precisely, consider the function $F(\boldsymbol{\beta}) = \sum_{j=i_s}^{i_t} f_j(\boldsymbol{\beta}^T \boldsymbol{x}_j)$. Then $\boldsymbol{\alpha} = \boldsymbol{\beta} - \boldsymbol{A}^{-1} \nabla F(\boldsymbol{\beta})$. By exploiting gradient Lipschitzness of f_j, the correction matrix $\boldsymbol{A}$ can be shown to satisfy the Hessian dominance $\nabla^2 F(\boldsymbol{\beta}) \preccurlyeq \boldsymbol{A}$. This Newton style update is shown to keep the term T_2 to be negative through the following generalized descent lemma:

Lemma 11. *We have that $T_2 \leq -\frac{1}{2} \|\nabla F(\boldsymbol{\beta})\|^2_{\boldsymbol{A}^{-1}}$.*

The negative descent term displayed in the above Lemma is similar to the standard (squared) Newton decrement (Nesterov, 2004) in the sense that it is also influenced by the local geometry through the norm induced by the inverse correction matrix $\boldsymbol{A}^{-1}$.

We then proceed to show that the negative T_2 can diminish the effect of T_3 by keeping $T_2 + T_3$ to be an $O(1)$ quantity. Thus the dynamic regret within the bin $[i_s, i_t] \in \mathcal{P}$ is controlled to $\tilde{O}(1)$. Adding the bound across all bins in $\mathcal{P}$ from Lemma 10 yields Theorem 3 in one dimension.

A major challenge in the analysis is to prove that the term $T_2 + T_3 = O(1)$ without imposing restrictive assumptions such as Self-Concordance or Hessian Lipschitzness as in the classical analysis of Newton method (see for eg.(Nesterov, 2004)). In the rest of this section, we outline the arguments leading to this result.

Lemma 12. *We have that $T_2 + T_3 = O(1)$ where T_2 and T_3 are as defined in Eq.(6)*

Here the main idea is to exploit the KKT conditions of the offline optimization problem in Eq.(5) to show $T_2 + T_3 = O(1)$ even if $|T_2|$ and $|T_3|$ can be very large individually. Though this is similar to the observation in Baby and Wang (2021), our regret decomposition in Eq.(6) and the KKT conditions (see Lemma 17 in Appendix B) are different. So showing this result requires non-trivial deviations from the proof of Baby and Wang (2021). Importantly, it was not apriori clear that $T_2 + T_3$ can be possibly bound by $O(1)$ for the current problem. The key novelty is that we bound $T_2 + T_3$ by introducing an auxiliary function that is concave in its arguments which allows us to systematically explore the properties of its maximizers. We refer the reader to Appendix B for a thorough overview on the construction and use of such auxiliary functions in proving the lemma.

As discussed before, the case of multiple dimensions is handled by constructing suitable reductions that will allow us to re-use much of the analytical machinery developed for the case of uni-variate setting. We refer the reader to Appendix B for an overview of the details of such reductions. However, we emphasize that this reduction happens only in the analysis, and we *do not* run d uni-variate FLH-SIONS algorithms for handling multi-dimensions (see Theorem 3).

We conclude this section by noting that our proof techniques also lead to a dynamic regret bound that *simultaneously* hold for any sub-interval of $[n]$.

Remark 13. *Assume the notations used in Lemma 10 and Proposition 5. Consider an interval $[a, b] \subseteq [n]$. By partitioning this sub-interval as per Lemma 10 and applying the regret decomposition of Eq.(6), one can show the following dynamic regret bound over the interval $[a, b]$:*

$$\sum_{t=a}^{b} f_t(\boldsymbol{p}_t) - f_t(\boldsymbol{w}_t) = \tilde{O}\left(d^2 \left(\left(\ell^{1/5} \left(\ell \|D^2 \boldsymbol{w}_{a:b}\|_1 \right)^{2/5} \right) \wedge \left(\ell^{1/3} \|D^1 \boldsymbol{w}_{a:b}\|_1^{2/3} \right) \right) \vee d^2 \right),$$

where $\ell := \ell_{a \to b}$.

5 CONCLUSION

In this work, we derived universal dynamic regret rate parametrized by a *novel* second-order path variational of the comparators. Such a path variational naturally captures the piecewise linear structures of the comparators and can be used to flexibly model many practical non-stationarities in the environment. Our results for the exp-concave losses achieved an adaptive universal dynamic regret of $\tilde{O}^* \left(\min\{n^{1/5}[\mathcal{TV}_1(\boldsymbol{w}_{1:n})]^{2/5}, n^{1/3}[\mathcal{TV}_0(\boldsymbol{w}_{1:n})]^{2/3}\} \vee 1 \right)$ which matches the minimax lower bound up to a factor that depends on d and $\log n$. This is the first result of such kind in the adversarial setting and the first that works with general exp-concave family of losses. We conjecture that a similar algorithm as in Fig.3 based on degree k monomial co-variates $[1, t, \ldots, t^k]$ can lead to optimal dynamic regret in terms of $\mathcal{TV}_k$.

Acknowledgements

The authors are grateful to the anonymous reviewers whose constructive comments helped to enhance the paper. The work is partially supported by NSF Awards #2007117 and a gift from Amazon.

Bibliography

Dmitry Adamskiy, Wouter M. Koolen, Alexey Chernov, and Vladimir Vovk. A closer look at adaptive regret. *Journal of Machine Learning Research*, 2016.

Dheeraj Baby and Yu-Xiang Wang. Online forecasting of total-variation-bounded sequences. In *Neural Information Processing Systems (NeurIPS)*, 2019.

Dheeraj Baby and Yu-Xiang Wang. Adaptive online estimation of piecewise polynomial trends. *Neural Information Processing Systems (NeurIPS)*, 2020.

Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in exp-concave online learning. In *COLT*, 2021.

Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in proper online learning with strongly convex losses and beyond. *AISTATS*, 2022a.

Dheeraj Baby and Yu-Xiang Wang. Optimal dynamic regret in LQR control. In *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=D0aqV81d0_k.

Dheeraj Baby, Xuandong Zhao, and Yu-Xiang Wang. An optimal reduction of tv-denoising to adaptive online learning. *AISTATS*, 2021.

Omar Besbes, Yonatan Gur, and Assaf Zeevi. Non-stationary stochastic optimization. *Operations research*, 63(5):1227–1244, 2015.

Nicolo Cesa-Bianchi and Gabor Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, NY, USA, 2006. ISBN 0521841089.

Xi Chen, Yining Wang, and Yu-Xiang Wang. Non-stationary stochastic optimization under lp, q-variation measures. 2018.

Albert Cohen, Ingrid Daubechies, and Pierre Vial. Wavelets on the interval and fast wavelet transforms. *Applied and Computational Harmonic Analysis*, 1(1):54 – 81, 1993. ISSN 1063-5203.

Ashok Cutkosky. Parameter-free, dynamic, and strongly-adaptive online learning. In *ICML*, 2020.

Amit Daniely, Alon Gonen, and Shai Shalev-Shwartz. Strongly adaptive online learning. In *International Conference on Machine Learning*, pages 1405–1411, 2015.

Ronald A. DeVore and George G. Lorentz. Constructive approximation. In *Grundlehren der mathematischen Wissenschaften*, 1993.

David Donoho, Richard Liu, and Brenda MacGibbon. Minimax risk over hyperrectangles, and implications. *Annals of Statistics*, 18(3):1416–1437, 1990.

David L Donoho and Iain M Johnstone. Minimax estimation via wavelet shrinkage. *The annals of Statistics*, 26(3):879–921, 1998.

Pierre Gaillard and Sébastien Gerchinovitz. A chaining algorithm for online nonparametric regression. In *Conference on Learning Theory*, pages 764–796, 2015.

Gautam Goel and Adam Wierman. An online algorithm for smoothed regression and lqr control. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, 2019.

Elad Hazan. Introduction to online convex optimization. *Foundations and Trends® in Optimization*, 2(3-4):157–325, 2016.

Elad Hazan and Comandur Seshadhri. Adaptive algorithms for online decision problems. In *Electronic colloquium on computational complexity (ECCC)*, volume 14, 2007.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. 2007.

Andrew Jacobsen and Ashok Cutkosky. Parameter-free mirror descent. *COLT*, 2022.

Ali Jadbabaie, Alexander Rakhlin, Shahin Shahrampour, and Karthik Sridharan. Online optimization: Competing with dynamic comparators. In *Artificial Intelligence and Statistics*, pages 398–406, 2015.

Rémi Jézéquel, Pierre Gaillard, and Alessandro Rudi. Efficient improper learning for online logistic regression. In *Proceedings of Thirty Third Conference on Learning Theory*, 2020.

Iain M. Johnstone. *Gaussian estimation: Sequence and wavelet models*. 2017.

Seung-Jean Kim, Kwangmoo Koh, Stephen Boyd, and Dimitry Gorinevsky. ℓ_1 trend filtering. *SIAM Review*, 51(2): 339–360, 2009.

Wouter M Koolen, Alan Malek, Peter L Bartlett, and Yasin Abbasi. Minimax time series prediction. In *Advances in Neural Information Processing Systems (NIPS'15)*, pages 2557–2565. 2015.

Wojciech Kotłowski, Wouter M. Koolen, and Alan Malek. Online isotonic regression. In *Annual Conference on Learning Theory (COLT-16)*, volume 49, pages 1165–1189. PMLR, 2016.

Haipeng Luo, Alekh Agarwal, Nicolò Cesa-Bianchi, and John Langford. Efficient second order online learning by sketching. In *NIPS*, 2016.

Aaron Meurer, Christopher P. Smith, Mateusz Paprocki, Ondřej Čertík, Sergey B. Kirpichev, Matthew Rocklin, AMiT Kumar, Sergiu Ivanov, Jason K. Moore, Sartaj Singh, Thilina Rathnayake, Sean Vig, Brian E.

Granger, Richard P. Muller, Francesco Bonazzi, Harsh Gupta, Shivam Vats, Fredrik Johansson, Fabian Pedregosa, Matthew J. Curry, Andy R. Terrel, Štěpán Roučka, Ashutosh Saboo, Isuru Fernando, Sumith Kulal, Robert Cimrman, and Anthony Scopatz. Sympy: symbolic computing in python. *PeerJ Computer Science*, January 2017.

Aryan Mokhtari, Shahin Shahrampour, A. Jadbabaie, and Alejandro Ribeiro. Online optimization in dynamic environments: Improved regret rates for strongly convex problems. *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 7195–7201, 2016.

Yurii Nesterov. Introductory lectures on convex optimization - a basic course. In *Applied Optimization*, 2004.

Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression. In *Conference on Learning Theory*, pages 1232–1264, 2014.

Alexander Rakhlin and Karthik Sridharan. Online nonparametric regression with general loss functions, 2015.

Ryan J. Tibshirani. Adaptive piecewise polynomial estimation via trend filtering. *Annals of Statistics*, 42(1): 285–323, 2014.

Sara van de Geer. Estimating a regression function. *Annals of Statistics*, 18(2):907—924, 1990.

Yu-Xiang Wang, James Sharpnack, Alex Smola, and Ryan J Tibshirani. Trend filtering on graphs. *Journal of Machine Learning Research*, 17(105):1–41, 2016.

Tianbao Yang, Lijun Zhang, Rong Jin, and Jinfeng Yi. Tracking slowly moving clairvoyant: optimal dynamic regret of online learning with true and noisy gradient. In *International Conference on Machine Learning (ICML-16)*, pages 449–457, 2016.

Jianjun Yuan and Andrew Lamperski. Trading-off static and dynamic regret in online least-squares and beyond. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 6712–6719, 2020.

Lijun Zhang, Shiyin Lu, and Zhi-Hua Zhou. Adaptive online learning in dynamic environments. In *Advances in Neural Information Processing Systems (NeurIPS-18)*, pages 1323–1333, 2018a.

Lijun Zhang, Tianbao Yang, Zhi-Hua Zhou, et al. Dynamic regret of strongly adaptive methods. In *International Conference on Machine Learning (ICML-18)*, pages 5877–5886, 2018b.

Lijun Zhang, G. Wang, Wei-Wei Tu, and Zhi-Hua Zhou. Dual adaptivity: A universal algorithm for minimizing the adaptive regret of convex functions. *NeurIPS*, 2021.

Zhiyu Zhang, Ashok Cutkosky, and Ioannis Ch. Paschalidis. Unconstrained dynamic regret via sparse coding. *arXiv preprint arXiv:2301.13349*, 2023.

Peng Zhao and Lijun Zhang. Improved analysis for dynamic regret of strongly convex and smooth functions. *L4DC*, 2021.

Peng Zhao, Y. Zhang, L. Zhang, and Zhi-Hua Zhou. Dynamic regret of convex and smooth functions. *NeurIPS*, 2020.

Peng Zhao, Yu-Xiang Wang, and Zhi-Hua Zhou. Non-stationary online learning with memory and non-stochastic control. *AISTATS*, 2022.

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

A More on Related Work

In this section, we elaborate upon the related works mentioned in Section 2. We inherit all the notations and terminologies introduced in Section 1.

Dynamic regret against $\mathcal{TV}^{(1)}(C_n)$ in stochastic setting. Perhaps, the most relevant to our work is that of Baby and Wang (2020). They consider an online protocol where at each round, the learner makes a prediction $\hat{\theta}_t \in \mathbb{R}$. Then a label $y_t = \theta_t + \mathcal{N}(0,1)$ is revealed. They assume that the ground truth sequence $\theta_{1:n} \in \mathcal{TV}^{(k)}(C_n)$ (see Eq.(3)). The goal of the learner is control the expected cumulative squared error of the learner namely $\sum_{t=1}^{n}(\hat{\theta}_t - \theta_t)^2$. In this setting, they propose policies that can attain a near optimal estimation error of $\tilde{O}(n^{\frac{1}{2k+3}} C_n^{\frac{2}{2k+3}})$ for any $k > 0$. In retrospect, in this work we consider the case where comparators belong to a $\mathcal{TV}^{(1)}(C_n)$ class (i.e, with k=1). Further, the absence of stochastic assumptions in our setting poses a significant challenge in controlling the universal dynamic regret (Eq.(1)).

Restricted dynamic regret minimization. In this line of work, we consider a similar learning setting as mentioned in Section 1. However, the goal of the learner is to control the dynamic regret against point-wise minimizers:

$$R_{n,\text{restrict}} = \sum_{t=1}^{n} f_t(\boldsymbol{p}_t) - f_t(\boldsymbol{w}_t^*),$$

where $\boldsymbol{w}_t^* \in \operatorname{argmin} x \in \mathcal{W} f_t(\boldsymbol{x})$. When the losses are strongly convex and gradient smooth, Mokhtari et al. (2016) proposes algorithms that can attain a restricted dynamic regret of $O(1 + C_n^*)$, where $C_n^* = \sum_{t=1}^{n-1} \|\boldsymbol{w}_t^* - \boldsymbol{w}_{t+1}^*\|_2$. However, as noted in Zhang et al. (2018a), such a guarantee can be sometimes overly pessimistic. For example, in the context of statistical learning where the losses are sampled iid from a distribution, the point-wise minimizers can incur a path length $C_n^* = O(n)$ due to random perturbations introduced by sampling.

Universal dynamic regret minimization. This is the same framework as considered in the introduction. Obtaining universal dynamic regret guarantees is challenging since we need to bound the dynamic regret for *any* comparator sequence from the bench mark set $\mathcal{W}$ while automatically adapting to their path length. When the losses are convex Zhang et al. (2018a); Cutkosky (2020) provides an optimal universal dynamic regret of $O(\sqrt{n(1+P_n)})$, where

$$P_n = \sum_{t=1}^{n-1} \|\boldsymbol{w}_t - \boldsymbol{w}_{t+1}\|_2.$$

For $C_n = \Omega(1)$, the embedding $\mathcal{TV}^{(1)}(C_n) \subseteq \mathcal{TV}^{(0)}(\kappa C_n)$ (see Proposition 37) where κ is a constant doesn't depend on n or C_n implies that work of Zhang et al. (2018a) yields a dynamic regret rate of $O^*(\sqrt{nC_n})$ for competing against comparator sequences in $\mathcal{TV}^{(1)}(C_n)$ class. (O^* hides dimension dependencies.) This is a sub-optimal rate when applied to our setting as expected, since they don't assume the losses are exp-concave.

When the losses are in-fact exp-concave, one can apply the result of Baby and Wang (2021) to produce a dynamic regret rate of $\tilde{O}^*(n^{1/3} C_n^{2/3} \vee 1)$. However, as noted in Section 1, this rate is sub-optimal. In Appendix F, we give an elaborate description on why their analysis lead to sub-optimal rate in our setting of competing against comparators from $\mathcal{TV}^{(1)}(C_n)$ class.

Dynamic regret based on functional variations. It is also common to measure the non-stationarity of the environment in terms of the variation incurred by the loss function sequence. Define

$$D_n = \sum_{t=2}^{n} \max_{\boldsymbol{w} \in \mathcal{W}} |f_t(\boldsymbol{w}) - f_{t-1}(\boldsymbol{w})|.$$

There are works such as (Besbes et al., 2015; Yang et al., 2016; Chen et al., 2018) that aims in controlling the dynamic(restricted / universal) in terms of D_n. Jadbabaie et al. (2015) proposes algorithms that can control dynamic regret simultaneously in the terms of D_n and P_n.

Static regret minimization. In classical OCO, a well known metric is to control the static regret of an algorithm namely, $\sum_{t=1}^{n} f_t(\boldsymbol{p}_t) - f_t(\boldsymbol{w})$. Algorithms such as Online Gradient Descent (Zinkevich, 2003) can attain an optimal $O(\sqrt{n})$ static regret when losses are convex. When the losses are exp-concave or strongly convex it is possible to attain logarithmic static regret (Hazan et al., 2007). However, static regret is not a good metric for measuring the performance of a leraner in non-stationary environments.

Strongly Adaptive (SA) regret minimization. This notion of regret is introduced by Daniely et al. (2015). In this framework, the learner aims in controlling its static regret in any local time window as a function of window-length (modulo factors of $\log n$). The algorithms in Daniely et al. (2015); Cutkosky (2020) provides a static regret of $\tilde{O}(\sqrt{|I|})$ across any local interval I whenever the losses are convex. When the losses are strongly convex or exp-concave, the algorithms in (Hazan and Seshadhri, 2007; Adamskiy et al., 2016; Zhang et al., 2021) yields logarithmic static regret in any local time window when the base learners are chosen appropriately.

Zhang et al. (2018b) shows that SA algorithms can be used to control the dynamic regret in terms of the functional variation D_n.

Online non-parametric regression. In section 1, we modelled the dynamics of the comparator sequence as a member of a non-parametric function class. Rakhlin and Sridharan (2014, 2015) studies the minimax rate of learning against a non-parametric function class. They establish the right minimax rate (in terms of dependencies wrt n) using arguments based on sequential Rademacher complexity in a non-constructive manner. In-fact, their results on Besov spaces imply that the minimax dynamic regret rate for our problem is indeed $O(n^{1/5})$ since the $\mathcal{TV}^{(1)}$ class is sandwiched between two Besov spaces having the same minimax rate (see for eg. (DeVore and Lorentz, 1993) and (Donoho and Johnstone, 1998)). There are other line of works that study the non-parametric regression problem against other function classes of interest such as Gaillard and Gerchinovitz (2015) (for Holder class of functions), Kotłowski et al. (2016) (for isotonic functions) and Koolen et al. (2015) (for Sobolev functions). These function classes doesn't capture the $\mathcal{TV}^{(1)}$ class that we study in this work. For instance, the (discretized) higher order Holder and Sobolev classes features sequences that are more regular than that of $\mathcal{TV}^{(1)}$ class (see for example Baby and Wang (2019)).

Further explanations on Remark 8. In rest of this section, we focus on proving Remark 8.

It is sufficient to focus on OGD algorithms with step-size $\eta < 1$. Otherwise if $\eta \geq 1$, one can come up with sequence of losses that can enforce linear regret. An example of this scenario is described as follows:

The loss at time t is given by $f_t(x) = \frac{(x-y_t)^2}{2}$ with $y_t = -1$ at odd rounds and $y_t = 1$ at even rounds. The decision set is given by $\mathcal{D} = [-1, 1]$. We focus on OGD algorithms which plays 0 at the first round, though similar arguments can be given for any valid initialization point. Suppose the step size is $\eta \geq 1$. The iterate at step $t+1$, denoted by w_{t+1}, is maintained as

$$w_{t+1} = \text{clip}_{[-1,1]}\left(w_t - \eta(w_t - y_t)\right),$$

where clip function clips the argument to $[-1, 1]$.

Recall that $\eta \geq 1$. So we have $w_1 = 0$, $w_2 = \text{clip}_{[-1,1]}(-\eta) = -1$, $w_3 = \text{clip}_{[-1,1]}(-1(1-\eta)+\eta) = 1$, $w_4 = \text{clip}_{[-1,1]}(1(1-\eta)-\eta) = -1$ and so on.

Thus the iterates oscillates between -1 and 1. However, the best fixed comparator for the sequence of losses is given by 0. Hence we have that

$$\begin{aligned}\sum_{t=1}^{n} f_t(w_t) - f_t(0) &= 1/2 + 2(n-1) - n/2\\ &\geq 3n/4,\end{aligned}$$

for all $n \geq 2$.

Thus choosing step-size $\eta \geq 1$ can be exploited by the adversary to enforce a linear regret even for the case of static comparators.

So it suffices to consider OGD algorithms with step size $\eta < 1$ as what is done in the following lemma.

First, let's define the comparator class:

$$\mathcal{TV}^{(1)}(1) := \{\theta_{1:n} : \mathcal{TV}_1(\theta_{1:n}) \leq 1\}.$$

Lemma 14. *There exist a choice of loss functions, comparator sequence $\theta_{1:n} \in \mathcal{TV}^{(1)}(1)$ and decision set such that OGD with steps size $\eta < 1$ necessarily suffers a dynamic regret of $\Omega(n^{1/4})$ for all $n \geq 35$.*

Proof. Consider a setup where the decision set $\mathcal{D} = [-1/(2\sqrt{2.2}), 1/(2\sqrt{2.2})]$. Let $y_t = \theta_t + \epsilon_t$ where $|\theta_t| \leq 1/(8\sqrt{2.2})$, and ϵ_t are iid uniformly chosen from $[-1/(8\sqrt{2.2}), 1/(8\sqrt{2.2})]$. Thus $y_t \in [-1/(4\sqrt{2.2}), 1/(4\sqrt{2.2})]$. Further $\theta_{1:n} \in \mathcal{TV}^{(1)}(1)$.

The loss at time t is $f_t(x) = \frac{2\sqrt{2.2}}{3}(y_t - x)^2$. So the Lispchitzness coefficient of these losses is bounded by $|\nabla f_t(x)| \leq \frac{4\sqrt{2.2}}{3}(|y_t| + |x|) \leq 1 := G$ as $|y_t| \leq 1/(4\sqrt{2.2})$ and $|x| \leq 1/(2\sqrt{2.2})$ for all $x \in \mathcal{D}$.

Let D be the diameter of $\mathcal{D}$. So $D = 1/\sqrt{2.2}$.

Under this setup, the projected online gradient descent (OGD) with learning rates $\eta < 1$ doesn't need to do any projection. This can be seen as follows. Assume that OGD till step t doesn't project. Let Π denote the projection to set $\mathcal{D}$. Then the iterate at time $t+1$ (denoted by x_{t+1}) is given by $x_{t+1} = \Pi(z_{t+1})$ where

$$z_{t+1} = \sum_{k=1}^{t} (1-\eta)^{t-k} \eta y_k. \tag{7}$$

We have that

$$|z_{t+1}| \leq 1/(4\sqrt{2.2}),$$

where we applied triangle inequality, summed up the infinite series and used the fact that $|y_t| \leq 1/(4\sqrt{2.2})$. So $z_{t+1} \in \mathcal{D}$ and therefore $x_{t+1} = z_{t+1}$. Hence by induction, we have that OGD with learning rate $\eta < 1$ doesn't need to project.

Looking at Eq.(7) we see that the OGD output at any time is a fixed linear function of the revealed labels y_t. Baby and Wang (2020) calls such forecasters to be linear forecasters. They provide the following proposition (para-phrased here for clarity) about such forecasters:

Proposition 15. *(Proposition 2 in Baby and Wang (2020) for $k = 1$) For any online estimator producing estimates $\hat{\theta}_t$ which is a fixed linear function of past labels $y_{1:t-1}$, $t \in [n]$ we have*

$$\sup_{\theta_{1:n} \in \in \mathcal{TV}^{(1)}(1)} \sum_{t=1}^{n} E[(\hat{\theta}_t - \theta_t)^2] = \Omega(n^{1/4}). \tag{8}$$

Thus we have

$$\begin{aligned}
\sup_{\theta_{1:n} \in \in \mathcal{TV}^{(1)}(1)} \sum_{t=1}^{n} E[(y_t - \hat{\theta}_t)^2] - E[(y_t - \theta_t)^2] &= \sup_{\theta_{1:n} \in \in \mathcal{TV}^{(1)}(1)} \sum_{t=1}^{n} E[(\hat{\theta}_t - \theta_t)^2] - 2E[\epsilon_t(\hat{\theta}_t - \theta_t)] + E[\epsilon_t^2] - E[\epsilon_t^2] \\
&=_{(a)} \sup_{\theta_{1:n} \in \in \mathcal{TV}^{(1)}(1)} \sum_{t=1}^{n} E[(\hat{\theta}_t - \theta_t)^2] - 2E[\epsilon_t] E[(\hat{\theta}_t - \theta_t)] \\
&= \sup_{\theta_{1:n} \in \in \mathcal{TV}^{(1)}(1)} \sum_{t=1}^{n} E[(\hat{\theta}_t - \theta_t)^2] \\
&=_{(b)} \Omega(n^{1/4}),
\end{aligned}$$

where line (a) is due to the fact that $\hat{\theta}_t$ and ϵ_t are mutually independent (because of online nature of algorithm) and line (b) is due to Eq.(8).

