Adaptive Realizable Regression from Two Online Learners

Dirk van der Hoeven
Mathematical Institute, Leiden University
Dirk@dirkvanderhoeven.com

Abstract

We study realizable regression with a finite or countably infinite function class $\mathcal{F}$. We are given $D_n = \{(X_t, Y_t)\}_{t=1}^n$, which is assumed to be i.i.d. with $f^\star \in \mathcal{F}$ such that $\mathbb{E}[Y_t|X_t] = f^\star(X_t)$. The goal is to estimate $f^\star$ with $\widehat{f}$ under the absolute loss $\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|]$. We propose a simple reduction that runs two instances of online learning algorithms: one minimizes a learned loss and the other minimizes the error induced by learning the loss. Our main result is a prior-dependent second-order bound: for any prior π over $\mathcal{F}$, the estimator achieves with probability at least $1-\delta$,

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \le C\sqrt{\frac{\mathbb{E}_X[\sigma(X)]\,\Gamma(f^\star)}{n}} + \frac{CB\,\Gamma(f^\star)}{n}\,,$$

where $\sigma(X) = \mathbb{E}_Y[(Y - f^\star(X))^2 \mid X]$ is the conditional variance, $\Gamma(f^\star) = \log(1/\delta) + \log((1+\log(Bn))/\pi(f^\star))$, $C > 0$ is an absolute constant, and B is an upper bound on $|Y|$ and $|f(X)|$. The bound adapts automatically to the noise level $\mathbb{E}_X[\sigma(X)]$ and range B without any prior knowledge of it, achieving the fast rate $O(B\Gamma(f^\star)/n)$ when $\sigma(X) = 0$ and recovering the slow rate $O(B\sqrt{\Gamma(f^\star)/n})$ in the worst case. The dependence on $\mathcal{F}$ enters only through $\log(1/\pi(f^\star))$, enabling faster rates for structured $\mathcal{F}$ or sparse $f^\star$ and extending naturally to countably infinite classes via summable priors. For finite $\mathcal{F}$ we further show that the same algorithm produces an empirically verifiable bound that allows the learner to evaluate the quality of $\widehat{f}$ from the data alone, without knowledge of $\mathbb{E}_X[\sigma(X)]$. Finally, we extend our reduction to general convex loss functions satisfying $\ell(f(x), f(x)) = 0$.

1 Introduction

We are given a dataset $D_n = \{(X_t, Y_t)\}_{t=1}^n$ where $X_t \in \mathcal{X}$ is the feature of the t-th datapoint and $Y_t \in [-B, B]$ is its label. We assume that $(X_t, Y_t) \sim \mathcal{D}_{X,Y}$ is identically and independently distributed with $\mathbb{E}[Y_t|X_t] = f^\star(X_t)$, where $f^\star \in \mathcal{F}$ and $|f(x)| \le B$ for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$.

The goal is to output an $\widehat{f}$ such that

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \tag{1}$$

is as small as possible, where the expectation is over unobserved X.

This realizable regression problem arises naturally as a subroutine in several important problems in decision making. For example, in contextual bandits, an agent must repeatedly select actions based on observed contexts, and the reduction of Simchi-Levi and Xu [2022] shows that an optimal contextual bandit algorithm can be constructed from a regression oracle that solves a closely related problem. Similarly, in reinforcement learning, regression oracles are used to estimate reward or value functions

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

from logged data, and the realizability assumption is natural when the function class is chosen to contain the true reward function.

The worst-case optimal algorithm (see e.g. Audibert [2007]) guarantees that with probability at least $1-\delta$, for some universal constant $C > 0$,

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \le CB\sqrt{\frac{\log(|\mathcal{F}|/\delta)}{n}}\,.$$

This result follows from minimizing the squared loss since $\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \le \sqrt{\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|^2]}$ by Jensen's inequality.

It is important to note that the objective (1) is a different objective than a common objective which is to control the risk of $\widehat{f}$. When we want to control the risk of $\widehat{f}$, the objective is to output an $\widehat{f}$ such that

$$\mathbb{E}_X[\mathbb{E}_Y[|\widehat{f}(X) - Y||X]]\,, \tag{2}$$

is small. In fact, (2) is an upper bound on (1) by Jensen's inequality. As a consequence, a bound on the risk leads to a bound on (1) that does not vanish to zero with increasing sample sizes unless there exists an $f \in \mathcal{F}$ such that $f(X) = Y$ for all $X \in \mathcal{X}$.

Contributions. We propose a simple reduction of realizable regression to online learning, described in Algorithm 1, which runs two online learning algorithms $\mathcal{A}$ and $\mathcal{B}$ in parallel. The first algorithm $\mathcal{A}$ predicts, while the second algorithm $\mathcal{B}$ is used to construct the losses for $\mathcal{A}$. The final estimator $\widehat{f} = \frac{1}{n}\sum_{t=1}^n f_t^A$ is the average of the predictions of $\mathcal{A}$. Our main contributions are as follows.