Thus we conclude that

$$\sup_{\theta_{1:n} \in \mathcal{TV}^{(1)}(1)} \sum_{t=1}^{n} E[f_t(\hat{\theta}_t) - f_t(\theta_t)] = \Omega(n^{1/4}),$$

where the losses f_t are as defined in the beginning of the proof. This concludes the lemma. ∎

B Overview of proof strategy

Remark 16. *(**reason behind faster rates**). We remark that in the low TV1 regime, the sequence assumes a piecewise linear structure with gradually changing slopes. This regularity of the comparator sequence is what enables to derive fast regret rates in low TV1 regimes.*

For the sake of clarity, we give an elaborate overview of our proof scheme before presenting the analysis in Appendix C. We adopt the notations introduced in Section 4.

We start with the KKT conditions of the offline optimization problem (Eq.(5)) defined in Section 4.1.

Lemma 17. *(**KKT conditions**) Let $u_1, \ldots, u_n$ be the optimal primal variables and let $\lambda \geq 0$ be the optimal dual variable corresponding to the constraint (5a). Further, let $\gamma_t^- \geq 0, \gamma_t^+ \geq 0$ be the optimal dual variables that correspond to constraints (5b) and (5c) respectively for all $t \in [n]$. By the KKT conditions, we have*

- ***stationarity:*** $\nabla f_t(u_t) = \lambda\left((s_{t-1} - s_t) - (s_{t-2} - s_{t-1})\right) + \gamma_t^- - \gamma_t^+$*, where* $s_t = \mathrm{sign}((u_{t+2} - u_{t+1}) - (u_{t+1} - u_t))$*. Here* $\mathrm{sign}(x) = x/|x|$ *if* $|x| > 0$ *and any value in* $[-1, 1]$ *otherwise. For convenience of notations, we also define* $s_{-1} = s_0 = s_{n-1} = s_n = 0$*.*
- ***complementary slackness:*** *(a)* $\lambda\left(\|D^2 u_{1:n}\|_1 - C_n/n\right) = 0$*; (b)* $\gamma_t^-(u_t + 1) = 0$ *and* $\gamma_t^+(u_t - 1) = 0$ *for all* $t \in [n]$

As mentioned in Section 4, a key step is proving Lemma 12 which we restate for convenience.

Lemma 12. *We have that $T_2 + T_3 = O(1)$ where T_2 and T_3 are as defined in Eq.(6)*

Proof Sketch. For the sake of explaining ideas, we consider a case where the offline optimal within a bin $[i_s, i_t] \in \mathcal{P}$ doesn't touch the boundary 1 but may touch boundary -1 at multiple time points. (In the full proof, we show that the partition $\mathcal{P}$ can be slightly modified so that in non-trivial cases, the offline optimal can only touch one of the boundaries due to the TV1 constraint within the bins described in Lemma 10.) Then by complementary slackness of Lemma 17 we conclude that $\gamma_j^+ = 0$ for all $j \in [i_s, i_t]$. Our analysis starts by considering a scenario where the offline optimal touches boundary -1 at precisely two points $r, w \in [i_s, i_t]$ with $r < w$ (see Fig.5). Again via complementary slackness, only γ_r^- and γ_w^- can be potentially non-zero in this case. Through certain careful bounding steps, we show that:

$$T_2 + T_3 \leq -B(\lambda, \gamma_r^-, \gamma_w^-; r, w), \tag{9}$$

where B is a function jointly convex in its arguments $\lambda, \gamma_r^-, \gamma_w^-$. We treat r and w to be fixed parameters. The exact form of the function B is present at Eq.(45) in Appendix. Then we consider the following convex optimization procedure:

$$\min_{\lambda, \gamma_r^-, \gamma_w^-} \quad B(\lambda, \gamma_r^-, \gamma_w^-; r, w) \tag{10a}$$
$$\text{s.t.} \quad \lambda \geq 0$$

First, we perform a partial minimization wrt γ_r^- and γ_w^- keeping λ fixed. Note that even-though $\gamma_r^- \geq 0$ and $\gamma_w^- \geq 0$ via Lemma 17, we choose to perform an *unconstrained* minimization wrt these variables as doing so can only increase the bound on $T_2 + T_3$.

Let the optimal solutions of the partial minimization procedure be denoted by $\hat{\gamma}_r^-$ and $\hat{\gamma}_w^-$. We find that:

$$B(\lambda, \hat{\gamma}_r^-, \hat{\gamma}_w^-; r, w) = \mathcal{L}(\lambda), \tag{11}$$

where $\mathcal{L}(\lambda)$ is a linear function of λ that *doesn't depend* on r or w (Eq.(48) in Appendix). The constrained minimum of this linear function is then found to be attained at $\lambda = 0$ and we show that

$$-B(0, \hat{\gamma}_r^-, \hat{\gamma}_w^-; r, w) = O(1)$$

This leaves us with an important question on how to handle more than two boundary touches at -1 where many of γ_j^-, $j \in [i_s, i_t]$ can potentially be non-zero. One could perform a similar unconstrained optimization as earlier wrt all γ_j^-. However, deriving the closed form expressions for the optimal $\hat{\gamma}_j^-$ becomes very cumbersome as it involves solving for a

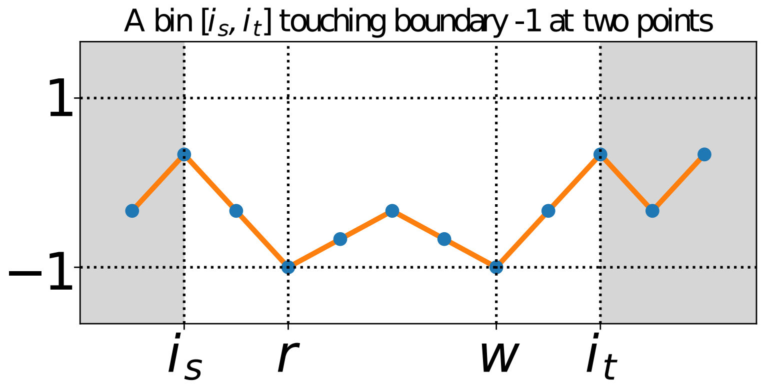


Figure 5: *A configuration referred in the proof sketch of Lemma 12. The blue dots represent the offline optimal sequence.*

complex system of linear equations. In the following, we argue that this general case can be handled via a reduction to the previous setting where only two dual variables γ_r^- and γ_w^- can be potentially non-zero. Specifically we show that the same auxiliary function B as in Eq.(9) can be used to obtain

$$T_2 + T_3 \leq -B(\lambda, \tilde{\gamma}_r^-, \tilde{\gamma}_w^-; \tilde{r}, \tilde{w}),$$

where $\tilde{r}, \tilde{w}, \tilde{\gamma}_r^-$ and γ_w^- can be computed from the sequence of dual variables $\gamma_{i_s:i_t}^-$. Now we can proceed to optimize similarly as in Eq.(10a) with the optimization variables being $\lambda, \tilde{\gamma}_r^-, \tilde{\gamma}_w^-$ and use the same arguments as earlier to bound $T_2 + T_3 = O(1)$. We remark that while doing so, it is an extremely fortunate fact that the partially minimized objective in Eq.(11) does not depend on the parameter values r and w. This fact in hindsight is what permitted us to fully eliminate the dependence of all γ_j^- where $j \in [i_s, i_t]$ on the bound via the method of reduction to the case of two non-zero dual variables considered earlier. □

Proof summary for Theorem 3 in multi-dimensions. In rest of this section, we focus on outlining the analysis ideas that facilitated the main result Theorem 3. The high-level idea is to construct a reduction that helps us to re-use much of the machinery developed in Section 4.1. We emphasize that this reduction happens only in the analysis, and we *do not* run d uni-variate FLH-SIONS algorithms for handling multi-dimensions. Following Lemma serves a key role in materializing the desired reduction.

Lemma 18. *Let $\boldsymbol{X}_j \in \mathbb{R}^{d \times 2d}$ be as defined as:*

$$\boldsymbol{X}_j^T = \begin{bmatrix} \boldsymbol{x}_j[1:2] & \boldsymbol{0} & \dots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{x}_j[3:4] & \dots & \boldsymbol{0} \\ \vdots & \ddots & & \vdots \\ \boldsymbol{0} & \dots & & \boldsymbol{x}_j[2d-1:2d] \end{bmatrix},$$

where $\boldsymbol{0} = [0,0]^T$ and $\boldsymbol{x}_j \in \mathbb{R}^{2d}$. The entries $\boldsymbol{x}_j[2k-1:2k] \in \mathbb{R}^2$ for $k \in [d]$. Let $\tilde{f}_j(\boldsymbol{v}) = f_j(\boldsymbol{X}_j \boldsymbol{v})$ for some $\boldsymbol{v} \in \mathbb{R}^{2d}$ and let $\boldsymbol{\Sigma} := \boldsymbol{X}_j^T \boldsymbol{X}_j \in \mathbb{R}^{2d \times 2d}$ which is a block diagonal matrix. We have that

$$\nabla^2 \tilde{f}_j(\boldsymbol{v}) \preccurlyeq \boldsymbol{\Sigma}.$$

In multi-dimensions also we form a partition $\mathcal{P}$ of the offline optimal similar to Lemma 10. Then we consider following regret decomposition for any bin $[i_s, i_t] \in \mathcal{P}$.

$$\sum_{j=i_s}^{i_t} f_j(\boldsymbol{p}_j) - f_j(\boldsymbol{u}_j) = \underbrace{\sum_{j=i_s}^{i_t} f_j(\boldsymbol{p}_j) - f_j(\boldsymbol{X}_j \boldsymbol{\alpha}_j)}_{T_1} + \underbrace{\sum_{j=i_s}^{i_t} f_j(\boldsymbol{X}_j \boldsymbol{\alpha}_j) - f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)}_{T_2} + \underbrace{\sum_{j=i_s}^{i_t} f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j) - f_j(\boldsymbol{u}_j)}_{T_3}, \quad (12)$$

where we shall shortly describe how to construct the quantities $\boldsymbol{X}_j \in \mathbb{R}^{d \times 2d}$, $\boldsymbol{\alpha}_j \in \mathbb{R}^{2d}$ and $\boldsymbol{\beta}_j \in \mathbb{R}^{2d}$. For compactness of notations later, let's define $\boldsymbol{\alpha}_{j,k} = \boldsymbol{\alpha}_j[2k-1:2k] \in \mathbb{R}^2$, $\boldsymbol{\beta}_{j,k} = \boldsymbol{\beta}_j[2k-1:2k] \in \mathbb{R}^2$ and $\boldsymbol{y}_{j,k} = \boldsymbol{x}_j[2k-1:2k] \in \mathbb{R}^2$

for some $\boldsymbol{x}_j \in \mathbb{R}^{2d}$ as in lemma 18. The Hessian dominance in Lemma 18 leads to:

$$\begin{aligned}\tilde{f}_j(\boldsymbol{\alpha}_j) - \tilde{f}_j(\boldsymbol{\beta}_j) &\leq \sum_{k=1}^{d} \langle \nabla f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)[k] \boldsymbol{y}_{j,k}, \boldsymbol{\alpha}_{j,k} - \boldsymbol{\beta}_{j,k} \rangle + \frac{1}{2} \sum_{k=1}^{d} \|\boldsymbol{\alpha}_{j,k} - \boldsymbol{\beta}_{j,k}\|^2_{\boldsymbol{y}_{j,k}\boldsymbol{y}_{j,k}^T} \\ &:= \sum_{k=1}^{d} t_{2,j,k}. \end{aligned} \tag{13}$$

Further, due to gradient Lipschitzness of f_j,

$$\begin{aligned}\tilde{f}_j(\boldsymbol{\beta}_j) - f_j(\boldsymbol{u}_j) &\leq \sum_{k=1}^{d} \nabla f_j(\boldsymbol{u}_j)[k] \cdot \left(\boldsymbol{\beta}_{j,k}^T \boldsymbol{y}_{j,k} - \boldsymbol{u}_j[k] \right) + \sum_{k=1}^{d} \frac{1}{2} \|\boldsymbol{\beta}_{j,k}^T \boldsymbol{y}_{j,k} - \boldsymbol{u}_j[k]\|_2^2 \\ &:= \sum_{k=1}^{d} t_{3,j,k} \end{aligned} \tag{14}$$

Combining Eq.(13) and (14), we see that $T_2 + T_3$ in any bin $[i_s, i_t]$ can be bounded coordinate-wise:

$$T_2 + T_3 \leq \sum_{k=1}^{d} \sum_{j=i_s}^{i_t} t_{2,j,k} + t_{3,j,k}.$$

This form allows one to bound $\sum_{j=i_s}^{i_t} t_{2,j,k} + t_{3,j,k} = O(1)$ separately for each coordinate by constructing $\boldsymbol{\alpha}_{j,k}, \boldsymbol{\beta}_{j,k}$ and $\boldsymbol{y}_{j,k}$ similar to Section 4.1. We then sum across all coordinates to bound $T_2 + T_3 = O(d)$. We remark that the situation is a bit more subtle here because in-order to handle certain combinatorial structures imposed by the KKT conditions, we had to use a sequence of comparators $\boldsymbol{\alpha}_{i_s}, \ldots, \boldsymbol{\alpha}_{i_t}$ (for linear predictors in Eq.(12)) that switches at-most $O(d)$ times . Finally by appealing to strong adaptivity of FLH-SIONS, we show that $T_1 = \tilde{O}(d^2)$ for each bin $[i_s, i_t] \in \mathcal{P}$ and Theorem 3 then follows by adding the $\tilde{O}(d^2)$ regret across all $O(n^{1/5} C_n^{2/5} \vee 1)$ bins in $\mathcal{P}$.

C Analysis

We start with the analysis in the uni-variate setting followed by the proof in multi-dimensions. The analysis requires very clumsy algebraic manipulations in certain places. We used Python's open-source simplification engine SymPy (Meurer et al., 2017) to assist with the algebraic manipulations.

A remark. The constants occurring in the proofs may be optimized further though we haven't aggressively focused on doing so. Throughout the analysis we compete with comparators whose TV1 distance is bounded by C_n. This quantity can be unknown to the algorithm. Hence the resulting regret rate of FLH-SIONS simultaneously holds for any value of C_n.

C.1 One dimensional setting

Lemma 17. *(KKT conditions) Let $u_1, \ldots, u_n$ be the optimal primal variables and let $\lambda \geq 0$ be the optimal dual variable corresponding to the constraint (5a). Further, let $\gamma_t^- \geq 0, \gamma_t^+ \geq 0$ be the optimal dual variables that correspond to constraints (5b) and (5c) respectively for all $t \in [n]$. By the KKT conditions, we have*

- ***stationarity:*** $\nabla f_t(u_t) = \lambda \left((s_{t-1} - s_t) - (s_{t-2} - s_{t-1}) \right) + \gamma_t^- - \gamma_t^+$, *where* $s_t = \mathrm{sign}((u_{t+2} - u_{t+1}) - (u_{t+1} - u_t))$. *Here* $\mathrm{sign}(x) = x/|x|$ *if* $|x| > 0$ *and any value in* $[-1, 1]$ *otherwise. For convenience of notations, we also define* $s_{-1} = s_0 = s_{n-1} = s_n = 0$.

- ***complementary slackness:*** *(a)* $\lambda \left(\|D^2 u_{1:n}\|_1 - C_n/n \right) = 0$*; (b)* $\gamma_t^-(u_t + 1) = 0$ *and* $\gamma_t^+(u_t - 1) = 0$ *for all* $t \in [n]$

Proof. By introducing auxiliary variables, we can re-write the offline optimization problem as:

$$\begin{aligned}
\min_{\tilde{u}_1,\dots,\tilde{u}_n} \quad & \sum_{t=1}^{n} f_t(\tilde{u}_t) \\
\text{s.t.} \quad & \tilde{z}_t = \tilde{u}_{t+2} - 2\tilde{u}_{t+1} + \tilde{u}_t \; \forall t \in [n-2] \\
& \sum_{t=1}^{n-2} |\tilde{z}_t| \leq C_n/n, \\
& -1 \leq \tilde{u}_t \; \forall t \in [n], \\
& \tilde{u}_t \leq 1 \; \forall t \in [n],
\end{aligned}$$

The Lagrangian of the optimization problem can be written as

$$\begin{aligned}
\mathcal{L}(\tilde{u}_{1:n}, \tilde{z}_{1:n-2}, \tilde{v}_{n-2}, \tilde{\gamma}^-_{1:n}, \tilde{\gamma}^+_{1:n}\tilde{\lambda}) &= \sum_{t=1}^{n} f_t(\tilde{u}_t) + \tilde{\lambda}\left(\sum_{t=1}^{n-2}|\tilde{z}_t| - C_n/n\right) \\
&+ \sum_{t=2}^{n-2} \tilde{v}_t(\tilde{u}_{t+2} - 2\tilde{u}_{t+1} + \tilde{u}_t - \tilde{z}_t) + \sum_{t=1}^{n} \gamma_t^+(\tilde{u}_t - 1) - \gamma_t^-(\tilde{u}_t + 1).
\end{aligned}$$

Due to stationary conditions wrt u_t, we have

$$\nabla f_t(u_t) = 2v_{t-1} - v_t - v_{t-2} + \gamma_t^- - \gamma_t^+,$$

where we define $v_{-1} = v_0 = v_{n-1} = v_n = 0$ and, due to staionarity conditions wrt v_t we have

$$v_t = \lambda \text{sign}(z_t).$$

Combining the above two equations and the complementary slackness rule now yields the Lemma.

□

Terminology. In what follows, we refer to $u_{1:n}$ from the Lemma above to be the offline optimal sequence.

Lemma 10. ***(key partition)*** *For some interval $[a, b] \in [n]$, define $\ell_{a\to b} := b - a + 1$. There exists a partitioning of the time horizon $\mathcal{P} := \{[1_s, 1_t], \dots, [i_s, i_t], \dots [M_s, M_t]\}$ where $M = |\mathcal{P}|$ such that for any bin $[i_s, i_t] \in \mathcal{P}$ we have: 1) $\|D^2 u_{i_s:i_t}\|_1 \leq 1/\ell^{3/2}_{i_s\to i_t}$; 2) $\|D^2 u_{i_s:i_t+1}\|_1 > 1/\ell^{3/2}_{i_s\to i_t+1}$ and 3) $M = O\left(n^{1/5}C_n^{2/5} \vee 1\right)$.*

Proof. Let the total number of bins formed be M. Consider the case where $M > 1$. We have that

$$\begin{aligned}
\|D^2 u_{1:n}\|_1 &\geq \sum_{i=1}^{M-1} \|D^2 u_{i_s\to i_t+1}\|_1 \\
&\geq_{(a)} 1/\ell^{3/2}_{i_s\to i_t+1} \\
&\geq_{(b)} \frac{(M-1)^{5/2}}{n^{3/2}},
\end{aligned}$$

where line (a) follows due to the construction of the partition and line (b) is due to Jensen's inequality applied to the convex fucntion $f(x) = 1/x^{3/2}$ for $x > 0$.

Rearranging and including the trivial case where $M = 1$ yields the lemma.

□

Proposition 19. *In the following analysis we will often use a useful represent offline optimal within a bin $[a, b]$ to be $m_a, m_a + m_{a+1}, \dots, \sum_{t=a}^{b} m_t$ WLOG. We can view this sequence to be samples obtained from a piece-wise linear signal that is continuous at every sampling point.*

Lemma 20. ***(residual bound)*** *Consider a bin $[a, b]$. Let $\ell := b - a + 1$. Define:*

$$X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ \vdots & \\ 1 & \ell \end{bmatrix}$$

Let $\boldsymbol{\beta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T u_{a:b}$ be the least square fit coefficient computed with labels u_t and co-variates $\boldsymbol{x}_t = [1, t - a + 1]^T$ where $t \in [a, b]$. Then we have that the residuals satisfy

$$|\boldsymbol{\beta}^T\boldsymbol{x}_t - u_t| \leq 20\ell \|D^2 u_{a:a+\ell-1}\|_1,$$

whenever $\ell \geq 6$.

Proof. We follow the notations of Proposition 19 for representing the offline optimal $u_a, \ldots, u_b$. The residual at time $i \in [a, b]$ can be computed through straight forward algebra as:

$$u_i - \boldsymbol{\beta}^T\boldsymbol{x}_i = \frac{1}{(\ell^2 - 1)} \sum_{j=2}^{\ell} \Bigg(6(1 + (1 - 2i)/\ell)(\ell - j + 1/)(\ell + j)/2 \\ + (6i + 6i/\ell - 4\ell - 6 - 2/\ell)(\ell - j + 1) + (\ell^2 - 1)\mathbb{I}\{j \leq i\} \Bigg) m_{a+j-1}, \tag{16}$$

where $\mathbb{I}\{\cdot\}$ is the indicator function assuming value 1 if the argument evaluates true and 0 otherwise. Now we note that if all m_k for $k \in [a + 1, b]$ are same, then the residuals $u_i - \boldsymbol{\beta}^T\boldsymbol{x}_i$ must be zero for all i as the least square fit exactly matches the labels in this case. In particular, this implies from Eq.(16) that

$$\frac{1}{(\ell^2 - 1)} \sum_{j=2}^{\ell} \Bigg(6(1 + (1 - 2i)/\ell)(\ell - j + 1/)(\ell + j)/2 \\ + (6i + 6i/\ell - 4\ell - 6 - 2/\ell)(\ell - j + 1) + (\ell^2 - 1)\mathbb{I}\{j \leq i\} \Bigg) m_{a+1} = 0. \tag{17}$$

Subtracting Eq.(17) from (16) we get,

$$\begin{aligned} u_i - \boldsymbol{\beta}^T\boldsymbol{x}_i &= \frac{1}{(\ell^2 - 1)} \sum_{j=2}^{\ell} \Bigg(6(1 + (1 - 2i)/\ell)(\ell - j + 1/)(\ell + j)/2 \\ &\quad + (6i + 6i/\ell - 4\ell - 6 - 2/\ell)(\ell - j + 1) + (\ell^2 - 1)\mathbb{I}\{j \leq i\} \Bigg) (m_{j+a-1} - m_{a+1}) \\ &\leq \frac{1}{(\ell^2 - 1)} \max_{j \in [a+2, b]} |m_j - m_{a+1}| \sum_{j=3}^{\ell} \Bigg| 6(1 + (1 - 2i)/\ell)(\ell - j + 1/)(\ell + j)/2 \\ &\quad + (6i + 6i/\ell - 4\ell - 6 - 2/\ell)(\ell - j + 1) + (\ell^2 - 1)\mathbb{I}\{j \leq i\} \Bigg|, \end{aligned}$$

where the last line is due to Holder's inequality. Further, we have $|m_j - m_{a+1}| \leq \sum_{t=j}^{a+2} |m_j - m_{j-1}| \leq \|D^2 u_{a:b}\|_1$ by the definition of the discrete difference operator D^2.

Now applying triangle inequality and the crude bounds $1 + (1 - 2i)/\ell \leq 3, (\ell - j + 1) \leq \ell, (\ell + j) \leq 2\ell, i/\ell \leq 1, 2\ell \geq 2/\ell$ and $-2/\ell \leq 0$ we obtain

$$\Bigg| 6(1+(1-2i)/\ell)(\ell-j+1/)(\ell+j)/2 + (6i+6i/\ell-4\ell-6-2/\ell)(\ell-j+1)+(\ell^2-1)\mathbb{I}\{j \le i\} \Bigg| \le 19\ell^2+2\ell.$$

So,

$$\begin{aligned}|u_i - \boldsymbol{\beta}^T\boldsymbol{x}_i| &\le \ell \cdot \frac{19\ell^2+2\ell}{\ell^2-1}\|D^2 u_{a:b}\|_1 \\ &\le 20\ell\|D^2u_{a:b}\|_1,\end{aligned}$$

where the last line is due to $19\ell^2 + 2\ell \le 20\ell^2 - 20$ for all $\ell \ge 6$.

□

Lemma 21. ***(bounding** T_3**)** *Consider a bin $[a,b]$ with length $\ell = b-a+1$ obtained from the scheme in Lemma 10. Assume the notations in Lemma 20. Let's represent the residual as $r_t := \boldsymbol{\beta}^T\boldsymbol{x}_t - u_t = (t-a+1)M_{t-1} + C_{t-1}$ for $t > a$ and $r_1 := \boldsymbol{\beta}^T\boldsymbol{x}_a - u_a = M_a + C_a$ with $M_b := M_{b-1} = M_{a+\ell-2}$ and $C_b := C_{b-1} = C_{a+\ell-2}$. Suppose $\|D^2u_{a:b}\|_1 \le \ell^{-3/2}$. We have,*

$$\begin{aligned}\sum_{t=a}^{b} f_t(\boldsymbol{\beta}^T\boldsymbol{x}_t) - f_t(u_t) \le 200 + \lambda\Bigg(&(s_{a-1}-s_{a-2})(M_a+C_a) - (s_b - s_{b-1})(\ell M_b + C_b) \\ &- s_{a-1}M_a + s_{b-1}M_{b-1} - \sum_{t=a+1}^{b}|M_t - M_{t-1}|\Bigg) \\ &+ 20\ell^{-1/2}\sum_{t=a}^{b}|\gamma_t^- - \gamma_t^+| \end{aligned} \tag{18}$$

Further we have $|M_a| \le \|D^2u_{a:b}\|_1$ and $|M_b| \le \|D^2u_{a:b}\|_1$ whenever $\ell \ge 2$.

Here the semantics is that each $M_t = r_{t+1} - r_t$ for all $t > a$ and $M_a = r_{a+1} - r_a$. Any two points r_t and r_{t+1} can be joined using a unique line segment which in turn defines C_t appropriately.

Proof. By gradient Lipschitzness of f we have

$$\sum_{t=a}^{b} f_t(\boldsymbol{\beta}^T\boldsymbol{x}_t) - f_t(u_t) \le \sum_{t=a}^{b}\langle \nabla f_t(u_t), \boldsymbol{\beta}^T\boldsymbol{x}_t - u_t\rangle + \sum_{t=a}^{b}\frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{x}_t - u_t)^2.$$

Now will focus on bounding the last two terms above.

From the construction of bins in Lemma 10, we know that $\ell\|D^2u_{a:b}\|_1 \le 1/\sqrt{\ell}$. Hence we obtain using Lemma 20 that

$$\sum_{t=a}^{b}\frac{1}{2}(\boldsymbol{\beta}^T\boldsymbol{x}_t - u_t)^2 \le 200.$$

Recall the representation of the residual $\boldsymbol{\beta}^T\boldsymbol{x}_t - u_t = tM_t + C_t$ mentioned in the lemma statement. Observe that in accordance with Proposition 19 this residual can also be viewed as samples of a piece-wise linear signal that is continuous at every sampled point. In particular observe that for every $t \in [a,b]$ we have:

$$(t-a+1)M_{t-1} + C_{t-1} = (t-a+1)M_t + C_t$$

Consequently

$$C_t - C_{t-1} = (t-a+1)(M_{t-1} - M_t) \tag{19}$$

From KKT conditions of Lemma 17 we have

$$\sum_{t=a}^{b} \langle \nabla f_t(u_t), \boldsymbol{\beta}^T \boldsymbol{x}_t - u_t \rangle = \underbrace{\sum_{t=a}^{b} \lambda \left(((s_{t-1} - s_{t-2}) - (s_t - s_{t-1})) \left((t-a+1) M_t + C_t \right) \right)}_{X_1}$$
$$+ \underbrace{\sum_{t=a}^{b} (\gamma_t^- - \gamma_t^+)(\boldsymbol{\beta}^T \boldsymbol{x}_t - u_t)}_{X_2}$$

$$\begin{aligned}
\frac{X_1}{\lambda} &= (s_{a-1} - s_{a-2})(M_a + C_a) - (s_b - s_{b-1})(\ell M_b + C_b) \\
&+ \sum_{t=a}^{b-1} (s_t - s_{t-1}) \left((t-a+2) M_{t+1} + C_{t+1} - ((t-a+1) M_t + C_t) \right) \\
&=_{(a)} (s_{a-1} - s_{a-2})(M_a + C_a) - (s_b - s_{b-1})(\ell M_b + C_b) + \sum_{t=a}^{b-1} (s_t - s_{t-1}) M_{t+1} \\
&= (s_{a-1} - s_{a-2})(M_a + C_a) - (s_b - s_{b-1})(\ell M_b + C_b) + \sum_{t=a}^{b-1} (M_{t+1} - M_{t+2}) s_t - s_{a-1} M_2 + s_{b-1} M_\ell \\
&=_{(b)} (s_{a-1} - s_{a-2})(M_a + C_a) - (s_b - s_{b-1})(\ell M_b + C_b) - s_{a-1} M_a + s_{b-1} M_{b-1} - \sum_{t=a+1}^{b} |M_t - M_{t-1}|,
\end{aligned}$$

where in line (a) we used Eq.(19) and in line (b) we used the fact that $s_t = \text{sign}((u_{t+2} - u_{t+1}) - (u_{t+1} - u_t)) = \text{sign}(M_{t+2} - M_{t+1})$ along with the fact that $M_a = M_{a+1}$ and $M_{b-1} = M_b$.

By Holder's inequality and Lemma 20, we have

$$\begin{aligned}
X_2 &\leq 20\ell \|D^2 u_{a:b}\|_1 \sum_{t=a}^{b} |\gamma_t^- - \gamma_t^+| \\
&\leq 20\ell^{-1/2} \sum_{t=a}^{b} |\gamma_t^- - \gamma_t^+|,
\end{aligned}$$

where the last line is due to $\|D^2 u_{a:b}\|_1 \leq \ell^{-3/2}$ as assumed in the lemma's statement. Putting everything together completes the proof.

Next, we proceed to give useful bounds on $|M_a|$ and $|M_{b-1}|$.

Since $M_a = M_{a+1}$ and $C_a = C_{a+1}$, we have $M_a = (u_{a+1} - \boldsymbol{\beta}^T \boldsymbol{x}_{a+1}) - (u_a - \boldsymbol{\beta}^T \boldsymbol{x}_a)$.So Eq.(16) we have,

$$\begin{aligned}
|M_a| &= \left| \sum_{j=2}^{\ell} \frac{6(\ell - j + 1)(1 - j)}{\ell^3 - \ell} (m_{j+a-1} - m_{a+1}) \right| \\
&\leq \|D^2 u_{a:b}\|_1 \sum_{j=3}^{\ell} \frac{6(\ell - j + 1)(j - 1)}{\ell^3 - \ell} \\
&= \|D^2 u_{a:b}\|_1 \frac{\ell^2 + \ell - 6}{\ell(\ell + 1)} \\
&\leq \|D^2 u_{a:b}\|_1,
\end{aligned}$$

where in the last line we used $\frac{\ell^2+\ell-6}{\ell(\ell+1)} \leq 1$ for all $\ell \geq 2$.

Similarly $M_{b-1} = u_b - \boldsymbol{\beta}^T \boldsymbol{x}_b - (u_{b-1} - \boldsymbol{\beta}^T \boldsymbol{x}_{b-1})$ by recalling that $M_b = M_{b-1}$ and $C_b = C_{b-1}$. Proceeding from Eq.(16) we obtain,

$$
\begin{aligned}
|M_{b-1}| &= \left| \sum_{j=2}^{\ell-1} \frac{6(\ell-j+1)(1-j)}{\ell^3-\ell} (m_{j+a-1} - m_b) \right| \\
&\leq \|D^2 u_{a:b}\|_1 \sum_{j=2}^{\ell-1} \frac{6(\ell-j+1)(j-1)}{\ell^3-\ell} \\
&= \|D^2 u_{a:b}\|_1 \frac{\ell^2+\ell-6}{\ell(\ell+1)} \\
&\leq \|D^2 u_{a:b}\|_1.
\end{aligned}
$$

□

Lemma 22. *Consider a bin $[a,b] \in \mathcal{P}$ of length ℓ from Lemma 10. Suppose $|u_a| < 1$. Then either $\gamma_j^- = 0$ or $\gamma_j^+ = 0$ for all $j \in [a,b]$.*

Proof. We will provide arguments for the case when the offline optimal first hits -1 before hitting 1 for some point in $[a,b]$. The arguments for the alternate case where it hits 1 first are similar.

If the offline optimal hits -1 at some point in $[a,b]$ it can only rise upto at-most $-1 + 1/\sqrt{l}$ afterwards. This is due to the constraint $\|D^2 u_{a:b}\|_1 \leq 1/\ell^{3/2}$.

Since $-1 + 1/\sqrt{l} < 1$ as $\ell > 1/4$, we have that the offline optimal never touches 1 within the bin $[a,b]$. Consequently $\gamma_j^+ = 0$ for all $j \in [a,b]$.

□

Definition 23. *The **slope** of the optimal solution at a time point t is defined to be $u_{t+1} - u_t$ for all $t \in [n-1]$.*

Proposition 24. *The bins in $\mathcal{P}$ can be further refined in such a way that each bin either satisfy the condition in Lemma 22 or has constant slope, meaning the L1 TV distance is zero. Further in doing so the size of partition $\mathcal{P}$ only gets increased by at-most 2.*

Proof. Suppose for a bin $[a,b] \in \mathcal{P}$, if the offline optimal starts at 1. Then we can split that bin into two bins $[a,c]$ and $[c+1,b]$ such that $u_c > -1$ and $\|D^2 u_{a:c}\|_1 = 0$. Similar splitting can also be done for bins that start from -1. Observe that this refinement only increases the partition size only by at-most 2. □

Corollary 25. *One powerful consequence of Lemma 22 and Proposition 24 when combined with the fact that γ_t^- and γ_t^+ are both non-negative (Lemma 17) is that $\sum_{t=a}^b |\gamma_t^- - \gamma_t^+|$ is either equal to $\sum_{t=a}^b \gamma_t^-$ or $\sum_{t=a}^b \gamma_t^+$ for all bins $[a,b]$ in the refined partition of Proposition 24 whenever the $\|D^2 u_{a:b}\|_1 > 0$.*

From here on WLOG we will assume that the bins $[a,b]$ in partition $\mathcal{P}$ will satisfy the conditions:

- $\|D^2 u_{a:b}\|_1 \leq 1/l^{3/2}$, where $\ell = b-a+1$.
- It satisfies the conditions mentioned in Proposition 24 and consequently satisfying the condition in Corollary 25.
- $|\mathcal{P}| = O(n^{1/5} C_n^{2/5})$.

Lemma 26. *(**bounding** T_2) Consider a bin $[a,b] \in \mathcal{P}$ with length $\ell = b-a+1$ that doesn't touch boundary 1. Let $\Gamma = \sum_{j=a}^b \gamma_j^-$ and $\tilde{\Gamma} = \sum_{j=a}^b j' \gamma_j^-$ where $j' := j-a+1$. Let $\boldsymbol{\beta}$ be as in Lemma 20.*

Let $F(\boldsymbol{\beta}) := \sum_{j=a}^{b} f_j(\boldsymbol{x}_j^T \boldsymbol{\beta})$. Define:

$$\boldsymbol{A} := \sum_{j=a}^{b} \boldsymbol{x}_j \boldsymbol{x}_j^T$$

Consider the following update:

$$\begin{aligned}\boldsymbol{\alpha} &= \boldsymbol{\beta} - \boldsymbol{A}^{-1} \sum_{j=a}^{b} f_j'(\boldsymbol{x}_j^T \boldsymbol{\beta}) \boldsymbol{x}_j \\ &= \boldsymbol{\beta} - \boldsymbol{A}^{-1} \nabla F(\boldsymbol{\beta})\end{aligned}$$

We have,

$$\begin{aligned}2L\left(F(\boldsymbol{\alpha}) - F(\boldsymbol{\beta})\right) \leq -\|\boldsymbol{g}\|_{\boldsymbol{A}^{-1}}^2 - \|\boldsymbol{h}\|_{\boldsymbol{A}^{-1}}^2 - 2 < \boldsymbol{g}, \boldsymbol{A}^{-1}\boldsymbol{h}\rangle \\ + 2\langle \boldsymbol{A}^{-1}(\boldsymbol{g}+\boldsymbol{h}), \sum_{j=a}^{b} \boldsymbol{x}_j \left(f_j'(\boldsymbol{x}_j^T \boldsymbol{\beta}) - f_j'(u_j)\right),\end{aligned}$$

where $\boldsymbol{g} = \lambda[-s_{a-2} + s_{a-1} + s_{b-1} - s_b, -s_{a-2} + (\ell+1)s_{b-1} - \ell s_b]^T$ and $\boldsymbol{h} = [\Gamma, \tilde{\Gamma}]^T$ so that $\sum_{j=a}^{b} f_j'(u_j)\boldsymbol{x}_j = \boldsymbol{g} + \boldsymbol{h}$. Further we have:

- $\|\boldsymbol{g}\|_{\boldsymbol{A}^{-1}}^2$ *as in Eq.* (23)
- $\|\boldsymbol{h}\|_{\boldsymbol{A}^{-1}}^2$ *as in Eq.* (29)
- $\langle \boldsymbol{A}^{-1}\boldsymbol{g}, \boldsymbol{h}\rangle$ *as in Eq.* (30)
- $\langle \boldsymbol{A}^{-1}\boldsymbol{g}, \sum_{j=a}^{b} \boldsymbol{x}_j (f_j'(\boldsymbol{x}_j^T \boldsymbol{\beta}) - f_j'(u_j))\rangle$ *bounded above by Eq.*(27)
- $\langle \boldsymbol{A}^{-1}\boldsymbol{h}, \sum_{j=a}^{b} \boldsymbol{x}_j (f_j'(\boldsymbol{x}_j^T \boldsymbol{\beta}) - f_j'(u_j))\rangle$ *bounded above by Eq.*(28)

Similar expressions can be derived for bins $[a, b]$ that may touch boundary 1 instead of -1.

Proof. We note that due to gradient Lipschitnzess of f,

$$\nabla^2 F(\boldsymbol{\beta}) = \sum_{j=a}^{b} f_j''(\boldsymbol{x}_j^T \boldsymbol{\beta}) \boldsymbol{x}_j \boldsymbol{x}_j^T \preccurlyeq \boldsymbol{A}$$

So by Taylor's theorem we have for some $\boldsymbol{z} = t\boldsymbol{\alpha} + (1-t)\boldsymbol{\beta}$

$$\begin{aligned}F(\boldsymbol{\alpha}) - F(\boldsymbol{\beta}) &= -\langle \nabla F(\boldsymbol{\beta}), \boldsymbol{A}^{-1}\nabla F(\boldsymbol{\beta})\rangle + \frac{1}{2}\|\boldsymbol{A}^{-1}\nabla F(\boldsymbol{\beta})\|_{\nabla^2 F(\boldsymbol{z})}^2 \\ &\leq -\langle \nabla F(\boldsymbol{\beta}), \boldsymbol{A}^{-1}\nabla F(\boldsymbol{\beta})\rangle + \frac{1}{2}\|\boldsymbol{A}^{-1}\nabla F(\boldsymbol{\beta})\|_{\boldsymbol{A}}^2 \\ &= -\frac{1}{2}\|\nabla F(\boldsymbol{\beta})\|_{\boldsymbol{A}^{-1}}^2, \qquad (20)\end{aligned}$$

where

$$\boldsymbol{A}^{-1} = \frac{2}{(\ell-1)\ell} \begin{bmatrix} 2\ell+1 & -3 \\ -3 & \frac{6}{\ell+1} \end{bmatrix}$$

Next we turn to lower bounding the above RHS

$$\begin{aligned}\|\nabla F(\boldsymbol{\beta})\|^2_{\boldsymbol{A}^{-1}} &= \|\sum_{j=a}^{b} f_j'(u_j)\boldsymbol{x}_j + \sum_{j=a}^{b}(f_j'(\boldsymbol{x}_j^T\boldsymbol{\beta}) - f_j'(u_j))\boldsymbol{x}_j\|^2_{\boldsymbol{A}^{-1}}\\ &\geq \|\sum_{j=a}^{b} f_j'(u_j)\boldsymbol{x}_j\|^2_{\boldsymbol{A}^{-1}} - 2\langle \boldsymbol{A}^{-1}\sum_{j=a}^{b} f_j'(u_j)\boldsymbol{x}_j, \sum_{j=a}^{b}(f_j'(\boldsymbol{x}_j^T\boldsymbol{\beta}) - f_j'(u_j))\boldsymbol{x}_j\rangle\end{aligned}$$

From the KKT conditions in Lemma 17, we have

$$\sum_{j=a}^{b} f_j'(u_j)\boldsymbol{x}_j = [\lambda(-s_{a-2} + s_{a-1} + s_{b-1} - s_b) + \Gamma, \lambda(-s_{a-2} + (\ell+1)s_{b-1} - \ell s_b) + \tilde{\Gamma}]^T, \tag{21}$$

where Γ and $\tilde{\Gamma}$ are as defined in the statement of the lemma.

For the sake of brevity let's denote $\boldsymbol{g} = \lambda[-s_{a-2} + s_{a-1} + s_{b-1} - s_b, -s_{a-2} + (\ell+1)s_{b-1} - \ell s_b]^T$ and $\boldsymbol{h} = [\Gamma, \tilde{\Gamma}]^T$ so that $\sum_{j=a}^{b} f_j'(u_j)\boldsymbol{x}_j = \boldsymbol{g} + \boldsymbol{h}$.

We have

$$\begin{aligned}\boldsymbol{A}^{-1}\boldsymbol{g} = \frac{2\lambda}{(\ell-1)\ell}[&(2-2\ell)s_{a-2} + (2\ell+1)s_{a-1} - (\ell+2)s_{b-1} + (\ell-1)s_b,\\ &\frac{3(\ell-1)}{\ell+1}s_{a-2} - 3s_{a-1} + 3s_{b-1} - \frac{3(\ell-1)}{\ell+1}s_b]^T,\end{aligned} \tag{22}$$

and so

$$\begin{aligned}\|\boldsymbol{g}\|^2_{\boldsymbol{A}^{-1}} = \frac{2\lambda^2}{(\ell-1)\ell(\ell+1)}\Bigg(&(2\ell^2 - 3\ell + 1)s_{a-2}^2 + (4 - 4\ell^2)s_{a-2}s_{a-1} - (2\ell^2 - 6\ell + 4)s_{a-2}s_b +\\ &(2\ell^2 - 2)s_{a-2}s_{b-1} + (2\ell^2 + 3\ell + 1)s_{a-1}^2 +\\ &(2\ell^2 - 2)s_{a-1}s_b - (2\ell^2 + 6\ell + 4)s_{a-1}s_{b-1} +\\ &(2\ell^2 - 3\ell + 1)s_b^2 + (4 - 4\ell^2)s_{b-1}s_b + (2\ell^2 + 3\ell + 1)s_{b-1}^2\Bigg)\end{aligned} \tag{23}$$

Using Eq. (22) we get

$$\begin{aligned}\langle \boldsymbol{A}^{-1}\boldsymbol{g}, \sum_{j=a}^{b}\boldsymbol{x}_j(f_j'(\boldsymbol{x}_j^T\boldsymbol{\beta}) - f_j'(u_j))\rangle &= \frac{2\lambda}{(\ell-1)\ell}\Bigg((2-2\ell)s_{a-2} + (2\ell+1)s_{a-1}\\ &\quad - (\ell+2)s_{b-1} + (\ell-1)s_b\Bigg)\sum_{j=a}^{b}(f_j'(\boldsymbol{x}_j^T\boldsymbol{\beta}) - f_j'(u_j))\\ &\quad + \frac{2\lambda}{(\ell-1)\ell}\Bigg(\frac{3(\ell-1)}{\ell+1}s_{a-2} - 3s_{a-1}\\ &\quad + 3s_{b-1} - \frac{3(\ell-1)}{\ell+1}s_b\Bigg)\sum_{j=a}^{b} j'(f_j'(\boldsymbol{x}_j^T\boldsymbol{\beta}) - f_j'(u_j)).\end{aligned} \tag{24}$$

Using gradient Lipschitzness, triangle inequality and Lemma 20 we have

$$\sum_{j=a}^{b}(f_j'(\boldsymbol{x}_j^T\boldsymbol{\beta}) - f_j'(u_j)) \leq \sum_{j=a}^{b}|\boldsymbol{x}_j^T\boldsymbol{\beta} - u_j| \leq 20\ell^2\|D^2 u_{a:b}\|_1, \tag{25}$$

and similarly

$$\sum_{j=a}^{b} j'(f_j'(\boldsymbol{x}_j^T\boldsymbol{\beta}) - f_j'(u_j)) \leq \sum_{j=a}^{b} Lj'|\boldsymbol{x}_j^T\boldsymbol{\beta} - u_j| \leq 20\ell^3\|D^2 u_{a:b}\|_1. \tag{26}$$

So continuing from Eq. (24),

$$\begin{aligned}\langle \boldsymbol{A}^{-1}\boldsymbol{g}, \sum_{j=a}^{b}\boldsymbol{x}_j(f_j'(\boldsymbol{x}_j^T\boldsymbol{\beta}) - f_j'(u_j))\rangle &\leq \frac{40\lambda\ell\|D^2u_{a:b}\|_1}{(\ell-1)}\left|(2-2\ell)s_{a-2} + (2\ell+1)s_{a-1} - (\ell+2)s_{b-1} + (\ell-1)s_b\right| \\ &+ \frac{40\lambda\ell^2\|D^2u_{a:b}\|_1}{(\ell-1)}\left|\frac{3(\ell-1)}{\ell+1}s_{a-2} - 3s_{a-1} + 3s_{b-1} - \frac{3(\ell-1)}{\ell+1}s_b\right| \\ &\leq \frac{40\lambda\ell^{-1/2}}{(\ell-1)}\left|(2-2\ell)s_{a-2} + (2\ell+1)s_{a-1} - (\ell+2)s_{b-1} + (\ell-1)s_b\right| \\ &+ \frac{40\lambda\ell^{1/2}}{(\ell-1)}\left|\frac{3(\ell-1)}{\ell+1}s_{a-2} - 3s_{a-1} + 3s_{b-1} - \frac{3(\ell-1)}{\ell+1}s_b\right|,\end{aligned} \tag{27}$$

where we used $\|D^2u_{a:b}\|_1 \leq \ell^{-3/2}$.

We have

$$\begin{aligned}\langle \boldsymbol{A}^{-1}\boldsymbol{h}, \sum_{j=a}^{b}\boldsymbol{x}_j\left(f_j'(\boldsymbol{\beta}^T\boldsymbol{x}_j) - f_j'(u_j)\right) &= \frac{2}{(\ell-1)\ell}\Bigg(((2\ell+1)\Gamma - 3\tilde{\Gamma})\sum_{j=a}^{b} f_j'(\boldsymbol{\beta}^T\boldsymbol{x}_j) - f_j'(u_j) \\ &+ \left(\frac{6\tilde{\Gamma}}{\ell+1} - 3\Gamma\right)\sum_{j=a}^{b}(j-a+1)\left(f_j'(\boldsymbol{\beta}^T\boldsymbol{x}_j) - f_j'(u_j)\right)\Bigg) \\ &\leq \frac{40\ell\|D^2u_{a:b}\|_1}{(\ell-1)}|(2\ell+1)\Gamma - 3\tilde{\Gamma}| + \frac{40\ell^2\|D^2u_{a:b}\|_1}{(\ell-1)}\left|\frac{6\tilde{\Gamma}}{\ell+1} - 3\Gamma\right| \\ &\leq \frac{40\ell^{-1/2}}{(\ell-1)}|(2\ell+1)\Gamma - 3\tilde{\Gamma}| + \frac{40\ell^{1/2}}{(\ell-1)}\left|\frac{6\tilde{\Gamma}}{\ell+1} - 3\Gamma\right|,\end{aligned} \tag{28}$$

where the last line is obtained by using similar arguments used for obtaining Eq.(27).

By substituting the expression for $\boldsymbol{A}^{-1}$ and simplifying,

$$\|\boldsymbol{h}\|^2_{\boldsymbol{A}^{-1}} = \frac{2}{(\ell-1)\ell(\ell+1)}\left((2\ell+1)(\ell+1)\Gamma^2 - 6\Gamma\tilde{\Gamma}(\ell+1) + 6\tilde{\Gamma}^2\right). \tag{29}$$

Using Eq.(22), we obtain

$$\langle \boldsymbol{A}^{-1}\boldsymbol{g}, \boldsymbol{h}\rangle = \sum_{j=a}^{b} \frac{2\lambda\gamma_j^-}{(\ell-1)\ell}\Bigg(\frac{-3j+3\ell j'-2\ell^2+2}{\ell+1}s_{a-2} + (-3j'+2\ell+1)s_{a-1}$$
$$+ (3j'-\ell-2)s_{b-1} + \frac{-3\ell j+3j'+\ell^2-1}{\ell+1}s_b\Bigg) \tag{30}$$

□

Lemma 27. *(**bounding** T_1) Consider a bin $[a,b]$. Let $\mathbf{p}_t$ be the predictions of FLH-SIONS algorithm with parameters $\epsilon = 2$, $C = 20$ and exp-concavity factor σ. Suppose $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are as defined in Lemma 26. For any $\boldsymbol{\mu} \in \{\boldsymbol{\alpha}, \boldsymbol{\beta}\}$ FLH-SIONS satsifies:*

$$\sum_{t=a}^{b} f_t(p_t) - f_t(\boldsymbol{\mu}^T\boldsymbol{x}_t) \le 256 + \frac{1}{2\sigma}\log(1+\sigma n/2) + \frac{4}{\sigma}\log n.$$

Proof. We will derive the guarantee for $\boldsymbol{\mu} = \boldsymbol{\alpha}$. The guarantee for $\boldsymbol{\mu} = \boldsymbol{\alpha}$ follows similarly.

Let's begin by calculating $\boldsymbol{v} := A^{-1}\sum_{j=a}^{b} f_j'(\boldsymbol{x}_j^T\boldsymbol{\beta})\boldsymbol{x}_j$.

We have,

$$\begin{aligned}|\boldsymbol{v}[1]| &= \left|\frac{2}{(\ell-1)\ell}\sum_{j=1}^{\ell}(2\ell+1-3j)f'_{(j+a-1)}(\boldsymbol{x}_{j+a-1}^T\boldsymbol{\beta})\right| \\ {}_{(a)} &\le \frac{2}{(\ell-1)\ell}\cdot 2\ell(\ell-1) \\ &= 4,\end{aligned} \tag{31}$$

where line (a) is obtained via Lipschitzness and Holder's inequality $\boldsymbol{x}^T\boldsymbol{y} \le \|\boldsymbol{x}\|_1\|\boldsymbol{y}\|_\infty$ and the fact that $|2\ell+1-3j| \le 2(\ell-1)$ for all $j \in [1,\ell]$.

Similarly

$$\begin{aligned}|\boldsymbol{v}[2]| &= \left|\frac{2}{(\ell-1)\ell(\ell+1)}\sum_{j=1}^{\ell}(-3(\ell+1)+6j)f'_{(j+a-1)}(\boldsymbol{x}_{j+a-1}^T\boldsymbol{\beta})\right| \\ &\le \frac{2}{(\ell-1)\ell(\ell+1)}\cdot 3\ell(\ell-1) \\ &= \frac{6}{(\ell+1)},\end{aligned} \tag{32}$$

where we used $|-3(\ell+1)+6j| \le 3(\ell-1)$ for all $j \in [1,\ell]$.