- **A simple and general reduction.** Theorem 1 in Section 2 shows that the estimation error of $\widehat{f}$ can be upper bounded by the sum of the regrets of $\mathcal{A}$ and $\mathcal{B}$ plus a martingale difference term. This reduction is agnostic to the choice of online learning algorithms and holds for any $\mathcal{A}$ and $\mathcal{B}$, allowing one to plug in any algorithm with a known regret bound.
- **Prior-dependent second-order bounds.** Theorem 2 shows that instantiating $\mathcal{A}$ and $\mathcal{B}$ with Squint+L [Mhammedi et al., 2019] with a prior π over a finite or countably infinite $\mathcal{F}$ achieves, with probability at least $1-\delta$,

  $$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \le C\sqrt{\frac{\mathbb{E}_X[\sigma(X)]\,\Gamma(f^\star)}{n}} + \frac{CB\,\Gamma(f^\star)}{n}\,,$$

  where $\Gamma(f^\star) = \log(1/\delta) + \log((1+\log(Bn))/\pi(f^\star))$. Crucially, the rate depends on $\mathcal{F}$ only through $\log(1/\pi(f^\star))$ rather than $\log|\mathcal{F}|$. This allows one to obtain faster rates for simple $f^\star$ by choosing π to concentrate on simple functions, and extends naturally to countably infinite function classes via summable priors. We also show that knowledge of B is not necessary and provide improved logarithmic dependence on n for finite $\mathcal{F}$ compared to prior work.
- **Empirical bounds.** Theorem 3 in Section 4 provides a version of the above bound expressed entirely in terms of observable quantities: the cumulative training loss $\sum_{t=1}^n \ell_t^A(f_t^A(X_t))$, where ℓ_t^A is the loss generated by $\mathcal{B}$ for $\mathcal{A}$ and the cumulative squared residuals $\sum_{t=1}^n (f_t^A(X_t) - Y_t)^2$. This allows the learner to evaluate the quality of $\widehat{f}$ from the data alone, without any knowledge of $\mathbb{E}_X[\sigma(X)]$.
- **Extensions.** We extend our results to general convex loss functions satisfying $\ell(f(x), f(x)) = 0$ (Section 5).

Related Work In a related setting called cost-sensitive classification Foster and Krishnamurthy [2021] provide first-order bounds, which were later improved and extended to the setting studied in this paper by Li and Jun [2025]. Li and Jun [2025] show that if $Y \in [0,1]$ and $f(X) \in [0,1]$ for all $X \in \mathcal{X}$ and $f \in \mathcal{F}$, minimizing the log-loss guarantees that with probability at least $1-\delta$, for some universal constant $C > 0$,

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \le C\sqrt{\mathbb{E}_X[f^\star(X)(1-f^\star(X))]\frac{\log(|\mathcal{F}|/\delta)}{n}} + \frac{\log(|\mathcal{F}|/\delta)}{n}\,.$$

Li and Jun [2025] also propose a novel loss function called the betting loss, which they show leads to variance-adaptive second-order generalization bounds without any knowledge of the conditional variances. Specifically, their algorithm guarantees that with probability at least $1-\delta$, for some universal constant $C > 0$,

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \le C\sqrt{\mathbb{E}_X[\mathbb{E}_Y[(f^\star(X)-Y)^2|X]]\frac{\log(n|\mathcal{F}|/\delta)}{n}} + \frac{\log(n|\mathcal{F}|/\delta)}{n}.$$

Ignoring $\log(n)$ terms, since $\mathbb{E}_X[\mathbb{E}_Y[(f^\star(X)-Y)^2|X]] \le \mathbb{E}_X[f^\star(X)(1-f^\star(X))]$, this latter result is strictly better. The results of Li and Jun [2025] are closely related in spirit to ours, but there are several structural reasons why their approach cannot obtain the results we derive in this paper.

Some of our bounds depend on the function class $\mathcal{F}$ only through $\log(1/\pi(f^\star))$. The approach of Li and Jun [2025] instead relies on the empirical risk minimizer of the betting loss and the analysis relies on uniform convergence arguments, which immediately leads to $\log(|\mathcal{F}|)$ scaling in the bounds.

Theorem 3 provides a bound expressed entirely in terms of observable quantities, which allows the learner to evaluate the quality of $\widehat{f}$ from the data alone. The approach of Li and Jun [2025] does not appear to yield an analogous empirical bound, since their analysis is based on uniform convergence arguments that involve expectations over the data-generating distribution and are therefore not directly computable from the observed data.