Combining Eq.(31) and (32) we conclude that

$$\begin{aligned}|\boldsymbol{v}^T\boldsymbol{x}_j| &\le 4 + (j-a+1)\frac{6}{(\ell+1)} \\ &\le 10,\end{aligned}$$

where the last line follows due to the fact $(j-a+1) \le \ell$.

Hence by Triangle inequality we have,

$$|\boldsymbol{\alpha}^T\boldsymbol{x}_j| \le |\boldsymbol{\beta}^T\boldsymbol{x}_j| + 10. \tag{33}$$

Further note that

$$\|\boldsymbol{v}\|_2 \le 8$$

Notice that $\boldsymbol{\beta} = \boldsymbol{A}^{-1}\sum_{j=a}^{\ell} u_j \boldsymbol{x}_j$ which have similar functional form as $\boldsymbol{v}$. Since $|u_j|\leq B$ for all $j \in [n]$, by following similar arguments used in bounding $\boldsymbol{v}$ we obtain $|\boldsymbol{\beta}^T\boldsymbol{x}_j|\leq 10$ and

$$\|\boldsymbol{\beta}\|_2\leq 8.$$

Continuing from (33) we get

$$|\boldsymbol{\alpha}^T\boldsymbol{x}_j| \leq 20. \tag{34}$$

Further,

$$\begin{aligned}\|\boldsymbol{\alpha}\|_2 &\leq \|\boldsymbol{\beta}\|_2+\|\boldsymbol{v}\|_2\\ &\leq 16 \end{aligned}\tag{35}$$

Since the losses f_t are σ exp-concave in $[-1,1]$, b+y Theorem 2 in (Luo et al., 2016) and Lemma 3.3 in (Hazan and Seshadhri, 2007), FLH-SIONS with parameters set as in the statement of the Lemma yields a regret of

$$\sum_{t=a}^{b} f_t(p_t) - f_t(\boldsymbol{\alpha}^T\boldsymbol{x}_t) \leq 256 + \frac{1}{2\sigma}\log(1+\sigma n/2) + \frac{4}{\sigma}\log n.$$

□

Lemma 28. ***(monotonic slopes)*** *Consider a bin $[i_s, i_t] \in \mathcal{P}$ such that the slopes are monotonic (i.e either non-decreasing or non-increasing). Let p_j be the predictions made by the FLH-SIONS algorithm with parameters as set in Lemma 27. Then we have,*

$$\begin{aligned}\sum_{j=i_s}^{i_t} f_j(p_j) - f_j(u_j) &\leq O\left(\frac{1}{2\sigma}\log(1+\sigma n/2) + \frac{4}{\sigma}\log n + 210408\right)\\ &= \tilde{O}(1)\end{aligned}$$

Proof. We will consider the case of non-decreasing slopes. The alternate case can be handled similarly.

Assume that the slope within the bin is not constant, otherwise we trivially get logarithmic regret as we need only to compete with the best fixed linear fit which is handled by the static regret of FLH-SIONS in any interval ($\boldsymbol{\mu} = \boldsymbol{\beta}$ in Lemma 27).

The optimal solution within a bin of $\mathcal{P}$ obtained via Proposition 24 which doesn't have constant slope may touch either -1 or 1 but not both. Consider the case where the optimal touches -1. Then as the slopes are non-decreasing, once it leaves -1, it never touches -1 again. So we can split the bin $[i_s, i_t]$ into at-most 3 bins $[a, b]$, $[b+1, c]$ and $[c+1, d]$ such that the optimal touches -1 only within $[b+1, c]$. (This bin can be empty if the optimal doesn't touch -1 anywhere within $[i_s, i_t]$).

Now we will bound the regret within bin $[a, b]$.

Suppose that $s_{a-1} = 1$ and $s_b = 1$. If this condition is not satisfied, we can refine the bin $[a, b]$ into at-most 3 bins $[a_1, b_1], [a_2, b_2], [a_3, b_3]$ such that the optimal has constant slope in the first and last bins and $s_{a_2-1} = s_{b_2} = 1$. This is possible because the slopes in $[a, b]$ are non-decreasing.

Let $\Delta := \|D^2 u_{a:b}\|_1$ and $\ell := b - a + 1$. Let p and q be two numbers in $[0, 2]$. Substituting $s_{a-2} = 1 - p$, $s_{a-1} = 1$, $s_{b-1} = 1 - q$ and $s_b = 1$ into Lemma 21 and using the fact that $|jM_j + C_j|\leq 20\ell\Delta$ for all $j \in [a, b]$ due to Lemma 20, we get

$$T_3 \leq 40\lambda(p+q)\ell\Delta + 200, \tag{36}$$

where we observed that a term arising from Lemma 21: $-M_a + M_{b-1} - \sum_{t=a+1}^{b} |M_t - M_{t-1}| = 0$ as the slopes are non-decreasing.

By making similar sign substitutions in Lemma 26 and noting that $\boldsymbol{h} = \boldsymbol{0}$, we get

$$
\begin{aligned}
T_2 &\leq \frac{-2\lambda^2}{2(\ell-1)\ell(\ell+1)} \left((2\ell^2 - 3\ell + 1)p^2 + (2\ell^2 + 3\ell + 1)q^2 + 2(\ell^2 - 1)pq \right) \\
&+ \frac{40\lambda\ell\Delta}{\ell - 1} \left(2p(\ell - 1) + q(\ell + 2) \right) + \frac{40\lambda\ell^2\Delta}{(\ell-1)(\ell+1)} \left(p(\ell - 1) + q(\ell + 1) \right) \\
&\leq \frac{-2\lambda^2}{2(\ell-1)\ell(\ell+1)} \left((2\ell^2 - 3\ell + 1)p^2 + (2\ell^2 + 3\ell + 1)q^2 + 2(\ell^2 - 1)pq \right) \\
&+ 160\lambda\ell\Delta(p+q) + 160\lambda\ell\Delta(p+q), \qquad (37)
\end{aligned}
$$

where in the last line we used the fact that $\ell - 1 \geq \ell/2$ and $\ell + 2 \leq 2\ell$ for all $\ell \geq 2$.

Now consider the case where $p \geq q$. Then,

$$
\begin{aligned}
(2\ell^2 - 3\ell + 1)p^2 + (2\ell^2 + 3\ell + 1)q^2 + 2(\ell^2 - 1)pq &\geq (2\ell^2 - 3\ell + 1)p^2 \\
&\geq \ell^2 p^2, \qquad (38)
\end{aligned}
$$

where the last line holds for all $\ell \geq 3$. (If $\ell \leq 3$, the regret within the bin is trivially $O(1)$ appealing to the Lipschitzness of the losses f_t and the boundedness of the predictions and the comparators (see proof of Lemma 27)). Thus by using $\ell - 1 \leq \ell$ and $\ell + 1 \leq 2\ell$, we get

$$
\frac{-2\lambda^2}{2(\ell-1)\ell(\ell+1)} \left((2\ell^2 - 3\ell + 1)p^2 + (2\ell^2 + 3\ell + 1)q^2 + 2(\ell^2 - 1)pq \right) \leq \frac{-\lambda^2 p^2}{2\ell}. \qquad (39)
$$

Combining Eq. (37) and (39) and using the fact that $p \geq q$, we have

$$
T_2 \leq \frac{-\lambda^2 p^2}{2\ell} + 640\lambda\ell\Delta p. \qquad (40)
$$

Similarly from (36) using $p \geq q$ we get

$$
\begin{aligned}
T_3 &\leq 40\lambda(p+q)\ell\Delta + 200 \\
&\leq 80\lambda p\ell\Delta + 200 \qquad (41)
\end{aligned}
$$

Combining Eq. (40) and (41) we have

$$
\begin{aligned}
T_2 + T_3 &\leq \frac{-\lambda^2 p^2}{2\ell} + 648\lambda p\ell\Delta + 200 \\
&= -\left(\frac{\lambda p}{\sqrt{2\ell}} - 648\sqrt{2}\ell^{3/2}\Delta \right) + 209952\ell^3\Delta^2 + 200 \\
&\leq 210152,
\end{aligned}
$$

where in the last line we dropped the negative term and used the facts that $\Delta \leq 1/\ell^{3/2}$.

For the case of $q \geq p$, we have

$$
\begin{aligned}
(2\ell^2 - 3\ell + 1)p^2 + (2\ell^2 + 3\ell + 1)q^2 + 2(\ell^2 - 1)pq &\geq (2\ell^2 - 3\ell + 1)q^2 \\
&\geq \ell^2 q^2,
\end{aligned}
$$

where the last line holds for all $\ell \geq 3$. This is the same expression as in Eq.(38) with p replaced by q. By replacing p with q in the arguments we detailed for the case of $p \geq q$ earlier, we arrive at the same conclusion that $T_2 + T_3 \leq 210152$ even when $q \geq p$. (If $\ell \leq 3$, the regret within the bin is trivially $O(1)$ appealing to the Lipschitzness of the losses f_t and the boundedness of the predictions and the comparators (see proof of Lemma 27))

Similar bound on $T_2 + T_3$ can be shown for bin $[c+1, d]$ by essentially the same arguments.

Hence through Lemma 27 we have the dynamic regret in bins $[a, b]$ to be:

$$\sum_{t=a}^{b} f_t(p_t) - f_t(u_t) \leq 256 + \frac{1}{2\sigma} \log(1 + \sigma n/2) + \frac{4}{\sigma} \log n + 210152$$
$$= \tilde{O}(1)$$

Similarly, the regret within bin $[c+1, d]$ is also bounded by the above expression.

As the slope within bin $[b+1, c]$ is constant, the regret incurred within this bin is trivially bounded by $256 + \frac{1}{2\sigma} \log(1 + \sigma n/2) + \frac{4}{\sigma} \log n$ due to Lemma 27.

Adding the regret incurred across the bins $[a, b]$, $[b+1, c]$ and $[c+1, d]$ together yields the lemma.

□

Next, we will focus on bounding $T_2 + T_3$ for general non-monotonic bins in $\mathcal{P}$.

Lemma 29. *(**non-monotonic slopes**) Consider a bin $[i_s, i_t] \in \mathcal{P}$ such that the slopes are not monotonic. Let p_j be the predictions made by the FLH-SIONS algorithm with parameters as set in Lemma 27. Then we have,*

$$\sum_{j=i_s}^{i_t} f_j(p_j) - f_j(u_j) \leq O\left(\frac{1}{\sigma} \log(1 + \sigma n) + \frac{12}{\sigma} \log n + 1\right)$$
$$= \tilde{O}(1)$$

Proof. Let $[a, b] \in \mathcal{P}$ be a bin where the slope is not monotonic and not constant.

Assume that $|s_{a-1}| = |s_b| = 1$. Otherwise we can split the original bin into at-most 3 bins $[a, b_1 - 1], [b_1, b_2], [b_2 + 1, b]$ such that $|s_{b_1 - 1}| = |s_{b_2}| = 1$ and slopes are constant in the the other two bins. This is possible because slope in $[a, b]$ is not constant or monotonic.

For a bin $[a, b]$ we define **boundary signs** to be $s_{a-2}, s_{a-1}, s_{b-1}$ and s_b.

First, we will study the case where the offline optimal touches the boundary -1 at two point r and w with $r < w$. The case of arbitrary number of boundary touches will be discussed towards the end. (All arguments can be mirrored appropriately for the case where optimal touches boundary 1).

In what follows we use the notations in the proof of Lemma 26. From Eq.(21) we have

$$\boldsymbol{g} + \boldsymbol{h} = \lambda \boldsymbol{\mu} + \gamma_r^- \boldsymbol{x}_r + \gamma_w^- \boldsymbol{x}_w, \tag{42}$$

where $\boldsymbol{\mu} \in \mathbb{R}^2$ is a vector depending on the boundary signs and the length $\ell := b - a + 1$. $\boldsymbol{x}_r = [1, r - a + 1]^T$ and $\boldsymbol{x}_w$ defined similarly.

Since $\boldsymbol{g} + \boldsymbol{h}$ is an affine map of $[\lambda, \gamma_r^-, \gamma_w^-]^T$ and since $\boldsymbol{A}$ is positive definite for $\ell \geq 2$, we conclude that $\|\boldsymbol{g} + \boldsymbol{h}\|^2_{\boldsymbol{A}^{-1}}$ is jointly convex in $\lambda, \gamma_r^-, \gamma_w^-$ via appealing to the convexity of squared L2 norm.

First let's focus on the case where boundary signs obey $s_{a-1} = 1$ and $s_b = -1$. Let $s_{a-2} = 1 - p$ and $s_{b-1} = -1 + q$ for some $p, q \in [0, 2]$.

Making these sign substitutions in Lemma 26, we get:

$$\|\boldsymbol{g}\|^2_{\boldsymbol{A}^{-1}} = \frac{2\lambda^2}{(\ell - 1)\ell(\ell + 1)} \left((2\ell^2 - 3\ell + 1)p^2 + (2\ell^2 + 3\ell + 1)q^2 - (2\ell^2 - 2)pq + 12(\ell - 1)p - 12(\ell + 1)q + 24\right).$$

$$\langle \boldsymbol{g}, \boldsymbol{A}^{-1}\boldsymbol{h}\rangle = \frac{\lambda}{(\ell-1)\ell(\ell+1)}\left(-24 - 6\ell(p-q) + 6(p+q)\right)(r'\gamma_r^- + w'\gamma_w^-) + \\ + \frac{\lambda}{(\ell-1)\ell(\ell+1)}\left(2\ell^2(2p-q) - 6\ell q - 4(p+q) + 12(\ell+1)\right)(\gamma_r^- + \gamma_w^-),$$

where $r' = r - a + 1$ and $w' = w - a + 1$.

Let $\Delta := \ell^{-3/2}$. By using equation (27) and the facts $\ell - 1 \geq \ell/2$, $\ell + 1 < \ell$, $p, q \in [0,2]$ and triangle inequality, we bound

$$\begin{aligned}\langle \boldsymbol{A}^{-1}\boldsymbol{g}, \sum_{j=a}^{b} \boldsymbol{x}_j(f_j'(\boldsymbol{x}_j^T\boldsymbol{\beta}) - f_j'(u_j))\rangle &\leq_{(a)} \frac{40\lambda\ell\Delta}{\ell-1}\left|2p(\ell-1) - q(\ell+2) + 6\right| \\ &\quad + \frac{40\lambda\ell^2\Delta}{\ell-1}\left|3q(\ell+1) + p(1-\ell) - 4\right| \\ &{}_{(b)} \leq 640\lambda\ell\Delta(p+q) + 800\lambda\Delta,\end{aligned}$$

where the line (a) is obtained by equation (27) and making the boundary sign substitutions. Line (b) is obtained using the facts $\ell - 1 \geq \ell/2$, $\ell + 2 \leq 2\ell$ whenever $\ell \geq 2$ and $p, q \in [0,2]$ along with triangle inequality.

From Eq.(28), by using similar triangle inequality based arguments and the fact that $|\tilde{\Gamma}| \leq \ell|\Gamma|$ by Holder's inequality and Corollary 25 in as above we obtain

$$\langle \boldsymbol{A}^{-1}\boldsymbol{h}, \sum_{j=a}^{b} \boldsymbol{x}_j(f_j'(\boldsymbol{x}_j^T\boldsymbol{\beta}) - f_j'(u_j))\rangle \leq 1200\ell\Delta(\gamma_r^- + \gamma_w^-).$$

To bound T3 we observe from Lemma 21

$$\begin{aligned}T_3 &= 200 + \lambda\Bigg((s_{a-1} - s_{a-2})(M_a + C_a) - (s_b - s_{b-1})(\ell M_b + C_b) \\ &\quad - s_{a-1}M_a + s_{b-1}M_{b-1} - \sum_{t=2}^{\ell}|M_t - M_{t-1}|\Bigg) + 20\ell\Delta(\gamma_r^- + \gamma_w^-) \\ &\leq 200 + \lambda\Bigg(|(s_{a-1} - s_{a-2})(M_a + C_a)| + |(s_b - s_{b-1})(\ell M_b + C_b)| \\ &\quad + |M_a| + |M_{b-1}| + \Delta\Bigg) + 20\ell\Delta(\gamma_r^- + \gamma_w^-) \\ &\leq 200 + 80\lambda\ell\Delta(p+q) + 3\lambda\Delta + 20\ell\Delta(\gamma_r^- + \gamma_w^-),\end{aligned}$$

where in the last line we used the fact that $|(j-a+1)M_j + C_j| \leq 20\ell\Delta$ from Lemma 20.

Recall that $\Delta = \ell^{-3/2}$. Combining all the above equations / inequalities above and Eq. (29), define:

$$\begin{aligned}T(\lambda, \gamma_r^-, \gamma_w^-) &:= \lambda^2\left((2\ell^2 - 3\ell + 1)p^2 + (2\ell^2 + 3\ell + 1)q^2 - (2\ell^2 - 2)pq + 12(\ell-1)p - 12(\ell+1)q + 24\right) \\ &\quad + \left((2\ell+1)(\ell+1)(\gamma_r^- + \gamma_w^-)^2 - 6(\gamma_r^- + \gamma_w^-)(r'\gamma_r^- + w'\gamma_w^-)(\ell+1) + 6(r'\gamma_r^- + w'\gamma_w^-)^2\right) \\ &\quad + \lambda\left(-24 - 6\ell(p-q) + 6(p+q)\right)(r'\gamma_r^- + w'\gamma_w^-) \\ &\quad + \lambda\left(2\ell^2(2p-q) - 6\ell q - 4(p+q) + 12(\ell+1)\right)(\gamma_r^- + \gamma_w^-) \\ &\quad - ((\ell-1)\ell(\ell+1))\left(720\lambda\ell\ell^{-3/2}(p+q) + 803\lambda\ell^{-3/2} + 1220\ell\ell^{-3/2}(\gamma_r^- + \gamma_w^-)\right).\end{aligned} \tag{43}$$

We have,

$$T_2 + T_3 \leq -\frac{T(\lambda, \gamma_r^-, \gamma_w^-)}{(\ell-1)\ell(\ell+1)} + 200. \tag{44}$$

The expression in Eq.(43) can be compactly written as:

$$\begin{aligned} T(\lambda, \check{\gamma}_r^-, \check{\gamma}_w^-) &= 0.5 \cdot (\ell-1)\ell(\ell+1)\|\boldsymbol{g}+\boldsymbol{h}\|^2_{\boldsymbol{A}^{-1}} + \Phi(\lambda, \check{\gamma}_r^- + \check{\gamma}_w^-) \\ &:= Q(\lambda, \gamma_r^- + \gamma_w^-, r\gamma_r^- + w\gamma_w^-), \end{aligned}$$

where $\boldsymbol{g}+\boldsymbol{h}$ is as in Eq.(42) (which only depends on the boundary signs and $\lambda, \gamma_r^- + \gamma_w^-$ and $r\gamma_r^- + w\gamma_w^-$) and $\Phi(\lambda, \check{\gamma}_r^-, \check{\gamma}_w^-)$ is a linear function of its arguments namely,

$$\Phi(\lambda, \check{\gamma}_r^- + \check{\gamma}_w^-) = -(\ell-1)\ell(\ell+1)\left(20\lambda\ell\ell^{-3/2}(p+q) + 803\lambda\ell^{-3/2} + 1220\ell\ell^{-3/2}(\gamma_r^- + \gamma_w^-)\right)$$

Since we have established earlier that $\|\boldsymbol{g}\|^2_{\boldsymbol{A}^{-1}}$ is convex in $\lambda, \gamma_r^-, \gamma_w^-$ we will certainly have $T(\lambda, \gamma_r^-, \gamma_w^-)$ as a function jointly convex in its arguments.

The function B referred in Appendix B is defined to be:

$$B(\lambda, \gamma_r^-, \gamma_w^-; r, w) := -\frac{T(\lambda, \gamma_r^-, \gamma_w^-)}{(\ell-1)\ell(\ell+1)} + 200, \tag{45}$$

with r' and w' in Eq.(43) to be taken as $r' = r - i_s + 1$ and $w' = w - i_s + 1$, $\ell = i_t - i_s + 1$ and $T(\lambda, \gamma_r^-, \gamma_w^-)$ is as in Eq.(43).

So we consider the following convex optimization problem:

$$\begin{aligned} \underset{\lambda, \gamma_r^-, \gamma_w^-}{min} \quad & T(\lambda, \gamma_r^-, \gamma_w^-) \\ \text{s.t.} \quad & \lambda \geq 0 \end{aligned}$$

Note that in the program above we do *unconstrained* minimization over γ_r^- and γ_w^-. Doing so can only further decrease the objective function leading to a valid upper bound on $T_2 + T_3$.

First we will perform a partial minimization wrt the variables γ_r^- and γ_w^-. Differentiating the objective wrt γ_r^- and setting to zero yields:

$$\begin{aligned} &\left(2(2\ell^2+3\ell+1) - 12(\ell+1)r' + 12(r')^2\right)\hat{\gamma}_r^- \\ &+ \left(2(2\ell^2+3\ell+1) - 6(\ell+1)(r'+w') + 12r'w'\right)\hat{\gamma}_w^- \\ &= \lambda r'\left(24 + 6\ell(p-q) - 6(p+q)\right) - \lambda\left(2\ell^2(2p-q) - 6\ell q - 4(p+q) + 12(\ell+1)\right) \\ &+ 1220\ell^2(\ell^2-1)\ell^{-3/2}. \end{aligned}$$

Similarly differentiating the objective wrt γ_w^- and setting to zero yields:

$$\begin{aligned} &\left(2(2\ell^2+3\ell+1) - 12(\ell+1)w' + 12(w')^2\right)\hat{\gamma}_w^- \\ &+ \left(2(2\ell^2+3\ell+1) - 6(\ell+1)(r'+w') + 12r'w'\right)\hat{\gamma}_r^- \\ &= \lambda w'\left(24 + 6\ell(p-q) - 6(p+q)\right) - \lambda\left(2\ell^2(2p-q) - 6\ell q - 4(p+q) + 12(\ell+1)\right) \\ &+ 1220\ell^2(\ell^2-1)\ell^{-3/2}. \end{aligned}$$

Solving the above two equations yields:

$$\hat{\gamma}_r^- = \frac{\begin{aligned} &w'\lambda\ell^2 p + w'\lambda\ell^2 q - w'\lambda p - w'\lambda q \\ &+ 1220w'\ell^{0.5} - 1220w'\ell^{2.5} - \lambda\ell^3 q - \lambda\ell^2 p \\ &- \lambda\ell^2 q + 2\lambda\ell^2 + \lambda\ell q + \lambda p \\ &+ \lambda q - 2\lambda - 610\ell^{0.5} - 610\ell^{1.5} + 610\ell^{2.5} + 610\ell^{3.5}\end{aligned}}{r'\ell^2 - r' - w'\ell^2 + w'},$$

and

$$\hat{\gamma}_w^- = \frac{\begin{aligned} &- r'\lambda\ell^2 p - r'\lambda\ell^2 q + r'\lambda p + r'\lambda q \\ &- 1220r'\ell^{0.5} + 1220r'\ell^{2.5} + \lambda\ell^3 q + \lambda\ell^2 p \\ &+ \lambda\ell^2 q - 2\lambda\ell^2 - \lambda\ell q - \lambda p - \lambda q \\ &+ 2\lambda + 610\ell^{0.5} + 610\ell^{1.5} - 610\ell^{2.5} - 610\ell^{3.5}\end{aligned}}{r'\ell^2 - r' - w'\ell^2 + w'}.$$

Substituting the above two expression we get:

$$T(\lambda, \hat{\gamma}_r^-, \hat{\gamma}_w^-) = \frac{\begin{aligned} &- 797\lambda\ell^{2.0} - 1780\lambda\ell^{3.0}p - 1780\lambda\ell^{3.0}q \\ &+ 2391\lambda\ell^{4.0} + 5340\lambda\ell^{5.0}p + 5340\lambda\ell^{5.0}q - 2391\lambda\ell^{6.0} \\ &- 5340\lambda\ell^{7.0}p - 5340\lambda\ell^{7.0}q + 797\lambda\ell^{8.0} + 1780\lambda\ell^{9.0}p + 1780\lambda\ell^{9.0}q \\ &+ 744200\ell^{3.5} - 2232600\ell^{5.5} + 2232600\ell^{7.5} - 744200\ell^{9.5}\end{aligned}}{\ell^{2.5} - 2\ell^{4.5} + \ell^{6.5}} \tag{47}$$

Looking at Eq.(47) we notice that it is a linear fucntion of λ which defined the function $\mathcal{L}(\lambda)$ mentioned in Appendix B:

$$\mathcal{L}(\lambda) = \frac{\begin{aligned} &- 797\lambda\ell^{2.0} - 1780\lambda\ell^{3.0}p - 1780\lambda\ell^{3.0}q \\ &+ 2391\lambda\ell^{4.0} + 5340\lambda\ell^{5.0}p + 5340\lambda\ell^{5.0}q - 2391\lambda\ell^{6.0} \\ &- 5340\lambda\ell^{7.0}p - 5340\lambda\ell^{7.0}q + 797\lambda\ell^{8.0} + 1780\lambda\ell^{9.0}p + 1780\lambda\ell^{9.0}q \\ &+ 744200\ell^{3.5} - 2232600\ell^{5.5} + 2232600\ell^{7.5} - 744200\ell^{9.5}\end{aligned}}{\ell^{2.5} - 2\ell^{4.5} + \ell^{6.5}} \tag{48}$$

We observe that the leading term (i.e terms whose magnitude is biggest) in the denominator is a positive quantity namely $\ell^{6.5}$. The leading term in the numerator that contains λ grows as $1780\lambda\ell^9(p+q) + 797\lambda\ell^8$. So the unconstrained minimum of this linear function is attained at $\lambda = -\infty$.

Hence the constrained minimum (with constraint $\lambda \geq 0$) of the optimization problem 46 is attained at $\lambda = 0$. We calculate the optimal objective to the constrained problem via Eq.(47) as

$$T(0, \hat{\gamma}_r^-, \hat{\gamma}_w^-) = \frac{744200\left(\ell^{1.5} - 3\ell^{3.5} + 3\ell^{5.5} - \ell^{7.5}\right)}{\ell^{0.5} - 2\ell^{2.5} + \ell^{4.5}},$$

where we consider bins with length $\ell \geq 14$.

Since $\ell^4 \geq 2\ell^2$ for all $\ell \geq 2$, we continue from the previous display to obtain:

$$\begin{aligned} T(0, \hat{\gamma}_r^-, \hat{\gamma}_w^-) &\geq -744200 \cdot (1+3+3+1)\frac{\ell^{7.5}}{\ell^{\ell.5}} \\ &= -5953600\ell^3, \end{aligned}$$

Hence we have

$$\begin{aligned} \frac{T(0, \hat{\gamma}_r^-, \hat{\gamma}_w^-)}{(\ell-1)\ell(\ell+1)} &\geq \frac{-5953600\ell^3}{(\ell-1)\ell(\ell+1)} \\ &\geq -11907200, \end{aligned}$$

where in the last line we used the fact that $\ell - 1 \geq \ell/2$ is satisfied for all $\ell \geq 14$ and $\ell + 1 > \ell$.

Hence continuing from Eq.(44) we conclude that

$$\begin{aligned} T_2 + T_3 &\leq (11907200 + 200) \\ &= 11907400. \end{aligned}$$

The term T_1 can be bound as

$$\begin{aligned} T_1 &\leq 256 + \frac{1}{2\sigma} \log(1 + \sigma n/2) + \frac{4}{\sigma} \log n \\ &= \tilde{O}(1), \end{aligned}$$

by Lemma 27.

Now suppose that the offline optimal within bin $[a, b]$ touches boundary -1 more than two times. In this case we propose a reduction to the previous type of analysis where only γ_r^- and γ_w^- are potentially non-zero.

The reduction is facilitated by two observations:

1. While performing the minimization of function $T(\lambda, \gamma_r^-, \gamma_w^-)$ in Eq.(43) via the optimization problem 46 we neither used the fact that r and w are integers nor constrained any bounds on them as well
2. The partially minimized objective in Eq.(47) fortunately doesn't depend on neither r nor w.

Now let's consider the case where arbitrary number of γ_j^-, $j \in [a, b]$ can be non-zero. We can then write,

$$\begin{aligned} \Gamma &= \sum_{j=a}^{b} \gamma_j^- \\ &= \check{\gamma}_r^- + \check{\gamma}_w^-, \end{aligned}$$

where $\check{\gamma}_r^- := \gamma_1^-$ and $\check{\gamma}_w^- = \Gamma - \check{\gamma}_1^-$.

Define $r' := 1$ and $w' := \frac{\sum_{j=a}^{b} j' \gamma_j^- - \check{\gamma}_r^-}{\check{\gamma}_w^-} = \frac{\tilde{\Gamma} - \check{\gamma}_r^-}{\check{\gamma}_w^-}$ where we assume that $\check{\gamma}_w^- > 0$ (otherwise, we fall back to the earlier analysis).

With these re-definitions we note that

$$T_2 + T_3 \leq -\frac{T(\lambda, \check{\gamma}_r^-, \check{\gamma}_w^-)}{(\ell - 1)\ell(\ell + 1)},$$

still holds. Further, $T(\lambda, \check{\gamma}_r^-, \check{\gamma}_w^-)$ is jointly convex in its arguments. This can be seen as follows: Note that $T(\lambda, \check{\gamma}_r^-, \check{\gamma}_w^-)$ assumes the form

$$T(\lambda, \check{\gamma}_r^-, \check{\gamma}_w^-) = 0.5 \cdot (\ell - 1)\ell(\ell + 1) \| \boldsymbol{g} + \boldsymbol{h} \|_{\boldsymbol{A}^{-1}}^2 + \Phi(\lambda, \check{\gamma}_r^- + \check{\gamma}_w^-),$$

where $\Phi(\lambda, \check{\gamma}_r^- + \check{\gamma}_w^-)$ is an affine function of its arguments and

$$\begin{aligned} \boldsymbol{h} &= [\Gamma, \tilde{\Gamma}]^T \\ &= [\check{\gamma}_r^- + \gamma_w^-, r' \check{\gamma}_r^- + w' \check{\gamma}_w^-]^T, \end{aligned}$$

where the last line follows due to our re-parametrizations. By following essentially same arguments as earlier for proving convexity of $T(\lambda, \gamma_r^-, \gamma_w^-)$ we conclude that $T(\lambda, \check{\gamma}_r^-, \check{\gamma}_w^-)$ is also jointly convex in its arguments.

This completes our reduction to the case of two-boundary touches and rest of analysis proceeds by minimizing $T(\lambda, \check{\gamma}_r^-, \check{\gamma}_w^-)$ as earlier.

We now consider the case where $s_{a-1} = s_b = 1$. We can split the original bin $[a, b]$ into two sub-bins $[a_1, b_1]$ and $[a_2, b_2]$ with $a_2 = b_1 + 1$ such that (i) $s_{b_1} = -1$ with $u_{b_1+1} - u_{b_1} > u_{a_2+1} - u_{a_2}$ and (ii) the slopes are non-decreasing within $[a_2, b_2]$. This can be achieved by picking b_1 as the last point within $[a, b]$ where $u_{b_1+1} - u_{b_1} > u_{b_1+2} - u_{b_1+1}$.

In the bin $[a_1, b_1]$ we apply the previous analysis to bound regret by $\tilde{O}(1)$. For the bin $[a_2, b_2]$ we resort to Lemma 28 to bound regret by $\tilde{O}(1)$.

The analysis for the case of boundary signs assignments $s_{a-1} = -1$ and $s_b = 1$ as well as $s_{a-1} = -1$ and $s_b = -1$ can be done similarly.

Adding the regret bounds across all newly formed bins due to potential splitting yields the lemma.

□

Next, we provide the full regret guarantee in a uni-variate setting.

Theorem 30. *Let p_t be the predictions of FLH-SIONS algorithm with parameters $\epsilon = 2$, $C = 20$ and exp-concavity factor σ. Under Assumptions A1-A4, we have that,*

$$\sum_{t=1}^{n} f_t(p_t) - f_t(w_t) = \tilde{O}(n^{1/5} C_n^{2/5} \vee 1),$$

for any comparator sequence $w_{1:n} \in \mathcal{TV}^{(1)}(C_n)$. Here $\tilde{O}$ hides poly-logarithmic factors of n and $a \vee b = \max\{a, b\}$.

Proof. The proof is complete by adding the $\tilde{O}(1)$ dynamic regret bound from Lemmas 28 and 29 across $O(n^{1/5} C_n^{2/5} \vee 1)$ bins in the partition $\mathcal{P}$.

□

The proof of Lemma 11 stated in Appendix B is similar to the arguments used to derive Eq.(20). We record it for the sake of completeness.

Lemma 11. *We have that $T_2 \leq -\frac{1}{2} \|\nabla F(\boldsymbol{\beta})\|^2_{\boldsymbol{A}^{-1}}$.*

Proof. We follow the same notations used in defining Lemma 11.

Let's begin by calculating $\boldsymbol{v} := \boldsymbol{A}^{-1} \sum_{j=a}^{b} f'_j(\boldsymbol{x}_j^T \boldsymbol{\beta}) \boldsymbol{x}_j$.

We have,

$$\begin{aligned}
|\boldsymbol{v}[1]| &= \left| \frac{2}{(\ell-1)\ell} \sum_{j=1}^{\ell} (2\ell + 1 - 3j) f'_{(j+a-1)}(\boldsymbol{x}_{j+a-1}^T \boldsymbol{\beta}) \right| \\
&\overset{(a)}{\leq} \frac{2}{(\ell-1)\ell} \cdot 2\ell(\ell-1) \\
&= 4, \qquad (49)
\end{aligned}$$

where line (a) is obtained via Lipschitzness and Holder's inequality $\boldsymbol{x}^T \boldsymbol{y} \leq \|\boldsymbol{x}\|_1 \|\boldsymbol{y}\|_\infty$ and the fact that $|2\ell + 1 - 3j| \leq 2(\ell - 1)$ for all $j \in [1, \ell]$.

Similarly

$$\begin{aligned}
|\boldsymbol{v}[2]| &= \left| \frac{2}{(\ell-1)\ell(\ell+1)} \sum_{j=1}^{\ell} (-3(\ell+1) + 6j) f'_{(j+a-1)}(\boldsymbol{x}_{j+a-1}^T \boldsymbol{\beta}) \right| \\
&\leq \frac{2}{(\ell-1)\ell(\ell+1)} \cdot 3\ell(\ell-1) \\
&= \frac{6}{(\ell+1)}, \qquad (50)
\end{aligned}$$

where we used $|-3(\ell+1) + 6j| \leq 3(\ell - 1)$ for all $j \in [1, \ell]$.

Combining Eq.(49) and (50) we conclude that

$$|\boldsymbol{v}^T \boldsymbol{x}_j| = 4 + (j - a + 1)\frac{6}{(\ell + 1)}$$
$$\leq 10, \tag{51}$$

where the last line follows due to the fact $(j - a + 1) \leq \ell$.

Hence by Triangle inequality we have

$$|\boldsymbol{\alpha}^T \boldsymbol{x}_j| \leq |\boldsymbol{\beta}^T \boldsymbol{x}_j| + 10. \tag{52}$$

Now we bound $|\boldsymbol{\beta}^T \boldsymbol{x}_j|$ using similar arguments. We have $\boldsymbol{v}' := \boldsymbol{A}^{-1} \sum_{j=a}^{b} u_j \boldsymbol{x}_j$. Now noting that $|u_j| \leq 1$ by Assumption A1 and using similar arguments used to obtain Eq.(51) we conclude that

$$|\boldsymbol{\beta}^T \boldsymbol{x}_j| \leq 10. \tag{53}$$

So continuing from Eq.(52) we have $|\boldsymbol{\alpha}^T \boldsymbol{x}_j| \leq 20$.

For some $\boldsymbol{z} = t\boldsymbol{\alpha} + (1 - t)\boldsymbol{\beta}$, $t \in [0, 1]$ we have by Taylor's theorem that

$$\begin{aligned} F(\boldsymbol{\alpha}) - F(\boldsymbol{\beta}) &= -\langle \nabla F(\boldsymbol{\beta}), \boldsymbol{A}^{-1} \nabla F(\boldsymbol{\beta}) \rangle + \frac{1}{2} \|\boldsymbol{A}^{-1} \nabla F(\boldsymbol{\beta})\|^2_{\nabla^2 F(\boldsymbol{z})} \\ &\leq -\langle \nabla F(\boldsymbol{\beta}), \boldsymbol{A}^{-1} \nabla F(\boldsymbol{\beta}) \rangle + \frac{1}{2} \|\boldsymbol{A}^{-1} \nabla F(\boldsymbol{\beta})\|^2_{\boldsymbol{A}} \\ &= -\frac{1}{2} \|\nabla F(\boldsymbol{\beta})\|^2_{\boldsymbol{A}^{-1}}, \end{aligned}$$

where in the first inequality we used that fact that $\nabla^2 F(\boldsymbol{z}) \preccurlyeq \boldsymbol{A}$ due to the fact that the functions f_j are 1 gradient Lipschitz in $[-20, 20]^d$ via Assumption A3. □

C.2 Multi-dimensional setting

`splitMonotonic`: Inputs- (1) offline optimal sequence (2) A bin $[i_s, i_t]$ (3) A coordinate $k \in [d]$

1. Compute $\boldsymbol{z}_j[k] = \boldsymbol{u}_{j+1}[k] - \boldsymbol{u}_j[k]$
2. If $\boldsymbol{z}[k]$ is constant in $[i_s, i_t]$ return $\{i_s, i_t\}$.
3. If $\boldsymbol{z}[k]$ is non-decreasing (non-increasing) across $[i_s, i_t]$: `//ensure equal boundary signs (see caption) for bin` $[b+1, c]$ `below.`
   (a) Split $[i_s, i_t]$ into at-most three bins $[i_s, b], [b + 1, c], [c + 1, i_t]$ such that $\boldsymbol{z}_j[k]$ remains constant in the first and last bins. Further $\boldsymbol{z}_{b+1}[k] > (<) \boldsymbol{z}_b[k]$ and $\boldsymbol{z}_{c+1}[k] > (<) \boldsymbol{z}_c[k]$.
   (b) Return $\{i_s, b, b + 1, c, c + 1, i_t\}$

Figure 6: `splitMonotonic` *procedure. If line 3 is replaced by "If* $\boldsymbol{z}[k]$ *is non-increasing ...", then we propagate that change by replacing the symbols* $>$ */*$<$ *in the lines below 3 by the bracketed statements next to it. For a bin* $[a, b]$*, we refer to* s_{a-1} *and* s_b *as the boundary signs.*

`generateBins`: Input- (1) offline optimal sequence

1. Form consecutive bins $[i_s, i_t]$ such that: `// coarse partition based on TV1 distance`
   (a) $\|D^2 \boldsymbol{u}_{i_s:i_t}\|_1 \leq 1/\ell_{i_s \to i_t}^{3/2}$
   (b) $\|D^2 \boldsymbol{u}_{i_s:i_t+1}\|_1 > 1/\ell_{i_s \to i_t+1}^{3/2}$,
   where $\ell_{a \to b} := b - a + 1$.
2. Let the partition of the time horizon be represented as $\mathcal{P}' := \{[1_s, 1_t], \ldots, [i_s, i_t], \ldots [M_s, M_t]\}$ where $M = |\mathcal{P}'|$.
3. Initialize $\mathcal{R} \leftarrow \Phi$.
4. For each bin $[i_s, i_t] \in \mathcal{P}'$: `// ensuring` $\boldsymbol{\gamma}_j^+[k]\gamma_j^-[k] = 0$ `for all` $k \in [d]$
   (a) $\mathcal{R} = \mathcal{R} \cup \{i_s, i_t\}$.
   (b) For each coordinate $k \in [d]$:
      i. If $\boldsymbol{u}_{i_s}[k] = 1(-1)$ and there exists a point $p \in [i_s, i_t]$ such that $\boldsymbol{u}_p = -1(1)$ then $\mathcal{R} \leftarrow \mathcal{R} \cup \{p-1, p\}$
      ii. If $\boldsymbol{u}_{i_t}[k] = 1(-1)$ and there exists a point $p \in [i_s, i_t]$ such that $\boldsymbol{u}_p = -1(1)$ then $\mathcal{R} \leftarrow \mathcal{R} \cup \{p-1, p\}$
5. Remove duplicates from $\mathcal{R}$ and form a partition $\mathcal{P}$ by splitting at each point in $\mathcal{R}$
6. Return $\mathcal{P}$

Figure 7: `generateBins` *procedure. If line 7(d) is replaced by "If $\boldsymbol{z}_p[k] < \boldsymbol{z}_{p-1}[k]$", then we propagate that change by replacing the symbols $> / <$ in the lines below 7(d) by the bracketed statements next to it. For a bin $[a, b]$, we refer to s_{a-1} and s_b as the boundary signs.*

Lemma 31. *Consider the following convex optimization problem.*

$$\begin{aligned}
\min_{\tilde{\boldsymbol{u}}_1, \ldots, \tilde{\boldsymbol{u}}_n, \tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_{n-1}} \quad & \sum_{t=1}^{n} f_t(\tilde{\boldsymbol{u}}_t) \\
\text{s.t.} \quad & \tilde{\boldsymbol{z}}_t = \tilde{\boldsymbol{u}}_{t+2} - 2\tilde{\boldsymbol{u}}_{t+1} + \tilde{\boldsymbol{u}}_t \; \forall t \in [n-2], \\
& \sum_{t=1}^{n-2} \|\tilde{\boldsymbol{z}}_t\|_1 \leq C_n/n, && (54a) \\
& -1 \leq \tilde{\boldsymbol{u}}_t[k] \; \forall t \in [n], \; \forall k \in [d] && (54b) \\
& \tilde{\boldsymbol{u}}_t[k] \leq 1 \; \forall t \in [n], \; \forall k \in [d] && (54c)
\end{aligned}$$

Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n, \boldsymbol{z}_1, \ldots, \boldsymbol{z}_{n-2}$ be the optimal primal variables and let $\lambda \geq 0$ be the optimal dual variable corresponding to the constraint (73a). Further, let $\boldsymbol{\gamma}_t^- \geq 0, \boldsymbol{\gamma}_t^+ \geq 0$ (coordinate-wise) be the optimal dual variables that correspond to constraints (73b) and (54c) respectively for all $t \in [n]$. Note that $\boldsymbol{\gamma}_t^-, \boldsymbol{\gamma}_t^+ \in \mathbb{R}^d$. By the KKT conditions, we have

- ***stationarity:*** $\nabla f_t(\boldsymbol{u}_t) = \lambda\left((\boldsymbol{s}_{t-1} - \boldsymbol{s}_t) - (\boldsymbol{s}_{t-2} - \boldsymbol{s}_{t-1})\right) + \boldsymbol{\gamma}_t^- - \boldsymbol{\gamma}_t^+$*, where* $\boldsymbol{s}_t[k] \in \partial|\boldsymbol{z}_t[k]|$ *(a subgradient) for* $k \in [d]$*. Specifically,* $\boldsymbol{s}_t[k] = \text{sign}((\boldsymbol{u}_{t+2}[k] - \boldsymbol{u}_{t+1}[k]) - (\boldsymbol{u}_{t+1}[k] - \boldsymbol{u}_t[k]))$ *if* $|(\boldsymbol{u}_{t+2}[k] - \boldsymbol{u}_{t+1}[k]) - (\boldsymbol{u}_{t+1}[k] - \boldsymbol{u}_t[k])| > 0$ *and* $\boldsymbol{s}_t[k]$ *is some value in* $[-1, 1]$ *otherwise. For convenience of notations, we also define* $\boldsymbol{s}_{-1} = \boldsymbol{s}_0 = \boldsymbol{0}$.
- ***complementary slackness:*** *(a)* $\lambda\left(\sum_{t=1}^{n-2} \|\boldsymbol{z}_t\|_1 - C_n/n\right) = 0$*; (b)* $\boldsymbol{\gamma}_t^-[k](\boldsymbol{u}_t[k] + 1) = 0$ *and* $\boldsymbol{\gamma}_t^+[k](\boldsymbol{u}_t[k] - 1) = 0$ *for all* $t \in [n]$.

The proof of above Lemma is similar to that of Lemma 17 and hence omitted.

Lemma 32. *(Luo et al., 2016) Consider an online learning setting where at each round t, we are given a feature vector $\boldsymbol{x}_t \in \mathbb{R}^2$. Define $\tilde{f}_t(\boldsymbol{v}) = f_t(\boldsymbol{x}_t^T \boldsymbol{v}[1:2], \dots, \boldsymbol{x}_t^T \boldsymbol{v}[2d-1:2d])$ for some vector $\boldsymbol{v} \in \mathbb{R}^{2d}$. Let the function $f(\boldsymbol{r})$ be σ exp-concave and G Lipschitz for $\boldsymbol{r} \in \mathbb{R}^d$ with $\|\boldsymbol{r}\|_\infty \leq C$. Define $\mathcal{K}_t := \{\boldsymbol{w} \in \mathbb{R}^{2d} : |\boldsymbol{x}_t^T \boldsymbol{w}[2k-1:2k]| \leq C \, \forall k \in [d]\}$. Let $\mathcal{K} := \cap_{t=1}^T \mathcal{K}_t$ and $\boldsymbol{g}_t := \nabla \tilde{f}_t(\boldsymbol{p}_t)$. Consider a variant of the algorithm proposed by (Luo et al., 2016) where the algorithm makes a prediction $\hat{\boldsymbol{p}}_{t+1} \in \mathbb{R}^d$ at round $t+1$ as:*

$$\boldsymbol{w}_{t+1} = \boldsymbol{p}_t - \boldsymbol{A}_t^{-1} \boldsymbol{g}_t$$
$$\boldsymbol{p}_{t+1} = \underset{\boldsymbol{w} \in \mathcal{K}_{t+1}}{\operatorname{argmin}} \|\boldsymbol{w} - \boldsymbol{w}_{t+1}\|_{\boldsymbol{A}_t}$$
$$\hat{\boldsymbol{p}}_{t+1} = \left[\boldsymbol{x}_{t+1}^T \boldsymbol{p}_{t+1}[1:2], \dots, \boldsymbol{x}_{t+1}^T \boldsymbol{p}_{t+1}[2d-1:2d]\right]^T$$

where $\boldsymbol{A}_t = \epsilon \boldsymbol{I} + \sum_{s=1}^t \sigma \boldsymbol{g}_s \boldsymbol{g}_s^T$ with $\boldsymbol{I}$ is the identity matrix and ϵ is an input parameter.

Then for any $\boldsymbol{w} \in \mathcal{K}$ we have the regret controlled as

$$\begin{aligned}\sum_{t=1}^T f_t(\hat{\boldsymbol{p}}_t) - \tilde{f}_t(\boldsymbol{w}) &= \sum_{t=1}^T \tilde{f}_t(\boldsymbol{p}_t) - \tilde{f}_t(\boldsymbol{w}) \\ &\leq \frac{\epsilon \|\boldsymbol{w}\|_2^2}{2} + \frac{2d}{\sigma} \log\left(1 + \frac{\sigma T G^2}{d\epsilon}\right).\end{aligned}$$

We will call this algorithm as SIONS (Scale Invariant Online Newton Step).

Proof. First we show that exp-concavity is invariant to affine transforms. Since f_t is σ exp-concave, we have

$$\begin{aligned}\tilde{f}_t(\boldsymbol{w}) \geq \tilde{f}_t(\boldsymbol{v}) &+ \Big\langle \nabla f_t(\boldsymbol{x}_t^T \boldsymbol{v}[1:2], \dots, \boldsymbol{x}_t^T \boldsymbol{v}[2d-1:2d]), \\ &[\boldsymbol{x}_t^T(\boldsymbol{w}[1:2] - \boldsymbol{v}[1:2]), \dots, \boldsymbol{x}_t^T(\boldsymbol{w}[2d-1:2d] - \boldsymbol{v}[2d-1:2d])]^T] \Big\rangle \\ &+ \frac{\sigma}{2}\Bigg(\Big\langle \nabla f_t(\boldsymbol{x}_t^T \boldsymbol{v}[1:2], \dots, \boldsymbol{x}_t^T \boldsymbol{v}[2d-1:2d]), \\ &[\boldsymbol{x}_t^T(\boldsymbol{w}[1:2] - \boldsymbol{v}[1:2]), \dots, \boldsymbol{x}_t^T(\boldsymbol{w}[2d-1:2d] - \boldsymbol{v}[2d-1:2d])]^T] \Big\rangle\Bigg)^2.\end{aligned}$$

For the sake of brevity let's denote $f_t^{(k)} := \nabla f_t(\boldsymbol{x}_t^T \boldsymbol{v}[1:2], \dots, \boldsymbol{x}_t^T \boldsymbol{v}[2d-1:2d])[k]$ for $k \in [d]$. Then we have

$$\nabla \tilde{f}_t(\boldsymbol{v}) = \left[f_t^{(1)} \boldsymbol{x}_t^T, \dots, f_t^{(d)} \boldsymbol{x}_t^T\right]^T.$$

Let

$$\begin{aligned}A = \Big\langle &\nabla f_t(\boldsymbol{x}_t^T \boldsymbol{v}[1:2], \dots, \boldsymbol{x}_t^T \boldsymbol{v}[2d-1:2d]), \\ &[\boldsymbol{x}_t^T(\boldsymbol{w}[1:2] - \boldsymbol{v}[1:2]), \dots, \boldsymbol{x}_t^T(\boldsymbol{w}[2d-1:2d] - \boldsymbol{v}[2d-1:2d])]^T] \Big\rangle.\end{aligned}$$

With this, we observe that,

$$A = (\boldsymbol{w} - \boldsymbol{v})^T \nabla \tilde{f}_t(\boldsymbol{v}).$$

Thus, we obtain the affine invariance of exp-concavity as:

$$\tilde{f}_t(\boldsymbol{w}) \geq \tilde{f}_t(\boldsymbol{v}) + (\boldsymbol{w} - \boldsymbol{v})^T \nabla \tilde{f}_t(\boldsymbol{v}) + \frac{\sigma}{2}\left((\boldsymbol{w} - \boldsymbol{v})^T \nabla \tilde{f}_t(\boldsymbol{v})\right)^2. \tag{55}$$

Note that the set $\mathcal{K}_t$ is convex. This can be seen as follows: if $\boldsymbol{v}, \boldsymbol{w} \in \mathcal{K}_t$, then we have $|\boldsymbol{x}_t^T \boldsymbol{v}[2k-1:2k]| \leq C$ and $|\boldsymbol{x}_t^T \boldsymbol{w}[2k-1:2k]| \leq C$ for all $k \in [d]$. Now for any $t \in [0,1]$ let $\boldsymbol{z} = t\boldsymbol{v} + (1-t)\boldsymbol{w}$. Then we have for any $k \in [d]$ that

$$\begin{aligned} |\boldsymbol{x}_t^T \boldsymbol{z}[2k-1:2k]| &\leq t|\boldsymbol{x}_t^T \boldsymbol{v}[2k-1:2k]| + (1-t)|\boldsymbol{x}_t^T \mathbf{2}[2k-1:2k]| \\ &\leq C, \end{aligned}$$

where the first inequality is via triangle inequality. Thus $\boldsymbol{z} \in \mathcal{K}_t$ so the set $\mathcal{K}_t$ is convex.