A related and classical technique for converting online learning guarantees into batch statistical estimation bounds is called online-to-batch conversion, which can be traced back to [Vapnik and Chervonenkis, 1974]. With online-to-batch conversion one outputs the uniform average of an online learner's predictions and shows that the excess risk is bounded by the average regret. The standard conversion, however, controls the risk $\mathbb{E}_X[\mathbb{E}_Y[|\widehat{f}(X) - Y||X]]$ rather than the estimation error $\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|]$, and as a consequence the resulting bound contains an irreducible noise term that does not vanish as $n \to \infty$; this is precisely the distinction between objectives (1) and (2) discussed in the introduction. Our two-algorithm reduction can be seen as a refined online-to-batch conversion that sidesteps this issue by constructing losses whose comparator term $\ell_t^A(f^\star)$ has conditional expectation zero, so the noise does not accumulate and the bound on $\mathbb{E}_X[|\widehat{f}(X)-f^\star(X)|]$ vanishes with n even in the presence of irreducible noise.

Our work fits naturally into a field referred to as algorithmic statistics [Clerico et al., 2025]. Broadly, in algorithmic statistics one reduces statistical estimation problems to sequential prediction by showing that low regret implies high-probability statistical guarantees. Representative works in this line include Waudby-Smith and Ramdas [2024], Orabona and Jun [2024], Clerico et al. [2025], Kirschner et al. [2025], and Lugosi and Neu [2023]. Lugosi and Neu [2023] provide an online-to-pac framework where a regret bound implies a generalization guarantee. Similarly, Waudby-Smith and Ramdas [2024], Orabona and Jun [2024], Clerico et al. [2025], Kirschner et al. [2025] reduce mean estimation of random variables and parameter estimation in generalized linear models to sequential prediction problems and construct martingales whose concentration properties yield confidence sequences. Our reduction is in the same spirit: Theorem 1 below shows that the estimation error $\mathbb{E}_X[|\widehat{f}(X)-f^\star(X)|]$ is controlled by the regrets of two online learning algorithms plus a martingale difference term, so that any pair of online learning algorithms with known regret bounds immediately yields a statistical guarantee.

Notation We denote by $\mathbb{E}_t[\cdot] = \mathbb{E}[\cdot|D_{t-1}]$ and use the convention that $\text{sign}(0) = 0$.

2 Algorithm

In this section we introduce our algorithm: Algorithm 1. Algorithm 1 has a simple structure: it runs two online learning algorithms $\mathcal{A}$ and $\mathcal{B}$ in parallel, and outputs the average of the predictions of $\mathcal{A}$. The role of the two algorithms is asymmetric. Algorithm $\mathcal{A}$ is the primary learner: it produces the predictions f_t^A whose average forms the final estimator $\widehat{f}$. Algorithm $\mathcal{B}$ plays a supporting role: it maintains a distribution p_t^B over $\mathcal{F}$ that is used to construct the loss ℓ_t^A fed to $\mathcal{A}$. Intuitively, $\mathcal{B}$ tries to identify functions in $\mathcal{F}$ that lie on the correct side of $f_t^A(X_t)$ relative to the observed label Y_t: the loss $\ell_t^B(f)$ is minimized when $\text{sign}(f_t^A(X_t) - f(X_t))$ matches $\text{sign}(f_t^A(X_t) - Y_t)$, so $\mathcal{B}$ is rewarded for placing mass on functions that point in the direction of the residual $f_t^A(X_t) - Y_t$.

Algorithm 1: Reduction to Online Learning

Input Online Learning Algorithms $\mathcal{A}, \mathcal{B}$
for $t = 1, \ldots, n$ **do**
  Obtain f_t^A from $\mathcal{A}$ and $p_t^B \in \triangle$ from $\mathcal{B}$.
  Send $\ell_t^A(f) = (f(X_t) - Y_t) \sum_{f' \in \mathcal{F}} p_t^B(f')\text{sign}(f_t^A(X_t) - f'(X_t))$ to algorithm $\mathcal{A}$ as the loss in round t.
  Send $\ell_t^B(f) = -(f_t^A(X_t) - Y_t)\text{sign}(f_t^A(X_t) - f(X_t))$ to algorithm $\mathcal{B}$ as the loss in round t.
end
Output $\widehat{f} = \frac{1}{n}\sum_{t=1}^n f_t^A$

The distribution p_t^B thus serves as a data-driven proxy for $\text{sign}(f_t^A(X_t) - f^\star(X_t))$, and $\mathcal{A}$ is in turn penalized through ℓ_t^A for residuals $f_t^A(X_t) - Y_t$ that align with this estimated direction of error.

The key identity underlying Theorem 1 below is that

$$\sum_{t=1}^n (f_t^A(X_t) - Y_t)\text{sign}(f_t^A(X_t) - f^\star(X_t)) = \sum_{t=1}^n \ell_t^A(f_t^A) + \sum_{t=1}^n \Big(\sum_{f \in \mathcal{F}} p_t^B(f)\ell_t^B(f) - \ell_t^B(f^\star) \Big),$$

where $\ell_t^A(f) = (f(X_t) - Y_t) \sum_{f' \in \mathcal{F}} p_t^B(f')\text{sign}(f_t^A(X_t) - f'(X_t))$ and $\ell_t^B(f) = -(f_t^A(X_t) - Y_t)\text{sign}(f_t^A(X_t) - f(X_t))$ are the losses for algorithms $\mathcal{A}$ and $\mathcal{B}$, respectively. The key idea is that we have now upper bounded $\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|]$ (up to a martingale error) by the cumulative loss of $\mathcal{A}$ plus the regret of $\mathcal{B}$. The regret of $\mathcal{A}$ then appears when one further subtracts $\sum_{t=1}^n \ell_t^A(f^\star)$ from both sides, which satisfies $\sum_{t=1}^n \mathbb{E}_t[\ell_t^A(f^\star)] = 0$. This decomposition is the heart of the reduction: it converts the estimation problem into two separate online learning problems, each of which can be solved with off-the-shelf algorithms. Assuming that $\mathcal{A}$ and $\mathcal{B}$ have $O(|\mathcal{F}|)$ per round computational complexity, the computational efficiency to find $\widehat{f}$ is $O(|\mathcal{F}|n)$.

Theorem 1. *Denote by*

$$R_n^{\mathcal{A}}(f^\star) = \sum_{t=1}^n \big(\ell_t^A(f_t^A) - \ell_t^A(f^\star)\big)$$

$$R_n^{\mathcal{B}}(f^\star) = \sum_{t=1}^n \big(\sum_{f \in \mathcal{F}} p_t^B(f)\ell_t^B(f) - \ell_t^B(f^\star)\big)$$

$$Z_t = \mathbb{E}_t[\big(f_t^A(X_t) - Y_t\big)\text{sign}(f_t^A(X_t) - f^\star(X_t))] - \big(f_t^A(X_t) - Y_t\big)\text{sign}(f_t^A(X_t) - f^\star(X_t)).$$

For any $\mathcal{A}$ and $\mathcal{B}$ Algorithm 1 guarantees that

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \le \frac{1}{n}\Big(R_n^{\mathcal{A}}(f^\star) + R_n^{\mathcal{B}}(f^\star) + \sum_{t=1}^n \big(\ell_t^A(f^\star) + Z_t\big)\Big).$$

Proof. We have that

$$\begin{aligned}
\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] &\le \mathbb{E}_X[\frac{1}{n}\sum_{t=1}^n |f_t^A(X) - f^\star(X)|] && \text{(Jensen's inequality)} \\
&= \mathbb{E}_X[\frac{1}{n}\sum_{t=1}^n \big(f_t^A(X) - f^\star(X)\big)\text{sign}(f_t^A(X) - f^\star(X))] \\
&= \mathbb{E}_{X,Y}[\frac{1}{n}\sum_{t=1}^n \big(f_t^A(X) - Y\big)\text{sign}(f_t^A(X) - f^\star(X))] && (\mathbb{E}[Y|X] = f(X)) \\
&= \frac{1}{n}\Big(\sum_{t=1}^n \big(f_t^A(X_t) - Y_t\big)\text{sign}(f_t^A(X_t) - f^\star(X_t)) + \sum_{t=1}^n Z_t\Big).
\end{aligned}$$

Next, we have that

$$\begin{aligned}
&\sum_{t=1}^{n} \big(f_t^A(X_t) - Y_t\big)\text{sign}(f_t^A(X_t) - f^\star(X_t)) \\
&= \sum_{t=1}^{n} \big(f_t^A(X_t) - Y_t\big) \sum_{f\in\mathcal{F}} p_t^B(f)\text{sign}(f_t^A(X_t) - f(X_t)) \\
&\quad + \sum_{t=1}^{n} \big(f_t^A(X_t) - Y_t\big)\Big(\text{sign}(f_t^A(X_t) - f^\star(X_t)) - \sum_{f\in\mathcal{F}} p_t^B(f)\text{sign}(f_t^A(X_t) - f(X_t))\Big) \\
&= \sum_{t=1}^{n} \ell_t^A(f_t^A) + \sum_{t=1}^{n} \Big(\sum_{f\in\mathcal{F}} p_t^B(f)\ell_t^B(f) - \ell_t^B(f^\star)\Big) \\
&= \sum_{t=1}^{n} \ell_t^A(f^\star) + R_n^{\mathcal{A}}(f^\star) + R_n^{\mathcal{B}}(f^\star)\,,
\end{aligned}$$