So by the properties of projection to convex sets (see for example, Lemma 16 in (Hazan et al., 2007)) and the definition of the algorithm, we have that

$$\begin{aligned} \|\boldsymbol{p}_{t+1} - \boldsymbol{w}\|_{\boldsymbol{A}_t}^2 &\leq \|\boldsymbol{w}_{t+1} - \boldsymbol{w}\|_{\boldsymbol{A}_t}^2 \\ &= \|\boldsymbol{p}_t - \boldsymbol{w}\|_{\boldsymbol{A}_t}^2 + \boldsymbol{g}_t^T \boldsymbol{A}_t^{-1} \boldsymbol{g}_t - 2\boldsymbol{g}_t^T(\boldsymbol{p}_t - \boldsymbol{w}). \end{aligned}$$

Let $R_T(\boldsymbol{w}) := \sum_{t=1}^T \tilde{f}_t(\boldsymbol{p}_t) - \tilde{f}_t(\boldsymbol{w})$. Since each f_t is exp-concave, we have by Eq.(55) and the previous inequality that

$$\begin{aligned} 2R_T(\boldsymbol{w}) &\leq \sum_{t=1}^T 2\boldsymbol{g}_t^T(\boldsymbol{p}_t - \boldsymbol{w}) - \sigma(\boldsymbol{g}_t^T(\boldsymbol{p}_t - \boldsymbol{w}))^2 \\ &\leq \sum_{t=1}^T \boldsymbol{g}_t^T \boldsymbol{A}_t^{-1} \boldsymbol{g}_t + \|\boldsymbol{p}_t - \boldsymbol{w}\|_{\boldsymbol{A}_t}^2 - \|\boldsymbol{p}_{t+1} - \boldsymbol{w}\|_{\boldsymbol{A}_t}^2 - \sigma(\boldsymbol{g}_t^T(\boldsymbol{p}_t - \boldsymbol{w}))^2 \\ &\leq \|\boldsymbol{w}\|_{\boldsymbol{A}_0}^2 + \sum_{t=1}^T \boldsymbol{g}_t^T \boldsymbol{A}_t^{-1} \boldsymbol{g}_t + (\boldsymbol{p}_t - \boldsymbol{w})^T(\boldsymbol{A}_t - \boldsymbol{A}_{t-1} - \sigma \boldsymbol{g}_t \boldsymbol{g}_t^T)(\boldsymbol{p}_t - \boldsymbol{w}) \\ &= \|\boldsymbol{w}\|_{\boldsymbol{A}_0}^2 + \sum_{t=1}^T \boldsymbol{g}_t^T \boldsymbol{A}_t^{-1} \boldsymbol{g}_t, \end{aligned}$$

where the last line is by the definition of $\boldsymbol{A}_t$.

By using the arguments of Lemma 12 of (Hazan et al., 2007) we have

$$\sum_{t=1}^T \boldsymbol{g}_t^T \boldsymbol{A}_t^{-1} \boldsymbol{g}_t \leq \frac{2d}{\sigma} \log\left(1 + \frac{\sigma T G^2}{d\epsilon}\right).$$

Thus overall we have,

$$R_T(\boldsymbol{w}) \leq \frac{\epsilon \|\boldsymbol{w}\|_2^2}{2} + \frac{2d}{\sigma} \log\left(1 + \frac{\sigma T G^2}{d\epsilon}\right)$$

□

Corollary 33. *(Hazan and Seshadhri, 2007) Consider the FLH algorithm from (Hazan and Seshadhri, 2007) with SIONS from Lemma 32 as the base experts with parameter $\epsilon = 2$ as described in Fig.3. Consider an arbitrary interval $[a,b] \subseteq [n]$. Then the regret of FLH-SIONS within this interval is controlled as:*

$$\sum_{j=a}^b f_j(\boldsymbol{y}_j) - \tilde{f}_j(\boldsymbol{w}) \leq \|\boldsymbol{w}\|_2^2 + \frac{2d}{\sigma} \log\left(1 + \frac{\sigma n^3 G^2}{d\epsilon}\right) + \frac{4 \log n}{\sigma},$$

where $\boldsymbol{w} \in \cap_{j=a}^b \mathcal{K}_j$ and $\tilde{f}$ is as defined in Lemma 32.

Proof. Since the loss functions f_j are σ exp-concave, by Lemma 3.3 in (Hazan and Seshadhri, 2007) we have that

$$\sum_{j=a}^b f_j(\boldsymbol{y}_j) \leq \frac{4 \log n}{\sigma} + \sum_{j=a}^b f_j(E_a(j)).$$

Subtracting $\tilde{f}_j(\boldsymbol{w})$ from both sides and using Lemma 32 now yields the result. □

Corollary 34. *The number of bins $M := |\mathcal{P}|$ formed via a call to* `generateBins`$(\boldsymbol{u}_{1:n})$ *is at-most $O(n^{1/5}C_n^{2/5} \vee 1)$.*

Proof. The proof is similar to that of Lemma 10. ☐

Lemma 35. *Let $[i_s, i_t] \in \mathcal{P}$ where $\mathcal{P}$ is the partition produced via the* `generateBins` *procedure. We have that the dynamic rgeret of FLH-SIONS within this bin controlled as*

$$\sum_{j=i_s}^{i_t} f_j(\hat{\boldsymbol{p}}_j) - f_j(\boldsymbol{u}_j) = \tilde{O}(d^2),$$

where $\hat{\boldsymbol{p}}_j \in \mathbb{R}^d$ are the predictions of the algorithm.

Proof. Consider a bin $[i_s, i_t]$. Let $\mathcal{Q} = \texttt{refineSplit}([i_s, i_t])$. Define $\tilde{f}_j(\boldsymbol{v}) := \tilde{f}_j(\boldsymbol{y}_j^T \boldsymbol{v})$ for $v \in \mathbb{R}^{2d}$.

Next, we proceed to construct the details of a regret decomposition within a bin $[i_s, i_t]$:

$$\sum_{j=i_s}^{i_t} f_j(\hat{\boldsymbol{p}}_j) - f_j(\boldsymbol{u}_j) = \underbrace{\sum_{j=i_s}^{i_t} f_j(\hat{\boldsymbol{p}}_j) - f_j(\boldsymbol{X}_j \boldsymbol{\alpha}_j)}_{T_1} + \underbrace{\sum_{j=i_s}^{i_t} f_j(\boldsymbol{X}_j \boldsymbol{\alpha}_j) - f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)}_{T_2} + \underbrace{\sum_{j=i_s}^{i_t} f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j) - f_j(\boldsymbol{u}_j)}_{T_3}, \quad (56)$$

where we will construct appropriate $\boldsymbol{y}_j, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j \in \mathbb{R}^{2d}$ and $\boldsymbol{X}_j \in \mathbb{R}^{d \times 2d}$ in what follows.

`AssignCo-variatesAndSlopes1`: Inputs- (1) offline optimal sequence (2) A bin $[a, b]$ (3) A coordinate $k \in [d]$

1. Let $\boldsymbol{\beta}_k$ be the least square fit coefficient computed with labels being $\boldsymbol{u}_a[k], \ldots, \boldsymbol{u}_b[k]$ and co-variates $\boldsymbol{x}_j := [1, j - a + 1]^T$ so that the fitted value at time j is given by $\hat{\boldsymbol{u}}_j[k] = \boldsymbol{\beta}^T \boldsymbol{x}_j$.
2. Set $\boldsymbol{\beta}_j[2k-1 : 2k] \leftarrow \beta_k$ for all $j \in [a, b]$.
3. Set $\boldsymbol{\alpha}_k \leftarrow \beta_k$
4. Set $\boldsymbol{\alpha}_j[2k-1 : 2k] \leftarrow \boldsymbol{\alpha}_k$ for all $j \in [a, b]$.
5. Set $\boldsymbol{y}_j[2k-1 : 2k] = \boldsymbol{x}_j$ for all $j \in [a, b]$.

Figure 8: `AssignCo-variatesAndSlopes1` *used to set the parameters in the regret decomposition of Eq.(56) whenever the offline optimal is constant across the specified coordinate k within the interval $[a, b]$. We use a 1-based indexing. i.e $\boldsymbol{v}[1]$ refers the first element of a vector $\boldsymbol{v}$.*

(A1): Consider a coordinate $k \in [d]$ such that $\boldsymbol{u}[k]$ is not monotonic in $[i_s, i_t]$ and do not touch boundary 1. Let $[i_s, i_t] = [i_s, a-1] \cup [a, b] \cup [b+1, c] \cup [c+1, i_t]$ such that $\boldsymbol{u}[k]$ is constant in bins $[i_s, a-1]$ and $[c+1, i_t]$. Further we consider the case where $s_{a-1} = 1$ and $s_b = -1$ with $\boldsymbol{u}[k]$ non-decreasing within $[b+1, c]$. (Note that this can be guaranteed by picking b as the last point with $\boldsymbol{u}_{b+1}[k] - \boldsymbol{u}_b[k] > \boldsymbol{u}_{b+2}[k] - \boldsymbol{u}_{b+1}[k]$.) The alternate case where $s_{a-1} = -1$ and $s_b = 1$ with $\boldsymbol{u}[k]$ non-increasing within $[b+1, c]$ can be handled similarly. All the arguments we explain for the case of offline optimal touching the boundary -1 can be mirrored to handle the case where the offline optimal touches the boundary 1. (The offline optimal can't touch both boundaries simultaneously along a coordinate, see Lemma 22)

We will use 1-based indexing. (i.e $\boldsymbol{v}[1]$ denotes the first element of a vector). For each $k \in [d]$:

- Call `AssignCo-variatesAndSlopes1`$(\boldsymbol{u}_{1:n}, [i_s, a-1], k)$.
- Call `AssignCo-variatesAndSlopes2`$(\boldsymbol{u}_{1:n}, [a, b], k)$.
- Let $[b+1, t_1 - 1], [t_1, t_2], [t_2 + 1, c]$ be the bins returned by a call to `splitMonotonic`$(\boldsymbol{u}_{1:n}, [b+1, c], k)$.
- Call `AssignCo-variatesAndSlopes1`$(\boldsymbol{u}_{1:n}, [b+1, t_1 - 1], k)$.
- Call `AssignCo-variatesAndSlopes2`$(\boldsymbol{u}_{1:n}, [t_1, t_2], k)$.

`AssignCo-variatesAndSlopes2`: Inputs- (1) offline optimal sequence (2) A bin $[a, b]$ (3) A coordinate $k \in [d]$

1. Let $\boldsymbol{\beta}_k$ be the least square fit coefficient computed with labels being $\boldsymbol{u}_a[k], \ldots, \boldsymbol{u}_b[k]$ and co-variates $\boldsymbol{x}_j := [1, j-a+1]^T$ so that the fitted value at time j is given by $\hat{\boldsymbol{u}}_j[k] = \boldsymbol{\beta}^T \boldsymbol{x}_j$.
2. Set $\boldsymbol{\beta}_j[2k-1:2k] \leftarrow \beta_k$ for all $j \in [a, b]$.
3. Set $\boldsymbol{y}_j[2k-1:2k] \leftarrow \boldsymbol{x}_j$ for all $j \in [a, b]$.
4. Define $\boldsymbol{A}_k := \sum_{j=a}^{b} \boldsymbol{x}_j \boldsymbol{x}_j^T$, $\tilde{f}_j(\boldsymbol{v}) := \tilde{f}_j(\boldsymbol{y}_j^T \boldsymbol{v})$ for some $v \in \mathbb{R}^{2d}$.
5. Set $\boldsymbol{\alpha}_k \leftarrow \boldsymbol{\beta}_k - \boldsymbol{A}_k^{-1} \sum_{j=a}^{b} \nabla \tilde{f}(\boldsymbol{\beta}_j)[2k-1:2k]$.
6. Set $\boldsymbol{\alpha}_j[2k-1:2k] \leftarrow \boldsymbol{\alpha}_k$ for all $j \in [a, b]$.

Figure 9: *`AssignCo-variatesAndSlopes2` used to set the parameters in the regret decomposition of Eq.(56) whenever the offline optimal may not be constant across the specified coordinate k within the interval $[a, b]$. We use a 1-based indexing. i.e $\mathbf{v}[1]$ refers the first element of a vector $\mathbf{v}$.*

- Call `AssignCo-variatesAndSlopes1`$(\boldsymbol{u}_{1:n}, [t_2+1, c], k)$.
- Call `AssignCo-variatesAndSlopes1`$(\boldsymbol{u}_{1:n}, [c+1, i_t], k)$.

For a vector $\boldsymbol{y}$ we treat $\boldsymbol{y}[m:n] = [\boldsymbol{y}[m], \ldots, \boldsymbol{y}[n]]^T$. Define $\boldsymbol{X}_j \in \mathbb{R}^{d \times 2d}$ as

$$\boldsymbol{X}_j^T = \begin{bmatrix} \boldsymbol{y}_j[1:2] & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{y}_j[3:4] & \ldots & \mathbf{0} \\ \vdots & \ddots & & \vdots \\ \mathbf{0} & \ldots & & \boldsymbol{y}_j[2d-1:2d] \end{bmatrix}, \tag{57}$$

where $\mathbf{0} = [0, 0]^T$ and $\boldsymbol{y}_j$ is set according to various calls of `AssignCo-variatesAndSlopes1` and `AssignCo-variatesAndSlopes2` as done previously.

We proceed to bound $T_2 + T_3$ in Eq.(56). First notice that due to Taylor's theorem,

$$\tilde{f}_j(\boldsymbol{\alpha}_j) - \tilde{f}_j(\boldsymbol{\beta}_j) = \langle \nabla \tilde{f}_j(\boldsymbol{\beta}_j), \boldsymbol{\alpha}_j - \boldsymbol{\beta}_j \rangle + \frac{1}{2} \|\boldsymbol{\alpha}_j - \boldsymbol{\beta}_j\|^2_{\nabla^2 \tilde{f}_j(\boldsymbol{v})},$$

where $\boldsymbol{v} = t\boldsymbol{\alpha}_j + (1-t)\boldsymbol{\beta}_j$ for some $t \in [0, 1]$. Now we use Lemma 36 to obtain,

$$\begin{aligned} \tilde{f}_j(\boldsymbol{\alpha}_j) - \tilde{f}_j(\boldsymbol{\beta}_j) &\leq \langle \nabla \tilde{f}_j(\boldsymbol{\beta}_j), \boldsymbol{\alpha}_j - \boldsymbol{\beta}_j \rangle + \frac{1}{2} \sum_{k'=1}^{d} \|\boldsymbol{\alpha}_j[2k'-1:2k'] - \boldsymbol{\beta}_j[2k'-1:2k']\|^2_{\boldsymbol{y}_j[2k'-1:2k']\boldsymbol{y}_j[2k'-1:2k']^T} \\ &= \sum_{k'=1}^{d} \langle \nabla f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)[k'] \boldsymbol{y}_j[2k'-1:2k'], \boldsymbol{\alpha}_j[2k'-1:2k'] - \boldsymbol{\beta}_j[2k'-1:2k'] \rangle \\ &\quad + \frac{1}{2} \sum_{k'=1}^{d} \|\boldsymbol{\alpha}_j[2k'-1:2k'] - \boldsymbol{\beta}_j[2k'-1:2k']\|^2_{\boldsymbol{y}_j[2k'-1:2k']\boldsymbol{y}_j[2k'-1:2k']^T} \end{aligned} \tag{58}$$

Further, due to gradient Lipschitzness,

$$\begin{aligned}
\tilde{f}_j(\boldsymbol{\beta}_j) - f_j(\boldsymbol{u}_j) &= f_j(\boldsymbol{X}_j\boldsymbol{\beta}_j) - f_j(\boldsymbol{u}_j) \\
&\le \langle \nabla f_j(\boldsymbol{u}_j), \boldsymbol{X}_j\boldsymbol{\beta}_j - \boldsymbol{u}_j \rangle + \frac{1}{2}\|\boldsymbol{X}_j\boldsymbol{\beta}_j - \boldsymbol{u}_j\|_2^2 \\
&= \sum_{k'=1}^{d} \nabla f_j(\boldsymbol{u}_j)[k'] \cdot \left(\boldsymbol{\beta}_j[2k'-1:2k']^T \boldsymbol{y}_j[2k'-1:2k'] - \boldsymbol{u}_j[k']\right) \\
&\quad + \sum_{k'=1}^{d} \frac{1}{2}\|\boldsymbol{\beta}_j[2k'-1:2k']^T \boldsymbol{y}_j[2k'-1:2k'] - \boldsymbol{u}_j[k']\|_2^2 \qquad (59)
\end{aligned}$$

Looking at Eq.(58) and (59), we see that they decompose across each coordinate k'. So we can bound $T_2 + T_3$ in any bin $[i_s, i_t]$ coordinate wise:

$$\begin{aligned}
T_2 + T_3 = \sum_{k'=1}^{d} \sum_{j=i_s}^{i_t} &\langle \nabla f_j(\boldsymbol{X}_j\boldsymbol{\beta}_j)[k'] \boldsymbol{y}_j[2k'-1:2k'], \boldsymbol{\alpha}_j[2k'-1:2k'] - \boldsymbol{\beta}_j[2k'-1:2k'] \rangle \\
&+ \frac{1}{2}\sum_{k'=1}^{d} \|\boldsymbol{\alpha}_j[2k'-1:2k'] - \boldsymbol{\beta}_j[2k'-1:2k']\|^2_{\boldsymbol{y}_j[2k'-1:2k']\boldsymbol{y}_j[2k'-1:2k']^T} \\
&+ \nabla f_j(\boldsymbol{u}_j)[k'] \cdot \left(\boldsymbol{\beta}_j[2k'-1:2k']^T \boldsymbol{y}_j[2k'-1:2k'] - \boldsymbol{u}_j[k']\right) \\
&+ \frac{1}{2}\|\boldsymbol{\beta}_j[2k'-1:2k']^T \boldsymbol{y}_j[2k'-1:2k'] - \boldsymbol{u}_j[k']\|_2^2 \\
:= \sum_{k'=1}^{d} \sum_{j=i_s}^{i_t} & t_{2,j,k'} + t_{3,j,k'}, \qquad (60)
\end{aligned}$$

where in the last line we define:

$$\begin{aligned}
t_{2,j,k'} &:= \langle \nabla f_j(\boldsymbol{X}_j\boldsymbol{\beta}_j)[k'] \boldsymbol{y}_j[2k'-1:2k'], \boldsymbol{\alpha}_j[2k'-1:2k'] - \boldsymbol{\beta}_j[2k'-1:2k'] \rangle \\
&\quad + \frac{1}{2}\|\boldsymbol{\alpha}_j[2k'-1:2k'] - \boldsymbol{\beta}_j[2k'-1:2k']\|^2_{\boldsymbol{y}_j[2k'-1:2k']\boldsymbol{y}_j[2k'-1:2k']^T},
\end{aligned}$$

and

$$\begin{aligned}
t_{3,j,k'} &:= \nabla f_j(\boldsymbol{u}_j)[k'] \cdot \left(\boldsymbol{\beta}_j[2k'-1:2k']^T \boldsymbol{y}_j[2k'-1:2k'] - \boldsymbol{u}_j[k']\right) \\
&\quad + \frac{1}{2}\|\boldsymbol{\beta}_j[2k'-1:2k']^T \boldsymbol{y}_j[2k'-1:2k'] - \boldsymbol{u}_j[k']\|_2^2.
\end{aligned}$$

Next, we proceed to bound $\sum_{j=i_s}^{i_t} t_{2,j,k} + t_{3,j,k}$ for the coordinate k with a structure as mentioned in Paragraph (A1).

Recall that $[i_s, i_t] = [i_s, a-1] \cup [a, b] \cup [b+1, t_1-1] \cup [t_1, t_2] \cup [t_2+1, c] \cup [c+1, i_t]$. So we will consider each of these sub-bins separately.

For bin $[i_s, a-1]$ we have $\boldsymbol{\alpha}_j[2k-1:2k] = \boldsymbol{\beta}_j[2k-1:2k]$ and $\boldsymbol{\beta}_j[2k-1:2k]^T \boldsymbol{y}_j[2k-1:2k] = \boldsymbol{u}_j[k]$. So we trivially have

$$\sum_{j=i_s}^{a-1} t_{2,j,k} + t_{3,j,k} = 0. \qquad (61)$$

Next, we focus on the bin $[a, b]$. We note that by construction, $\boldsymbol{\alpha}_j[2k-1:2k]$ and $\boldsymbol{\beta}_j[2k-1:2k]$ are fixed for all $j \in [a, b]$. Let's denote these fixed values by $\boldsymbol{\alpha}_k$ and $\boldsymbol{\beta}_k$ respectively. For the sake of brevity let's denote $\boldsymbol{x}_j := \boldsymbol{y}_j[2k-1:2k]$ and

$\boldsymbol{A}_k = \sum_{j=a}^b \boldsymbol{x}_j \boldsymbol{x}_j^T$. We have the relation,

$$\begin{aligned}\boldsymbol{\alpha}_k &= \boldsymbol{\beta}_k - \boldsymbol{A}_k^{-1} \sum_{j=a}^b \nabla \tilde{f}(\boldsymbol{\beta}_j)[2k-1:2k] \\ &= \boldsymbol{\beta}_k - \boldsymbol{A}_k^{-1} \sum_{j=a}^b \nabla f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)[k] \boldsymbol{x}_j. \end{aligned} \tag{62}$$

By the new compact notations, we have

$$t_{2,j,k} = \langle \nabla f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)[k] \boldsymbol{x}_j, \boldsymbol{\alpha}_k - \boldsymbol{\beta}_k \rangle + \frac{1}{2} \|\boldsymbol{\alpha}_k - \boldsymbol{\beta}_k\|^2_{\boldsymbol{x}_j \boldsymbol{x}_j^T},$$

and

$$t_{3,j,k} = \nabla f_j(\boldsymbol{u}_j)[k] \cdot \left(\boldsymbol{\beta}_k^T \boldsymbol{x}_j - \boldsymbol{u}_j[k]\right) + \frac{1}{2} \|\boldsymbol{\beta}_k^T \boldsymbol{x}_j - \boldsymbol{u}_j[k]\|_2^2.$$

From Eq.(62) we have,

$$\begin{aligned}\sum_{j=a}^b t_{2,j,k} &= -\left\|\sum_{j=a}^b \nabla f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)[k] \boldsymbol{x}_j\right\|^2_{\boldsymbol{A}_k^{-1}} + \frac{1}{2}\left\|\boldsymbol{A}_k^{-1} \sum_{j=a}^b \nabla f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)[k] \boldsymbol{x}_j\right\|^2_{\boldsymbol{A}_k} \\ &= -\frac{1}{2}\left\|\sum_{j=a}^b \nabla f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)[k] \boldsymbol{x}_j\right\|^2_{\boldsymbol{A}_k^{-1}} \\ &\le -\frac{1}{2}\left\|\sum_{j=a}^b \nabla f_j(\boldsymbol{u}_j)[k] \boldsymbol{x}_j\right\|^2_{\boldsymbol{A}_k^{-1}} + 2\langle \boldsymbol{A}_k^{-1} \sum_{j=a}^b \nabla f_j(\boldsymbol{u}_j)[k] \boldsymbol{x}_j, \sum_{j=a}^b \boldsymbol{x}_j \left(\nabla f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)[k] - \nabla f_j(\boldsymbol{u}_j)[k]\right)\rangle. \end{aligned}$$

Now we define $\boldsymbol{g}_k = \lambda[-\boldsymbol{s}_{a-2}[k] + \boldsymbol{s}_{a-1}[k] + \boldsymbol{s}_{b-1}[k] - \boldsymbol{s}_b[k], -\boldsymbol{s}_{a-2}[k] + (\ell+1)\boldsymbol{s}_{b-1}[k] - \ell \boldsymbol{s}_b[k]]^T$ and $\boldsymbol{h}_k = [\boldsymbol{\Gamma}[k], \tilde{\boldsymbol{\Gamma}}[k]]^T$ where $\boldsymbol{\Gamma} = \sum_{j=a}^b \gamma_j^- - \gamma_j^+$ and $\tilde{\boldsymbol{\Gamma}} = \sum_{j=a}^b j'(\gamma_j^- - \gamma_j^+)$ where $j' = j - a + 1$ so that $\sum_{j=a}^b \nabla f_j(\boldsymbol{u}_j)[k] \boldsymbol{x}_j = \boldsymbol{g}_k + \boldsymbol{h}_k$ via the KKT conditions in Lemma 31.

With these, we can bound:

$$\begin{aligned}\sum_{j=a}^b 2 \cdot t_{2,j,k} &\le -\|\boldsymbol{g}_k\|^2_{\boldsymbol{A}_k^{-1}} - \|\boldsymbol{h}_k\|^2_{\boldsymbol{A}_k^{-1}} - 2 < \boldsymbol{g}_k, \boldsymbol{A}_k^{-1} \boldsymbol{h}_k \rangle \\ &\quad + 2\langle \boldsymbol{A}_k^{-1}(\boldsymbol{g}_k + \boldsymbol{h}_k), \sum_{j=a}^b \boldsymbol{x}_j \left(\nabla f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)[k] - \nabla f_j(\boldsymbol{u}_j)[k]\right)\rangle \end{aligned} \tag{63}$$

Proceeding similarly to Eq.(25) and (26) by gradient Lipschitzness we obtain,

$$\begin{aligned}\sum_{j=a}^b \nabla f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)[k] - \nabla f_j(\boldsymbol{u}_j)[k] &\le \sum_{j=a}^b \|\boldsymbol{X}_j \boldsymbol{\beta}_j - \boldsymbol{u}_j\|_1 \\ &\le 20\ell^2 \ell^{-3/2}, \end{aligned}$$

where in the last line we used Lemma 20 coordinate-wise and the fact that $\|D^2 \boldsymbol{u}_{a:b}\|_1 \le \ell^{-3/2}$.