where the last equality is due to definition of $R_n^{\mathcal{A}}(f^\star)$ and $R_n^{\mathcal{B}}(f^\star)$. Thus, we have that

$$\frac{1}{n}\sum_{t=1}^{n} \mathbb{E}_t[|f_t^A(X_t) - f^\star(X_t)|] = \frac{1}{n}\Big(\sum_{t=1}^{n}\big(\ell_t^A(f^\star) + Z_t\big) + R_n^{\mathcal{A}}(f^\star) + R_n^{\mathcal{B}}(f^\star)\Big)\,, \tag{3}$$

which concludes the proof. □

To see why the guarantee of Theorem 1 is meaningful consider the following. First, there exist many algorithms such as the exponential weights algorithm [Vovk, 1990, Littlestone and Warmuth, 1994] that can guarantee that $R_n^{\mathcal{A}}(f^\star)$ and $R_n^{\mathcal{B}}(f^\star)$ are of order $B\sqrt{n\log(|\mathcal{F}|)}$. Second, observe that since $\mathbb{E}_Y[f^\star(X) - Y|X] = 0$ we have that $\mathbb{E}_t[\ell_t^A(f^\star) + Z_t] = 0$ and thus by standard concentration inequalities for martingale difference sequences such as Freedman's inequality (Lemma 6) we have that with probability at least $1-\delta$,

$$\sum_{t=1}^{n}\big(\ell_t^A(f^\star) + Z_t\big) = O(B\sqrt{n\log(1/\delta)})\,.$$

Therefore, by Theorem 1, for $\mathcal{A}$ and $\mathcal{B}$ such that $R_n^{\mathcal{A}}(f^\star) \le B\sqrt{n\log(|\mathcal{F}|)}$ and $R_n^{\mathcal{B}}(f^\star) \le B\sqrt{n\log(|\mathcal{F}|)}$ we have that with probability at least $1-\delta$,

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] = O\left(\frac{B\sqrt{\log(|\mathcal{F}|/\delta)}}{n}\right).$$

While such a result has been known, the generality of Theorem 1 allows us to improve upon known results and derive new results.

3 Improved Second-Order Bounds

In this section we turn Theorem 1 into a concrete statistical guarantee by instantiating $\mathcal{A}$ and $\mathcal{B}$ with Squint+L [Mhammedi et al., 2019] (see Appendix F for a description). Squint+L with prior π has a second-order property that is precisely what is needed to convert the regret bounds of Theorem 1 into a variance-adaptive estimation guarantee: the regret terms inherit the noise dependence of the underlying losses, the martingale term contributes a $\sqrt{\mathbb{E}_X[\sigma(X)]}$ factor via Freedman's inequality, and the two combine to give a bound that interpolates between the fast rate $O(B\log(\log(nB)/(\pi(f^\star)\delta))/n)$ when $\sigma(X) = 0$ and the slow rate $O(\sqrt{\mathbb{E}_X[\sigma(X)]\log(\log(nB)/(\pi(f^\star)\delta))/n})$ in the high-noise regime, with the function-class complexity entering only through $\log(1/\pi(f^\star))$.

Theorem 2. *Let $\mathcal{A}$ and $\mathcal{B}$ be Squint+L with prior π. Algorithm 1 guarantees that with probability at least $1-\delta$, for some absolute constant $C > 0$,*

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \le \frac{CB\,\Gamma(f^\star)}{n} + C\sqrt{\frac{\mathbb{E}_X[\sigma(X)]\,\Gamma(f^\star)}{n}}\,,$$

where $\Gamma(f^\star) = \log(1/\delta) + \log\big((1+\log(Bn))/\pi(f^\star)\big)$.

Proof sketch. Write $S_A = \sum_{t=1}^n \mathbb{E}_t[|f_t^A(X_t) - f^\star(X_t)|]$ and $\widetilde{\Gamma} = \log((2+\log(2Bn))/\pi(f^\star)) + \frac{1}{2}$. By the regret guarantee of Squint+L (Theorem 12), the regrets of $\mathcal{A}$ and $\mathcal{B}$ are controlled by $\sqrt{\sum_{t=1}^n (f_t^A(X_t) - f^\star(X_t))^2}$ and $\sqrt{\sum_{t=1}^n (f_t^A(X_t) - Y_t)^2}$ respectively, each scaled by $1 + \sqrt{2\widetilde{\Gamma}}$ and shifted by an additive $O(B\widetilde{\Gamma})$. The first square root reduces to $\sqrt{2B\sum_{t=1}^n |f_t^A(X_t) - f^\star(X_t)|}$, which AM-GM converts into $\frac{1}{8}S_A + O(B\widetilde{\Gamma})$. For the second, the decomposition $(f_t^A - Y_t)^2 \le 2(f_t^A - f^\star)^2 + 2(f^\star - Y_t)^2$ splits the regret into a piece absorbed by another $\frac{1}{8}S_A$ and a residual proportional to $(1+\sqrt{\widetilde{\Gamma}})\sqrt{\sum_{t=1}^n (f^\star(X_t) - Y_t)^2}$. The sum $\sum_{t=1}^n (f^\star(X_t) - Y_t)^2$ concentrates around $n\,\mathbb{E}_X[\sigma(X)]$ by Freedman's inequality, turning this residual into the $\sqrt{n\,\mathbb{E}_X[\sigma(X)]\,\Gamma(f^\star)}$ term that gives the bound its second-order character.