Similarly,

$$\begin{aligned}\sum_{j=a}^b j'(\nabla f_j(\boldsymbol{X}_j \boldsymbol{\beta}_j)[k]) - \nabla f_j(\boldsymbol{u}_j)[k] &\le \sum_{j=a}^b j' \|\boldsymbol{X}_j \boldsymbol{\beta}_j - \boldsymbol{u}_j\|_1 \\ &\le 20\ell^3 \ell^{-3/2}. \end{aligned}$$

Hence by KKT conditions in Lemma 31, we can further bound

$$\langle \boldsymbol{A}_k^{-1}\boldsymbol{g}_k, \sum_{j=a}^{b} \boldsymbol{x}_j \left(\nabla f_j(\boldsymbol{X}_j\boldsymbol{\beta}_j)[k] - \nabla f_j(\boldsymbol{u}_j)[k]\right)\rangle \leq \frac{40\lambda\ell^{-1/2}}{(\ell-1)} \Big| (2-2\ell)\boldsymbol{s}_{a-2}[k] + (2\ell+1)\boldsymbol{s}_{a-1}[k] - (\ell+2)\boldsymbol{s}_{b-1}[k] + (\ell-1)\boldsymbol{s}_b[k] \Big| + \frac{40\lambda\ell^{1/2}}{(\ell-1)} \Big| \frac{3(\ell-1)}{\ell+1}\boldsymbol{s}_{a-2}[k] - 3\boldsymbol{s}_{a-1}[k] + 3\boldsymbol{s}_{b-1}[k] - \frac{3(\ell-1)}{\ell+1}\boldsymbol{s}_b[k] \Big|, \tag{64}$$

and

$$\langle \boldsymbol{A}_k^{-1}\boldsymbol{h}_k, \sum_{j=a}^{b} \boldsymbol{x}_j \left(\nabla f_j(\boldsymbol{X}_j\boldsymbol{\beta}_j)[k] - \nabla f_j(\boldsymbol{u}_j)[k]\right)\rangle \leq \frac{40\ell^{-1/2}}{(\ell-1)} |(2\ell+1)\boldsymbol{\Gamma}[k] - 3\tilde{\boldsymbol{\Gamma}}[k]| + \frac{40\ell^{1/2}}{(\ell-1)} \left| \frac{6\tilde{\boldsymbol{\Gamma}}[k]}{\ell+1} - 3\boldsymbol{\Gamma}[k] \right|. \tag{65}$$

We observe that Eq.(63),(64),(65) are semantically same as Eq.(60), (25) and (26) respectively in the 1D case.

Next, we proceed to setup a similar observation for bounding $\sum_{j=a}^{b} t_{3,j,k}$. From KKT conditions in Lemma 31 and proceeding similar to the arguments in Lemma 21 we get,

$$\begin{aligned}\sum_{j=a}^{b} \nabla f_j(\boldsymbol{u}_j)[k] \cdot \left(\boldsymbol{\beta}_j[2k-1:2k]^T \boldsymbol{y}_j[2k-1:2k] - \boldsymbol{u}_j[k]\right) &= \sum_{j=a}^{b} \lambda \Big(((\boldsymbol{s}_{j-1}[k] - \boldsymbol{s}_{j-2}[k]) - (\boldsymbol{s}_j[k] - \boldsymbol{s}_{j-1}[k])) \times \\ &\quad ((j-a+1)\boldsymbol{M}_j[k] + \boldsymbol{C}_j[k]) \Big) \\ &\quad + \sum_{j=a}^{b} (\boldsymbol{\gamma}_j^-[k] - \boldsymbol{\gamma}_j^+[k]) \times \\ &\quad (\boldsymbol{\beta}_j[2k-1:2k]^T \boldsymbol{y}_j[2k-1:2k] - \boldsymbol{u}_j[k]) \\ &\leq \sum_{j=a}^{b} \lambda \Big(((\boldsymbol{s}_{j-1}[k] - \boldsymbol{s}_{j-2}[k]) - (\boldsymbol{s}_j[k] - \boldsymbol{s}_{j-1}[k])) \times \\ &\quad ((j-a+1)\boldsymbol{M}_j[k] + \boldsymbol{C}_j[k]) \Big) \\ &\quad + 20\ell^{-1/2} \sum_{j=a}^{b} \left| \boldsymbol{\gamma}_j^-[k] - \boldsymbol{\gamma}_j^+[k] \right|,\end{aligned}$$

where similar to Lemma 21, we represent $\boldsymbol{\beta}_j[2k-1:2k]^T \boldsymbol{y}_j[2k-1:2k] - \boldsymbol{u}_j[k] = (j-a+1)\boldsymbol{M}_j[k] + \boldsymbol{C}_j[k]$ with $\boldsymbol{M}_a[k] = \boldsymbol{M}_{a+1}[k]$, $\boldsymbol{C}_a[k] = \boldsymbol{C}_{a+1}[k]$, $\boldsymbol{M}_b[k] = \boldsymbol{M}_{b-1}[k]$ and $\boldsymbol{C}_b[k] = \boldsymbol{C}_{b-1}[k]$. The last line is obtained due to Lemma 20.

Further, by using Lemma 20 we obtain,

$$\sum_{k'=1}^{d} \sum_{j=a}^{b} \frac{1}{2} \|\boldsymbol{\beta}_j[2k'-1:2k']^T \boldsymbol{y}_j[2k'-1:2k'] - \boldsymbol{u}_j[k']\|_2^2 \leq 200.$$

Combining the last two inequalities yields,

$$\sum_{j=a}^{b} t_{3,j,k} \leq 200 + \sum_{j=a}^{b} \lambda \Bigg(\left((\boldsymbol{s}_{j-1}[k] - \boldsymbol{s}_{j-2}[k]) - (\boldsymbol{s}_{j}[k] - \boldsymbol{s}_{j-1}[k]) \right) \left((j-a+1) \boldsymbol{M}_j[k] + \boldsymbol{C}_j[k] \right) \Bigg) + 20\ell^{-1/2} \sum_{j=a}^{b} \left| \boldsymbol{\gamma}_j^{-}[k] - \boldsymbol{\gamma}_j^{+}[k] \right|. \tag{66}$$

We observe that the last inequality is semantically similar to Eq.(18) for 1D case. Recall that Eq.(63),(64),(65) are also semantically same as Eq.(60), (27) and (28) respectively in the 1D case.

Hence we can proceed to bound

$$\sum_{j=a}^{b} t_{2,j,k} + t_{3,j,k} = O(1), \tag{67}$$

using the same arguments as in Lemma 29.

Observe that by construction, the slopes across coordinate k are constant in the bins $[b+1, t_1-1]$, $[t_2+1, c]$ and $[c+1, i_t]$. So by using similar arguments used for handling the bin $[i_s, a-1]$ we obtain,

$$\sum_{j \in \mathcal{I}} t_{2,j,k} + t_{3,j,k} = 0, \tag{68}$$

where $\mathcal{I} \in \{[b+1, t_1-1], [t_2+1, c], [c+1, i_t]\}$.

By appealing to our reduction to 1D case facilitated by Eq.(63) and (66) and using similar arguments used to handle the monotonic slopes case as in Lemma 28 we obtain,

$$\sum_{j=t_1}^{t_2} t_{2,j,k} + t_{3,j,k} = O(1). \tag{69}$$

So far we have discussed bounding $\sum_{j=i_s}^{i_t} t_{2,j,k} + t_{3,j,k}$ for a bin with structure across coordinate k as described in Paragraph (A1). We remark that if the slopes across a coordinate k assumes a monotonic structure across $[i_s, i_t]$, we can handle it in the same way as we handled the sub-bin $[t_1, t_2]$ above.

We pause to remark that Eq.(61),(67),(68) and (69) together gives a way to bound to $\sum_{j=i_s}^{i_t} t_{2,j,k'} + t_{3,j,k'}$ across any coordinate k' as we comprehensively considered all the possible structures across a coordinate k'. (The alternate cases where where $s_{a-1} = -1$ and $s_b = 1$ with $\boldsymbol{u}[k']$ non-increasing within $[b+1, c]$ can be handled similarly to the case described in Paragraph (A1). Finally the case where the offline optimal touches boundary 1 instead of -1 can be handled using similar arguments.)

Thus overall we obtain that for any bin $[i_s, i_t] \in \mathcal{P}$ we have:

$$\begin{aligned} T_2 + T_3 &\leq \sum_{k'=1}^{d} \sum_{j=i_s}^{i_t} t_2 t_{2,j,k'} + t_{3,j,k'} \\ &= O(d), \end{aligned}$$

where T_2 and T_3 are as defined in Eq.(56).

Next, we proceed to control T_1. Recall that

$$T_1 = \sum_{j=i_s}^{i_t} f_j(\boldsymbol{p}_j) - f_j(\boldsymbol{X}_j \boldsymbol{\alpha}_j).$$

Let's revisit bin $[i_s, i_t]$ with structure as described in Paragraph (A1) across coordinate k. First we consider the bin $[a, b]$. Through the call to `AssignCo-variatesAndSlopes2`$(\boldsymbol{u}_{1:n}, [a, b], k)$ we set $\boldsymbol{\alpha}_k$. Further $\boldsymbol{\alpha}_j[2k-1 : 2k] = \boldsymbol{\alpha}_k$ for all $j \in [a,b]$. By using similar arguments as in the proof of Lemma 27 which lead to Eq.(34), we have that $|\boldsymbol{y}_j[2k-1:2k]^T \boldsymbol{\alpha}_j| \leq 20$. For other bins such as $[i_s, a-1]$, $[b+1, t_1-1]$, $[t_2+1, c]$, $[c+1, i_t]$ where the slope of the offline optimal across coordinate k remains constant, we set $\alpha_j[2k-1:2k]$ for $j \in \mathcal{I}$ with $\mathcal{I} \in \{[i_s, a-1], [b+1, t_1-1], [t_2+1, c], [c+1, i_t]\}$ to be a constant value obtained as the least square fit coefficients with co-variates $\boldsymbol{y}_j[2k-1:2k]$ and labels set appropriately via the call to `AssignCo-variatesAndSlopes1`. Hence in this case also we have $|\boldsymbol{y}_j^T[2k-1:2k]\boldsymbol{\alpha}_j[2k-1:2k]| \leq 10$ via the arguments in Lemma 27.

For the alternate cases (i) where $s_{a-1} = -1$ and $s_b = 1$ with $\boldsymbol{u}[k']$ non-increasing within $[b+1, c]$ as described in Paragraph (A1) (ii) case where the offline optimal touches boundary 1 instead of -1 (iii) The offline optimal across coordinate k is non-decreasing within $[i_s, i_t]$ and (iv) The offline optimal across coordinate k is non-increasing within $[i_s, i_t]$. In all these cases we can set the quantities $\boldsymbol{\alpha}_j[2k-1:2k]$, $\boldsymbol{y}_j[2k-1:2k]$ by similar calls to `AssignCo-variatesAndSlopes1` or `AssignCo-variatesAndSlopes2` such that $\boldsymbol{y}_j[2k-1:2k]^T\boldsymbol{\alpha}_j[2k-1:2k] \leq 20$ for all $j \in [i_s, i_t]$. For example, for case (iii) we can resort to similar arguments used for handling sub-bin $[t_1, t_2]$ which is again similar to how we handled the bin $[a, b]$. (see Paragraph (A1)).

Further, even-though we create at-most 6 sub-bins across each coordinate for an interval $[i_s, i_t] \in \mathcal{P}$ (see Paragraph (A1) and the sequence of calls beneath), doing so for each coordinate can result in at-most $6d$ partitions of $\boldsymbol{u}_{i_s:i_t}$ overall. However, if we consider any sub-bin $[p,q]$ of this partition, we have that $\boldsymbol{\alpha}_j[2k-1:2k]$ is fixed and $\boldsymbol{\beta}_j[2k-1:2k]$ is fixed for all $j \in [p,q]$ across any coordinate $k \in [d]$ and $y_j[2k-1:2k][2]$ is monotonically increasing wrt $j \in [p,q]$ for all coordinates $k \in [d]$. Now suppose that $k' \in [d]$ is such that $\boldsymbol{y}_j[2k'-1:2k'][1] \leq \boldsymbol{y}_j[2k-1:2k][1]$ for all $k \neq k'$ and for all $j \in [p,q]$. With a change of variables we have that $\tilde{\boldsymbol{\alpha}}_j[2k-1:2k]^T\boldsymbol{y}_j[2k'-1:2k'] = \boldsymbol{\alpha}_j[2k-1:2k]^T\boldsymbol{y}_j[2k-1:2k]$ by setting $\tilde{\boldsymbol{\alpha}}_j[2k-1:2k][2] = \boldsymbol{\alpha}_j[2k-1:2k][2]$ and $\tilde{\boldsymbol{\alpha}}_j[1] = \boldsymbol{\alpha}_j[1] + (\boldsymbol{y}_j[2k-1:2k][2] - \boldsymbol{y}_j[2k'-1:2k'][2])\boldsymbol{\alpha}_j[2k-1:2k][2]$ for $k \neq k'$ within the bin $[p,q]$. Since $(\boldsymbol{y}_j[2k-1:2k][2] - \boldsymbol{y}_j[2k'-1:2k'][2]) \leq \boldsymbol{y}_j[2k-1:2k][2]$ by Eq.(32) we have that

$$|(\boldsymbol{y}_j[2k-1:2k][2] - \boldsymbol{y}_j[2k'-1:2k'][2])\boldsymbol{\alpha}_j[2k-1:2k][2]| \leq 6. \tag{70}$$

Further we have from Eq.(34) that $|\boldsymbol{\alpha}_j[2k-1:2k][2]\boldsymbol{y}_j[2k'-1:2k'][2] + \boldsymbol{\alpha}_j[2k-1:2k][1]| \leq 20$ due to the fact that $\boldsymbol{\alpha}_j[2k-1:2k]$ remains fixed from a time point $j^* \leq p$ such that $\boldsymbol{y}_{j^*}[2k-1:2k] = [1,1]^T$. Further we have that

$$\begin{aligned}
\|\tilde{\boldsymbol{\alpha}}_j[2k-1:2k]\|_2^2 &= (\boldsymbol{\alpha}_j[2k-1:2k][2])^2 \\
&\quad + (\boldsymbol{\alpha}_j[2k-1:2k][1] + (\boldsymbol{y}_j[2k-1:2k][2] - \boldsymbol{y}_j[2k'-1:2k'][2])\boldsymbol{\alpha}_j[2k-1:2k][2])^2 \\
&\leq 2\left(\boldsymbol{\alpha}_j[2k-1:2k][2] + \boldsymbol{\alpha}_j[2k-1:2k][1]\right)^2 \\
&\quad + 2\left((\boldsymbol{y}_j[2k-1:2k][2] - \boldsymbol{y}_j[2k'-1:2k'][2])\boldsymbol{\alpha}_j[2k-1:2k][2]\right)^2 \\
&\leq 584,
\end{aligned} \tag{71}$$

where the last line is due to Eq.(35) and (70).

Let's represent $\boldsymbol{\mu} \in \mathbb{R}^{2d}$ such that $\boldsymbol{\mu}[2k'-1:2k'] = \boldsymbol{\alpha}[2k'-1:2k']$ and $\boldsymbol{\mu}[2k-1:2k] = \boldsymbol{\alpha}[2k-1:2k]$ for all other $k \in [d]$.

Thus within the sub-bin $[p,q]$, we have that $|\boldsymbol{\mu}^T[2k-1:2k]\boldsymbol{y}_j[2k'-1:2k']| \leq 20$ for all $k \in [d]$. Further, due to Eq.(71) we have that $\|\boldsymbol{\mu}\|_2^2 \leq 584d$. Hence we can use a base expert that starts at time p which gives the co-variate $\boldsymbol{y}_j[2k'-1:2k']$ to all coordinates where $j \in [p,q]$. Note that the sub-bin $[p,q]$ must have been resulted via a splitting across coordinate k' at time p. So by the calls to `AssignCo-variatesAndSlopes1` or `AssignCo-variatesAndSlopes2` we set $\boldsymbol{y}_p[2k'-1:2k'] = [1,1]^T$. Thus there exists a base expert in FLH-SIONS (Fig.3) that provides the co-variate $\boldsymbol{y}_j[2k'-1:2k']$ to all coordinates where $j \in [p,q]$.

This expert will have a regret of $\tilde{O}(d)$ against $\boldsymbol{\mu}$ via Lemma 32. By using Strong Adaptivity from Corollary 33 (set $\boldsymbol{w} = \boldsymbol{\mu}$ there and recall that $\|\boldsymbol{\mu}\|_2^2 \leq 584d$) and adding the regret across all $6d$ sub-bins of $[i_s, i_t]$ lead to an $\tilde{O}(d^2)$ on T_1 in Eq.(56). Thus for any bin in $\mathcal{P}$ produced by generate bins procedure, we have its dynamic regret bounded by $\tilde{O}(d^2)$.

□

Proof of Theorem 3. The proof is now complete by adding the $\tilde{O}(d^2)$ dynamic regret bound across all $O(n^{1/5}C_n^{2/5} \vee 1)$ bins in $\mathcal{P}$ from Corollary 34.

□

The proof of Lemma 18 is same as that of the lemma below, albeit with slightly different notations for $\boldsymbol{X}_j$.

Lemma 36. *Let $\boldsymbol{X}_j$ be as defined in Eq.(57). Let $\tilde{f}_j(\boldsymbol{v}) = f_j(\boldsymbol{X}_j \boldsymbol{v})$ for some $\boldsymbol{v} \in \mathbb{R}^{2d}$ and let $\boldsymbol{\Sigma} := \boldsymbol{X}_j^T \boldsymbol{X}_j \in \mathbb{R}^{2d \times 2d}$. We have,*

$$\nabla^2 \tilde{f}_j(\boldsymbol{v}) \preccurlyeq \boldsymbol{\Sigma}$$

Proof. We have,

$$\tilde{f}_j(\boldsymbol{v}) = f_j\left(\langle \boldsymbol{y}_j[1:2], \boldsymbol{v}[1:2]\rangle, \ldots, \langle \boldsymbol{y}_j[2d-1:2d], \boldsymbol{v}[2d-1:2d]\rangle\right).$$

Let

$$f''_{jk} := \nabla^2 f_j\left(\langle \boldsymbol{y}_j[1:2], \boldsymbol{v}[1:2]\rangle, \ldots, \langle \boldsymbol{y}_j[2d-1:2d], \boldsymbol{y}_j[2d-1:2d]\rangle\right)[j][k],$$

be the Hessian of f evaluated at the vector $[\langle \boldsymbol{y}[1:2], \boldsymbol{v}[1:2]\rangle, \ldots, \langle \boldsymbol{y}[2d-1:2d], \boldsymbol{v}[2d-1:2d]\rangle]^T \in \mathbb{R}^d$.

By straightforward calculations, we obtain

$$\nabla^2 \tilde{f}_j(\boldsymbol{v}) = \begin{bmatrix} f''_{11}\boldsymbol{y}_j[1:2]\boldsymbol{y}_j[1:2]^T & \ldots & f''_{1d}\boldsymbol{y}_j[1:2]\boldsymbol{y}_j[2d-1:2d]^T \\ \vdots & \ddots & \vdots \\ f''_{d1}\boldsymbol{y}_j[2d-1:2d]\boldsymbol{y}_j[1:2]^T & \ldots & \boldsymbol{f}''_{dd}\boldsymbol{y}_j[2d-1:2d]\boldsymbol{y}_j[2d-1:2d]^T \end{bmatrix},$$

Let $\boldsymbol{I} \in \mathbb{R}^{d \times d}$ be the identity matrix and $\boldsymbol{1} \in \mathbb{R}^{2 \times 2}$ be the matrix of all ones. Further let's denote $\boldsymbol{b} := [\langle \boldsymbol{y}[1:2], \boldsymbol{v}[1:2]\rangle, \ldots, \langle \boldsymbol{y}[2d-1:2d], \boldsymbol{v}[2d-1:2d]\rangle]^T$ We can succinctly write:

$$\boldsymbol{\Sigma} - \nabla^2 \tilde{f}_j(\boldsymbol{v}) = \left(\left(\boldsymbol{I} - \nabla^2 f(\boldsymbol{b})\right) \otimes \boldsymbol{1}\right) \circ \boldsymbol{y}_j \boldsymbol{y}_j^T,$$

where $\otimes$ denotes the Kronecker product and $\circ$ denotes the Hadamard product.

Recall that the loss functions f_j are 1-gradient Lipschitz. So we have $\left(\boldsymbol{I} - \nabla^2 f(\boldsymbol{b})\right)$ is Positive Semi Definite (PSD). The matrices $\boldsymbol{1}$ and $\boldsymbol{y}_j \boldsymbol{y}_j^T$ are also PSD. Since both Kronecker and Hadamard products preserves positive semidefiniteness, we have $\nabla^2 \tilde{f}_j(\boldsymbol{v}) \preccurlyeq \boldsymbol{\Sigma}$ which proves the lemma.

□

Proposition 37. *Consider the sequence class $\mathcal{TV}^1(C_n)$ as per Eq.(3). Under Assumption A1 (see Section 4) we have that $\mathcal{TV}^{(1)}(C_n) \subseteq \mathcal{TV}^{(0)}(2C_n + 20d)$.*

Proof. We start by considering a 1D setting. Consider a sequence $w_{1:n} \in \mathcal{TV}^{(1)}(C_n)$. We can represent it as sum (point-wise) of two sequences as

$$w_{1:n} = p_{1:n} + q_{1:n}, \tag{72}$$

where $q_{1:n} = \boldsymbol{\beta}^T \boldsymbol{x}_t$ where $\boldsymbol{x}_t = [1, t]^T$ and $\boldsymbol{\beta}$ is the least square fit coefficnts computed by using covariates $\boldsymbol{x}_t$ and labels w_t, $t \in [n]$. Here the $p_{1:n}$ is the residual sequence obtained by subtracting the least square fit sequence from the true sequence.

Following the terminology in Lemma 21, we can represent $p_t = tM_t + C_t$. Further, due to Eq.(19) (with $a = 1$) we have that $p_{t+1} - p_t = M_{t+1}$.

Applying triangle inequality to Eq.(72) we have

$$\|Dw_{1:n}\|_1 \leq \|Dp_{1:n}\|_1 + \|Dq_{1:n}\|_1.$$

Further,

$$\begin{aligned}
\|Dp_{1:n}\|_1 &= \sum_{t=2}^{n} |M_t| \\
&= \sum_{t=2}^{n} \left| M_1 + \sum_{j=1}^{t-1} M_{j+1} - M_j \right| \\
&\le \sum_{t=2}^{n} |M_1| + D^2 \|p_{1:n}\|_1 \\
&=_{(a)} n|M_1| + nD^2 \|w_{1:n}\|_1 \\
&\le_{(b)} 2nD^2 \|w_{1:n}\|_1,
\end{aligned}$$

where in line (a) we used the fact that $\|D^2 p_{1:n}\|_1 = \|D^2 w_{1:n}\|_1$ as subtracting a linear sequence doesn't affect the TV1 distance. In line (b) we applied $|M_1| \le \|D^2 w_{1:n}\|_1$ as shown in Lemma 21.

It remains to bound $\|Dq_{1:n}\|_1$. For this we note that $\|q_t\| \le 10$ for all $t \in [n]$ due to Eq.(53). Since $q_{1:n}$ is a monotonic sequence we have that its variation $\|Dq_{1:n}\|_1 \le 20$.

Thus overall we obtain that

$$\begin{aligned}
\|Dw_{1:n}\|_1 &\le 2nD^2 \|w_{1:n}\|_1 + 20 \\
&\le 2C_n + 20.
\end{aligned}$$

For multiple dimensions we apply the same argument across each dimension and add them up to yield the lemma.

□

D Proof of Proposition 5

In this section, we first prove the following result.

Theorem 38. *Let $\boldsymbol{p}_t$ be the predictions of FLH-SIONS algorithm with parameters $\epsilon = 2$, $C = 20$ and exp-concavity factor σ. Under Assumptions A1-A4, we have that,*

$$\sum_{t=1}^{n} f_t(\boldsymbol{p}_t) - f_t(\boldsymbol{w}_t) = \tilde{O}(d^2 n^{1/3} C_n^{2/3} \vee d^2),$$

for any $C_n > 0$ and any comparator sequence $\boldsymbol{w}_{1:n} \in \mathcal{TV}^{(0)}(C_n)$. Here $\tilde{O}$ hides poly-logarithmic factors of n and $a \vee b = \max\{a, b\}$.

Proof. The proof follows almost directly from the arguments in Baby and Wang (2021). First, we use the partition $\mathcal{P}$ mentioned in Lemma 30 in Baby and Wang (2021). Let the partition be $cP = \{[1_s, 1_t], \ldots, [M_s, M_t]\}$, with $|\mathcal{P}| = M$.

Consider the following convex optimization problem.

$$\begin{aligned}
\min_{\tilde{\boldsymbol{u}}_1, \ldots, \tilde{\boldsymbol{u}}_n, \tilde{\boldsymbol{z}}_1, \ldots, \tilde{\boldsymbol{z}}_{n-1}} \quad & \sum_{t=1}^{n} f_t(\tilde{\boldsymbol{u}}_t) \\
\text{s.t.} \quad & \tilde{\boldsymbol{z}}_t = \tilde{\boldsymbol{u}}_{t+1} - \tilde{\boldsymbol{u}}_t \;\forall t \in [n-1], \\
& \sum_{t=1}^{n-1} \|\tilde{\boldsymbol{z}}_t\|_1 \le C_n, \qquad (73a) \\
& \|\tilde{\boldsymbol{u}}_t\|_\infty \le B \;\forall t \in [n], \qquad (73b)
\end{aligned}$$

Let $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ be the optimal solution to the above problem. Let $\boldsymbol{w}_j$ be the prediction of the FLH-SIONS algorithm at time j. Define:

$$R_n(C_n) = \sum_{t=1}^{n} f_j(\boldsymbol{w}_t) - f_t(\boldsymbol{u}_t).$$

Define $\bar{\boldsymbol{u}}_i = \frac{1}{n_i}\sum_{j=i_s}^{i_t} \boldsymbol{u}_j$ and $\dot{\boldsymbol{u}}_i = \bar{\boldsymbol{u}}_i - \frac{1}{n_i}\sum_{j=i_s}^{i_t} \nabla f_j(\bar{\boldsymbol{u}}_i)$. We can use the regret decomposition of Baby and Wang (2021).

$$R_n(C_n) \le \sum_{i=1}^{M} \underbrace{\sum_{j=i_s}^{i_t} f_j(\boldsymbol{w}_j) - f_j(\dot{\boldsymbol{u}}_i)}_{T_{1,i}} + \sum_{i=1}^{M} \underbrace{\sum_{j=i_s}^{i_t} f_j(\dot{\boldsymbol{u}}_i) - f_j(\bar{\boldsymbol{u}}_i)}_{T_{2,i}} + \sum_{i=1}^{M} \underbrace{\sum_{j=i_s}^{i_t} f_j(\bar{\boldsymbol{u}}_i) - f_j(\boldsymbol{u}_j)}_{T_{3,i}}.$$

For any bin $[i_s, i_t] \in \mathcal{P}$, we can bound $T_{2,i} + T_{3,i} = O(1)$ by using the arguments in the proof of Theorem 14 of Baby and Wang (2021) since the losses in our case are also gradient-Lipschitz as per Assumption A3. So we only need to consider the term $T_{1,i}$. Observe that

$$\begin{aligned}\|\dot{\boldsymbol{u}}_i\|_\infty &\le \|\bar{\boldsymbol{u}}_i\|_\infty + \frac{1}{n_i}\sum_{j=i_s}^{i_t} \|\nabla f_j(\bar{\boldsymbol{u}}_i)\|_\infty \\ &\le 2,\end{aligned}$$

as per Assumptions A1-A2. Further we can view the comparator $\dot{\boldsymbol{u}}_i$ as a linear predictor with slope zero. The output of this linear predictor is bounded in magnitude by 2 which is less that 20. Hence FLH-SIONS under the setting of the current theorem leads to $T_{1,i} = \tilde{O}(d)$. Since $M = O(dn^{1/3}C_n^{2/3} \vee d)$ for the partition in Lemma 30 of adding the regret across all bins results in the theorem. ☐

Theorem38 when combined with Theorem 3 now directly leads to Proposition 5.

E Proof of Proposition 2

The result proven in this section is mainly due to the geometric arguments in Donoho et al. (1990); Donoho and Johnstone (1998) (or see Johnstone (2017) for a comprehensive monograph) with an extra technicality of handling boundedness constraint as per Assumption A1 (in Section 4).

In the proof we make extensive use of wavelet theory and refer readers to Johnstone (2017) for necessary preliminaries.

Proposition 2. *Under Assumptions A1-A4, any online algorithm necessarily suffers* $\sup_{\boldsymbol{w}_{1:n} \text{ with } \mathcal{TV}_1(\boldsymbol{w}_{1:n}) \le C_n} R_n(\boldsymbol{w}_{1:n}) = \Omega(d^{3/5}n^{1/5}C_n^{2/5} \vee d)$.

Proof. We consider a uni-variate setting with the losses $f_t(w) = (d_t - w)^2$ where $d_t = u_t + \mathcal{N}(0,1)$ with $u_{1:n} \in \mathcal{TV}^{(1)}(C_n)$. At each step, d_t is revealed to the learner as doing so can only make learning easier.

Let $\mathbb{W}$ be the set of whole numbers. For the purposes of analysis, we start with an abstract observation model:

$$y_j = \theta_j + \epsilon\mathcal{N}(0,1),\ j \in \mathbb{W} \tag{74}$$

where θ_j are the wavelet coefficients in a regularity-three CDJV multi-resolution basis (Cohen et al., 1993) of a function in $\mathcal{F}_1(C_n)$ from which the discrete samples $u_{1:n}$ are generated.

In what follows we will show that for any procedure estimating the wavelet coefficients (let the estimate be $\hat{\theta}_j$, $j \in \mathbb{W}$) we have that

$$\sum_{j\in\mathbb{W}} (\hat{\theta}_j - \theta_j)^2 = \Omega(C^{2/5}\epsilon^{8/5}).$$

Due to Section 15.5 of (Johnstone, 2017), by taking $\epsilon = 1/\sqrt{n}$, such a guarantee will then imply a lower bound of $\Omega(n^{-4/5}C^{2/5})$ for $\frac{1}{n}\sum_{t=1}^{n}(u_t - \hat{u}_t)^2$, where $\hat{u}_t$ is the estimate produced by observing the data d_t (assume $C = \Omega(1/\sqrt{n})$ for now). This rate will finally imply a dynamic regret lower bound in the following manner:

$$
\begin{aligned}
E\left[\sup_{r_{1:n}\in\mathcal{TV}^{(1)}(C)} \sum_{t=1}^{n} f_t(\hat{u}_t) - f_t(r_t)\right] &\geq \sup_{r_{1:n}\in\mathcal{TV}^{(1)}(C)} E\left[\sum_{t=1}^{n} f_t(\hat{u}_t) - f_t(r_t)\right] \\
&=_{(a)} \sup_{r_{1:n}\in\mathcal{TV}^{(1)}(C)} \sum_{t=1}^{n} E[(\hat{u}_t - u_t)^2] - (r_t - u_t)^2 \\
&= \sum_{t=1}^{n} E[(\hat{u}_t - u_t)^2], \qquad (75)
\end{aligned}
$$

where in line (a) we used the bias variance decomposition and the fact that $\hat{u}_t$ is independent of d_t for online algorithms.

In what follows we use a dyadic indexing scheme for referring to wavelet coefficients in Eq.(74) as θ_{jk} which means the k^{th} wavelet coefficient in resolution $j \geq 0$. There are 2^j wavelet coefficients in resolution j. We will also use $\theta_{j\cdot}$ to denote a sequence of 2^j wavelet coefficients at resolution j.

Let β be the subset of wavelet coefficients at resolutions less than or equal to 2. i.e, $\beta = [\theta_{0\cdot}, \theta_{1\cdot}, \theta_{2\cdot}]$ which has a length of 7.

Define a Besov norm as follows:

$$\|\theta\|_{b^{3/2}_{1,1}} := \|\beta\|_1 + \sum_{j\geq 3} 2^{3j/2}\|\theta_{j\cdot}\|_1.$$

Define a Besov space as:

$$\mathcal{A}(B) := \{\theta : \|\theta\|_{b^{3/2}_{1,1}} \leq B\}.$$

It is known that $A(\kappa C) \subseteq \mathcal{F}_1(C)$ for some constant $0 < \kappa \leq 1$. (see for eg. Eq.(33) in (Tibshirani, 2014) along with Theorem 1 in (Donoho and Johnstone, 1998)).

Since the space $\mathcal{A}(B)$ is solid and orthosymmetric (see Section 4.8 in Johnstone (2017)) we have that the risk of estimating coefficients from $\mathcal{A}$ is lower bounded by the risk (i.e $\sum_{j\geq 0}(\hat{\theta}_j - \theta_j)^2$) of the hardest rectangular sub-problem as shown by Donoho et al. (1990).

A hyper-rectangle is defined as follows:

$$\Theta(\tau) = \{\theta : |\theta_j| \leq \tau_j,\ j \geq 0\}.$$

From Donoho et al. (1990), the minimax risk over a hyper-rectangle under the observation model Eq.(74) is known to be:

$$
\begin{aligned}
R^*(\tau) &:= \min_{\hat{\theta}} \max_{\theta\in\Theta(\tau)} \sum_{j\geq 0}(\hat{\theta}_j - \theta)^2 \\
&\geq \sum_{j\geq 0} \min\{\tau_j^2, \epsilon^2\}.
\end{aligned}
$$

So all we need to show is an appropriate hyper-rectangle (which is identified by τ) within $\mathcal{A}(B)$ whose minimax risk is sufficiently large.

We next proceed to give such a hyper-rectangle. Let $j_* \in \mathbb{W}$ be the smallest number such that

$$2^{j_*} \geq \frac{C^{2/5}}{\epsilon^{2/5}}.$$

For simplicity, from now on-wards, let's assume that j_* is an integer that satisfy $2^{j_*} = \frac{C^{2/5}}{\epsilon^{2/5}}$.

Define the hyper-rectangle coordinates by

$$\tau_{j_* k} = \frac{\kappa C}{2^{5 j_*/2}}, \tag{76}$$

for all $k = 0, 1, \ldots, 2^{j_*} - 1$ and $\tau_{j \cdot} = 0$ for all other resolutions.

Note that $\frac{\kappa C}{2^{5 j_*/2}} = \epsilon$. The minimax risk over such a hyper-rectangle then becomes

$$\begin{aligned} R^*(\tau) &= 2^{j_*}\epsilon \\ &= (\kappa C)^{2/5}\epsilon^{8/5}. \end{aligned}$$

Now it remains to verify that

1. The hyper-rectangle in Eq.(76) is indeed in $\mathcal{A}(\kappa C)$.
2. The function produced by the coefficients in that hyper rectangle is bounded by 1 point-wise in magnitude.

First we notice that by taking $\epsilon = 1/\sqrt{n}$ as mentioned earlier, we have

$$2^{j_*} > 4,$$

whenever $C > 4^{5/2}/\sqrt{n}$. We first consider the case where C is within this regime.

For the first item, we have that

$$\begin{aligned} \|\tau\|_{b_{1,1}^{3/2}} &= 2^{3j_*/2} \cdot 2^{j_*} \frac{\kappa C}{2^{5j_*/2}} \\ &= \kappa C, \end{aligned}$$

where we used the fact that $j_* > 2$ in the regime $C > 4^{5/2}/\sqrt{n}$.

Hence $\Theta(\tau) \subseteq \mathcal{A}(\kappa C)$.

For the second item, we notice that due to Lemma B.18 in Johnstone (2017), it is sufficient to show that $2^{j_*/2}\|\theta_{j_* \cdot}\|_\infty = O(1)$. Taking $\epsilon = 1/\sqrt{n}$ as mentioned earlier, we have that

$$\begin{aligned} 2^{j_*/2}\|\theta_{j_* \cdot}\|_\infty &= \frac{\kappa C}{2^{2j_*}} \\ &= \kappa^{1/5} C^{1/5} \epsilon^{4/5} \\ &= \frac{\kappa^{1/5} C^{1/5}}{n^{2/5}} \\ &\leq 1, \end{aligned}$$

in the non-trivial regime of $C \leq n^2$ where we recall that $\kappa \leq 1$.

For the regime where $C \leq 1/\sqrt{n}$, the trivial lower bound of $\Omega(1)$ estimation error kicks in. Thus overall we have shown that for any online algorithm producing estimates $\hat{u}_t$ we have that

$$\sum_{t=1}^{n} E[(\hat{u}_t - u_t)^2] = \Omega(n^{1/5} C^{2/5} \vee 1),$$

thus obtaining a lower bound on the dynamic regret as per Eq.(75).

In multiple-dimensions we can consider a similar setup as before with losses $f_t(\boldsymbol{w}) = \|\boldsymbol{d}_t - \boldsymbol{w}\|_2^2$ with $\boldsymbol{d}_t[k] = \boldsymbol{u}_t[k] + \mathcal{N}(0,1)$ where $\boldsymbol{u}_{1:n} \in \mathcal{TV}^{(1)}(C)$. We can consider a sequence $\boldsymbol{u}_{1:n}$ such that $\|nD^2 \boldsymbol{u}_{1:n}[k]\|_1 = C/d$ across each coordinate $k \in [d]$.

$$\begin{aligned} \min_{\boldsymbol{p}_{1:n}} \max_{\boldsymbol{w}_{1:n} \in \mathcal{TV}^{(1)}(C)} \sum_{t=1}^{n} f_t(\boldsymbol{p}_t) - f_t(\boldsymbol{w}_t) &= \sum_{k=1}^{d} \sum_{t=1}^{n} \Omega(n^{1/5} (C/d)^{2/5} \vee 1) \\ &= \Omega(d^{3/5} n^{1/5} C^{2/5} \vee d). \end{aligned}$$

This completes the proof of the proposition.

Next, we verify the fact in Remark 6 that the rate of $n^{1/3}[\mathcal{TV}_0]^{1/3}$ is of the same order as $n^{1/5}[\mathcal{TV}_1]^{2/5}$ for the comparator sequence constructed above.

Define the following norm:

$$\|\theta\|_{b_{1,1}^{1/2}} := \|\beta\|_1 + \sum_{j\geq 2} 2^{j/2} \|\theta_{j\cdot}\|_1.$$

Define another Besov space as:

$$\mathcal{G}(B) := \{\theta : \|\theta\|_{b_{1,1}^{1/2}} \leq B\}.$$

Let $w_{1:n}$ be the sequence of reals and let $\theta_{1:n}$ be its wavelet coefficients. It is known from DeVore and Lorentz (1993) that

$$\|\theta\|_{b_{1,1}^{1/2}} \asymp \mathcal{TV}_0(w_{1:n}),$$

where $\asymp$ means that the quantities have similar scaling.

So we only need to compute the norm $\|\theta\|_{b_{1,1}^{1/2}}$ for the hard instance created above. We have that:

$$\begin{aligned}\|\theta\|_{b_{1,1}^{1/2}} &= 2^{j_*/2} \cdot 2^{j_*} \kappa C / 2^{5j_*/2} \\ &= \kappa C / 2^{j_*} \\ &= \kappa C^{3/5} / n^{1/5}.\end{aligned}$$

Thus the $\mathcal{TV}_0$ of the sequence scales as $C^{3/5}/n^{1/5}$. Hence the rate:

$$n^{1/3}[\mathcal{TV}_0(w_{1:n})]^{2/3} \asymp n^{1/5} C^{2/5}.$$

Thus for the hard instance constructed in the proof, both the rates grow with similar scale.

□

F Why the analysis of Baby and Wang (2021) leads to sub-optimal regret?

For simplicity, we consider a uni-variate setting. First we derive a tighter regret guarantee (than one implied by Proposition 37) of $O(n^{1/3} C_n^{2/3} \vee 1)$ for the results of Baby and Wang (2021) when applied to our setting. Then we explain the source of sub-optimality in their analysis. Throughout this section, we assume that the condition of low TV1 regime defined in Section 1 is satisfied.

First, let's define the comparator classes:

$$\mathcal{TV}^{(1)}(C) := \{\theta_{1:n} : \mathcal{TV}_1(\theta_{1:n}) \leq C\},$$

and

$$\mathcal{TV}^{(0)}(C) := \{\theta_{1:n} : \mathcal{TV}_0(\theta_{1:n}) \leq C\}.$$

Let $u_{1:n}$ be the offline optimal sequence as per Lemma 17.

In accordance to the details in Section 3, we can interpret a comparator sequence $u_{1:n} \in \mathcal{TV}^{(1)}(C_n)$ as a continuous piece-wise linear sequence. Then the dynamic regret can be expressed as:

$$\begin{aligned} R_n(u_{1:n}) &= \sum_{t=1}^{n} f_t(p_t) - f_t(u_t) \\ &=_{(a)} \sum_{t=1}^{n} f_t(\boldsymbol{\alpha}_t^T \boldsymbol{x}_t) - f_t(\boldsymbol{\beta}_t^T \boldsymbol{x}_t) \\ &:=_{(b)} \sum_{t=1}^{n} \tilde{f}_t(\boldsymbol{\alpha}_t) - \tilde{f}_t(\boldsymbol{\beta}_t), \end{aligned}$$

where in Line (a) we define $\boldsymbol{x}_t = [1, t/n]^T$ and $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are chosen such that $p_t = \boldsymbol{\alpha}_t^T \boldsymbol{x}_t$ and $u_t = \boldsymbol{\beta}_t^T \boldsymbol{x}_t$. Further the predictors $\boldsymbol{\beta}_t$ are chosen to satisfy $\boldsymbol{\beta}_t^T \boldsymbol{x}_t = \boldsymbol{\beta}_{t+1}^T \boldsymbol{x}_t$ so that the sequence $u_{1:n}$ can be interpreted as a piece-wise linear signal that is also continuous at every transition point where the slope changes (see Definition 1).

In Line (b) we define $\tilde{f}_t(\boldsymbol{v}) = f_t(\boldsymbol{v}^T \boldsymbol{x}_t)$. We chose the co-variates as $\boldsymbol{x}_t = [1, t/n]^T$ instead of $\boldsymbol{x}_t = [1, t]^T$ so that the losses $\tilde{f}_t(\boldsymbol{v})$ remains Lipschitz and gradient Lipschitz whenever $|\boldsymbol{v}^T \boldsymbol{x}_t| = O(1)$ which is a requirement for the results in (Baby and Wang, 2021).

By using similar line of arguments used to derive (19), we obtain

$$\boldsymbol{\beta}_{t+1}[1] - \boldsymbol{\beta}_t[1] = \frac{t}{n} \left(\boldsymbol{\beta}_t[2] - \boldsymbol{\beta}_{t+1}[2] \right).$$

Hence we have that

$$\begin{aligned} \sum_{t=1}^{n-1} |\boldsymbol{\beta}_{t+1}[1] - \boldsymbol{\beta}_t[1]| &\leq \sum_{t=1}^{n-1} |\boldsymbol{\beta}_t[2] - \boldsymbol{\beta}_{t+1}[2]| \\ &= n \|D^2 u_{1:n}\|_1, \end{aligned}$$

where we used the fact that the sum of difference of the slopes (see Definition 1) in the linear representation of $u_{1:n}$ with co-variates $\boldsymbol{x}_t = [1, t/n]^T$ is exactly equal to $n\|D^2 u_{1:n}\|_1$.

Thus overall, we obtain that

$$\begin{aligned} \sum_{t=1}^{n-1} \|\boldsymbol{\beta}_t - \boldsymbol{\beta}_{t+1}\|_1 &\leq 2n \|D^2 u_{1:n}\|_1 \\ &\leq 2C_n, \end{aligned} \tag{77}$$

as $u_{1:n} \in \mathcal{TV}^{(1)}(C_n)$.

Hence by the results of Baby and Wang (2021) we have that

$$R_n(u_{1:n}) = \tilde{O}(n^{1/3} C_n^{2/3} \vee 1).$$

Next, we proceed to explain source of this sub-optimality in the analysis of Baby and Wang (2021).

In Baby and Wang (2021) (Lemma 5) they form a partition $\mathcal{P}'$ of $\boldsymbol{\beta}_{1:n}$ so that in the i^{th} bin (represented by $[i_s, i_t]$) we have that:

- $\|D\boldsymbol{\beta}_{i_s:i_t}\| \leq 1/\sqrt{\ell_{i_s \to i_t}}$
- $\|D\boldsymbol{\beta}_{i_s:i_t+1}\| > 1/\sqrt{\ell_{i_s \to i_t+1}}$

where we recall that $\ell_{a \to b} = b - a + 1$.

So we have that within bin $[i_s, i_t] \in \mathcal{P}'$, $\|D\boldsymbol{\beta}_{i_s:i_t}[2]\|_1 \leq 1/\sqrt{\ell_{i_s \to i_t}}$. This amounts to saying that

$$\|D^2 u_{i_s:i_t}\|_1 \leq 1/(n\sqrt{\ell_{i_s \to i_t}}).$$

While in the partition $\mathcal{P}$ that we construct in Lemma 10 we have that

$$\|D^2 u_{i_s:i_t}\|_1 \leq 1/\ell_{i_s \to i_t}{}^{3/2}.$$

Comparing the previous two inequalities, we conclude that the sequence within each bin of $\mathcal{P}'$ is much smoother than that of $\mathcal{P}$.

This will result in the formation of $|\mathcal{P}'| = O(n^{1/3} C_n^{2/3} \vee 1)$ bins overall as per Eq.(77) (see Lemma 5 in Baby and Wang (2021)) which is larger than the $O(n^{1/5} C_n^{2/5} \vee 1)$ bins in $\mathcal{P}$ under the low TV1 regime.

Within each bin $[i_s, i_t] \in \mathcal{P}'$ Baby and Wang (2021) uses a three term regret decomposition as follows:

$$\begin{aligned} T_{[i_s,i_t]} &:= \sum_{j=i_s}^{i_t} \tilde{f}_j(\boldsymbol{\alpha}_j) - \tilde{f}_j(\boldsymbol{\beta}_j) \\ &= \underbrace{\sum_{j=i_s}^{i_t} \tilde{f}_j(\boldsymbol{\alpha}_j) - \tilde{f}_j(\dot{\boldsymbol{\beta}})}_{T_1'} + \underbrace{\sum_{j=i_s}^{i_t} \tilde{f}_j(\dot{\boldsymbol{\beta}}) - \tilde{f}_j(\bar{\boldsymbol{\beta}})}_{T_2'} + \underbrace{\sum_{j=i_s}^{i_t} \tilde{f}_j(\bar{\boldsymbol{\beta}}) - \tilde{f}_j(\boldsymbol{\beta}_j)}_{T_3'}, \end{aligned} \tag{78}$$

where $\bar{\boldsymbol{\beta}} = \frac{1}{n}\sum_{j=i_s}^{i_t} \boldsymbol{\beta}_j$ and $\dot{\boldsymbol{\beta}} = \bar{\boldsymbol{\beta}} - \frac{1}{\ell_{i_s \to i_t}} \sum_{j=i_s}^{i_t} \nabla \tilde{f}_j(\bar{\boldsymbol{\beta}})$.

Then Baby and Wang (2021) proceed to show that this one step gradient descent based decomposition is sufficient to keep $T_{[i_s,i_t]} = O(1)$ leading to an overall regret of $O(n^{1/3} C^{2/3} \vee 1)$ when summed across all bins.

In our case the main challenge is to keep $T_{[i_s,i_t]} = \tilde{O}(1)$ for $[i_s, i_t] \in \mathcal{P}$ while dealing with the fact that sequence within each bin of $\mathcal{P}$ is much less smooth than that in $\mathcal{P}'$. We accomplish this via a newton step based decomposition with a careful analysis as detailed in Section 4.1 (It was found that the one-step gradient descent as in Eq.(78) doesn't keep T_2' negative enough to make $T_2' + T_3' = O(1)$ for bins in $\mathcal{P}$). Eventhough the sequence in bins $\mathcal{P}$ is wigglier than that of bins in $\mathcal{P}'$, overall the sequence, $u_{1:n}$ from a $\mathcal{TV}^{(1)}$ class is much smoother than the sequences from $\mathcal{TV}^{(0)}$ class in the low TV1 regime due to sufficiently slowly changing piecewise linear structure. This extra smoothness property is what allowed us to consider *larger* (in terms of mean bin width) bins and hence smaller partition size (when compared to $\mathcal{P}'$) and still bound the regret within each bin to be $\tilde{O}(1)$. Adding this bound across all bins in $\mathcal{P}$ then lead to the optimal rate of $\tilde{O}(n^{1/5} C_n^{2/5} \vee 1)$.

G More examples from the low TV1 regime

We list some examples where the low TV1 regime defined in Section 1 is satisfied. Under this regime, the rate of $\tilde{O}(n^{1/5}[\mathcal{TV}_1(\boldsymbol{w}_{1:n})]^{2/5} \vee 1)$ attained by FLH-SIONS via Theorem 3 is faster than the rate of $\tilde{O}(n^{1/3}[\mathcal{TV}_0(\boldsymbol{w}_{1:n})]^{2/3} \vee 1)$ attained by Baby and Wang (2021). This is a non-exhaustive list of examples and one can construct many other examples as well. All the examples we consider here are for uni-variate setting, through the extension to multi-dimensions is a straight-forward replication of the sequence generating process across each coordinate.

We begin by a minimalist example yielding logarithmic dynamic regret rate.

Example 39. *Consider a sequence $\theta_{1:n}$ such that $\theta_t = t/n$ for $t \in \{0, 1, \ldots, n\}$. This is a sequence obtained via descretizing a linear signal. We have that $\mathcal{TV}_1(\theta_{1:n}) = 0$ and $\mathcal{TV}_0(\theta_{1:n}) = 1$. So by Theorem 3 we have that the rate attained by FLH-SIONS is $O(\log n)$ while the rate attained by Baby and Wang (2021) is $O(n^{1/3})$.*

Next, we give an example where both $\mathcal{TV}_1$ and $\mathcal{TV}_0$ distance of a sequence is growing with n.

Example 40. *For an integer $s < n$, define $a_{1:s} = 0, s/n, 2s/n, \ldots, \frac{s(n/s-1)}{n}$. Let $b_{1:s}$ be the mirror image of $a_{1:s}$.i.e $b_{1:s} = \frac{s(n/s-1)}{n}, \ldots, s/n, 0$. For simplicity lets' assume that n/s is an integer. Form a sequence $\theta_{1:n} := a_{1:s}, b_{1:s}, a_{1:s}, b_{1:s}, \ldots, a_{1:s}, b_{1:s}$ by concatenating the sequences $a_{1:s}, b_{1:s}$ for $s/2$ times. This sequence transitions between 0 and 1 through linear sections. For this sequence, we have that $\mathcal{TV}_0(\theta_{1:n}) = s$ and $\mathcal{TV}_1(\theta_{1:n}) = 2s^2$. Let $s = n^\alpha$ for some $0 < \alpha < 1$. Thus Theorem 3 yields a rate of $\tilde{O}(n^{\frac{4\alpha+1}{5}})$ while the results in Baby and Wang (2021) yields only a rate of $\tilde{O}(n^{\frac{2\alpha+1}{3}})$ which is a slower rate for all $\alpha < 1$.*