The martingale term $\sum_{t=1}^n (\ell_t^A(f^\star) + Z_t)$ is handled analogously, contributing another $\frac{1}{4}S_A$ to be absorbed along with a $\sqrt{n\,\mathbb{E}_X[\sigma(X)]\log(1/\delta)}$ residual. Plugging these into Theorem 1 (specifically, see (3)) yields

$$S_A \le \tfrac{1}{2}S_A + O\big(\sqrt{n\,\mathbb{E}_X[\sigma(X)]\,\Gamma(f^\star)}\big) + O(B\,\Gamma(f^\star))\,,$$

and rearranging gives the claimed rate. The full proof is deferred to Appendix A. □

The bound in Theorem 2 has several noteworthy features. First, the rate adapts automatically to the noise level $\mathbb{E}_X[\sigma(X)]$ without any prior knowledge of it: in the noiseless case $\sigma(X) = 0$ the bound reduces to the fast rate $O(B\Gamma(f^\star)/n)$, while in the high-noise case $\mathbb{E}_X[\sigma(X)] = \Theta(B^2)$ it recovers the slow rate $O(B\sqrt{\Gamma(f^\star)/n})$, and interpolates smoothly between the two. Second, the dependence on the function class enters only through $\log(1/\pi(f^\star))$, which for a uniform prior over a finite class reduces to $\log|\mathcal{F}|$, recovering the classical rate. However, for a non-uniform prior π that assigns more mass to simpler functions, the rate can be substantially faster: if $\pi(f^\star)$ is bounded away from zero, the $\log(1/\pi(f^\star))$ term is $O(1)$ and the bound reduces to $O(\sqrt{\mathbb{E}_X[\sigma(X)]/n})$, which does not depend on $|\mathcal{F}|$ at all. This is particularly useful for large or infinite function classes equipped with a summable prior. Third, in contrast to Li and Jun [2025], the algorithm does not require B as an input parameter yet it has an improved logarithmic dependence on n: there appears a $\log\log(n)$ term rather than a $\log(n)$ term. Finally, we note that the bound involves $\mathbb{E}_X[\sigma(X)]$, the average conditional variance, rather than the worst-case variance $\sup_x \sigma(x)$. This means the bound can be much tighter when the noise is concentrated on a small region of $\mathcal{X}$, and reflects the true difficulty of the problem rather than a worst-case proxy.

3.1 Example applications

Prior-dependent rates have proven crucial in a number of statistical applications, including PAC-Bayesian bounds for sparse regression [Alquier and Lounici, 2011], oracle inequalities for the aggregation of affine estimators [Dalalyan and Salmon, 2012], and inference in compound functional models [Dalalyan et al., 2014]. Below, we provide two concrete applications of our results.

Nested function classes. A natural application is when $\mathcal{F} = \bigcup_{m=1}^\infty \mathcal{F}_m$ is a countably infinite union of finite classes $\mathcal{F}_m$ of increasing complexity, for instance polynomial function classes of increasing degree or neural networks of increasing width. Setting

$$\pi(f) = \frac{6}{\pi^2 m^2} \cdot \frac{1}{|\mathcal{F}_m|}\,, \qquad f \in \mathcal{F}_m\,,$$

gives $\log(1/\pi(f^\star)) = O(\log m^\star + \log|\mathcal{F}_{m^\star}|)$ for $f^\star \in \mathcal{F}_{m^\star}$, so the rate adapts automatically to the smallest class $\mathcal{F}_{m^\star}$ that contains $f^\star$, up to a logarithmic overhead in the index $m^\star$. This recovers an oracle inequality over the union of classes without any prior knowledge of which class contains $f^\star$. Specifically, let us consider generalized linear models with an unknown norm. Let $\mathcal{X} \subseteq \mathbb{R}^d$ with $\|x\|_2 \le 1$ and consider the generalized linear model class

$$\mathcal{F} = \{f_{\boldsymbol{w}} : x \mapsto \mu(\langle \boldsymbol{w}, x\rangle) \mid \boldsymbol{w} \in \mathbb{R}^d\}\,,$$

where $\mu : \mathbb{R} \to [0,1]$ is a G-Lipschitz link function and we place no bound on $\|\boldsymbol{w}\|_2$. For $k = 1, 2, \ldots$, define the shells

$$\mathcal{S}_k = \{\boldsymbol{w} \in \mathbb{R}^d : k - 1 < \|\boldsymbol{w}\|_2 \le k\}\,,$$

and for each shell construct a cover $\mathcal{F}_k$ of the corresponding functions in sup-norm at resolution $\varepsilon = 1/n$, induced by a $1/(Gn)$-cover of $\mathcal{S}_k$ in ℓ_2-norm. By a standard volumetric argument (see e.g. Vershynin [2026, Section 4.2]),

$$\log |\mathcal{F}_k| \le d \log(3Gkn) \,.$$

We place the prior $\pi(f_{\boldsymbol{w}_\varepsilon}) = \frac{6}{\pi^2 k(f_{\boldsymbol{w}_\varepsilon})^2} \cdot \frac{1}{|\mathcal{F}_{k(f_{\boldsymbol{w}_\varepsilon})}|}, \quad f_{\boldsymbol{w}_\varepsilon} \in \mathcal{F}_{k(f_{\boldsymbol{w}_\varepsilon})}$, where $k(f_{\boldsymbol{w}_\varepsilon})$ denotes the shell index of $f_{\boldsymbol{w}_\varepsilon}$. This is a valid prior since

$$\sum_{k=1}^{\infty} \sum_{f_{\boldsymbol{w}_\varepsilon} \in \mathcal{F}_k} \pi(f_{\boldsymbol{w}_\varepsilon}) = \sum_{k=1}^{\infty} \frac{6}{\pi^2 k^2} = 1 \,.$$

For $f^\star$ with $k^\star = \lceil \|\boldsymbol{w}^\star\|_2 \rceil$, the nearest cover point $f^\star_\varepsilon \in \mathcal{F}_{k^\star}$ satisfies $\|f^\star - f^\star_\varepsilon\|_\infty \le 1/n$ and

$$\log(1/\pi(f^\star_\varepsilon)) = \log\left(\frac{\pi^2 (k^\star)^2}{6}\right) + \log |\mathcal{F}_{k^\star}| \le 2\log(k^\star) + d\log(3Gk^\star n) + O(1) \,.$$

Since $k^\star \le \|\boldsymbol{w}^\star\|_2 + 1$, we have $\log(1/\pi(f^\star_\varepsilon)) = O(d \log(G\|\boldsymbol{w}^\star\|_2 n))$, where the $2\log(k^\star)$ term from the prior is absorbed into the $d\log(G\|\boldsymbol{w}^\star\|_2 n)$ term for $d \ge 1$. Applying Theorem 2, we may conclude that with probability at least $1-\delta$,

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \le C\sqrt{\frac{\mathbb{E}_X[\sigma(X)]\,\Gamma^\star}{n}} + \frac{C\,\Gamma^\star}{n} \,, \tag{4}$$

where $\Gamma^\star = O(d\log(G\|\boldsymbol{w}^\star\|_2 n) + \log(1/\delta))$ and $C > 0$ is an absolute constant. This matches the rate one would obtain with an oracle that knows $\|\boldsymbol{w}^\star\|_2$ in advance, up to the replacement of $\|\boldsymbol{w}^\star\|_2$ by $\|\boldsymbol{w}^\star\|_2 + 1$ inside the logarithm.

Norm-adaptive rates for linear without prior knowledge of $\|\boldsymbol{w}^\star\|_2$ have been studied extensively in the online convex optimization literature, see Chapter 10 in [Orabona, 2019]. However, these results do not lead to (4), as they are designed for convex loss minimization.

Example application: description length. If $\mathcal{F}$ can be equipped with a description length function $\kappa : \mathcal{F} \to \mathbb{N}$ such that $\sum_{f\in\mathcal{F}} 2^{-\kappa(f)} \le 1$ (Kraft inequality, [Cover and Thomas, 1991]), then setting $\pi(f) = 2^{-\kappa(f)}/Z$ where $Z = \sum_{f\in\mathcal{F}} 2^{-\kappa(f)} \le 1$ defines a valid prior satisfying $\sum_{f\in\mathcal{F}} \pi(f) = 1$. Since $Z \le 1$, we have

$$\log(1/\pi(f^\star)) = \kappa(f^\star)\log 2 + \log Z \le \kappa(f^\star)\log 2 \,,$$

so the rate for $f^\star$ is governed by its description length $\kappa(f^\star)$, and the normalization constant Z only improves the bound. This connects our framework to the minimum description length (MDL) literature [Grünwald, 2007].

4 Empirical Bounds

The guarantee of Theorem 2 is distribution-dependent: it bounds the estimation error in terms of the unobservable conditional variance $\mathbb{E}_X[\sigma(X)]$. We now show that the same algorithm produces an empirically verifiable bound that is never much larger than the distribution-dependent one. Concretely, the bound we derive in Theorem 3 controls $\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|]$ in terms of the cumulative training loss and the cumulative squared residuals, both computable at runtime, while still recovering the $\sqrt{\mathbb{E}_X[\sigma(X)]/n}$ rate. This makes the bound usable for downstream tasks such as for example model selection and the construction of contextual bandit algorithms.

To provide these empirical bounds, we need to make additional assumptions: we assume that B is known to the learner and that $\mathcal{F}$ is finite. The result can be found below. The proof is deferred to Appendix B.

Theorem 3. *Suppose that $n \ge 4$, let $\mathcal{A}$ and $\mathcal{B}$ be Squint+L with uniform priors over $\mathcal{F}$, and for any $\delta \in (0, \frac{1}{e}]$ denote by*

$$\Gamma_\delta = 10\log(|\mathcal{F}|(2 + \log(2Bn))) + 10\log(1/\delta) + 2 \,.$$

Algorithm 2: Reduction for General Convex Losses

Input Online Learning Algorithms $\mathcal{A}, \mathcal{B}$

for $t = 1, \dots, n$ **do**

Obtain f_t^A from $\mathcal{A}$ and $p_t^B \in \triangle$ from $\mathcal{B}$.

Send $\ell_t^A(f) = (f(X_t) - Y_t)^\top \Big(\sum_{f' \in \mathcal{F}} p_t^B(f') \nabla \ell(f_t^A(X_t), f'(X_t)) \Big)$ to algorithm $\mathcal{A}$ as the loss in round t.

Send $\ell_t^B(f) = -(f_t^A(X_t) - Y_t)^\top \nabla \ell(f_t^A(X_t), f(X_t))$ to algorithm $\mathcal{B}$ as the loss in round t.

end

Output $\widehat{f} = \frac{1}{n} \sum_{t=1}^n f_t^A$

Then Algorithm 1 guarantees that with probability at least $1 - (1 + \log(n))\delta$*,*

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \le \frac{1}{n} \left(\sum_{t=1}^n \ell_t^A(f_t^A) + 4 \sqrt{\Gamma_\delta \sum_{t=1}^n (f_t^A(X_t) - Y_t)^2 + 4B\Gamma_\delta} \right) = \mathcal{R}_n \,. \quad (5)$$

Furthermore, with probability at least $1 - \delta$*,*

$$\mathcal{R}_n \le \frac{CB\Gamma_\delta}{n} + C\sqrt{\Gamma_\delta \, \mathbb{E}_X[\sigma(X)]/n} \,,$$

for some absolute constant $C > 0$*.*

Theorem 3 provides a bound on $\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|]$ that, unlike Theorem 2, is expressed entirely in terms of quantities observable from the data. Specifically, the right-hand side of (5) involves only $\sum_{t=1}^n \ell_t^A(f_t^A)$, the cumulative loss of the algorithm on the training data, and $\sum_{t=1}^n (f_t^A(X_t) - Y_t)^2$, the cumulative squared residuals, both of which are computable, which potentially leads to improved contextual bandits algorithms.

The second statement of Theorem 3 shows that the empirical bound is never much larger than the distribution-dependent bound of Theorem 2, so nothing is lost by using the empirical bound. More precisely, with high probability the empirical bound implies the same $O(\sqrt{\mathbb{E}_X[\sigma(X)]/n})$ rate as Theorem 2, but it additionally adapts to instances where the observed residuals $\sum_{t=1}^n (f_t^A(X_t) - Y_t)^2$ are smaller than $n \, \mathbb{E}_X[\sigma(X)]$, yielding a tighter guarantee on such favorable instances.

The reason we chose a uniform prior is the following. We could obtain the same empirical bound with $\log(1/\pi(f^\star))$ instead of $\log(|\mathcal{F}|)$ by replacing the uniform prior with an arbitrary prior. However, since we cannot observe $\log(1/\pi(f^\star))$, this would render the bound useless.

5 General loss functions

The reduction can be extended beyond the absolute loss to any loss function $\ell : \mathcal{W} \times \mathcal{W} \to \mathbb{R}$ that is convex in its first argument and satisfies $\ell(f(x), f(x)) = 0$ for all $x \in \mathcal{X}$, where $\mathcal{W} \subseteq \mathbb{R}^d$. As before, we assume that there exists an $f^\star \in \mathcal{F}$ such that $\mathbb{E}[Y|X] = f^\star(X)$, where $Y \in \mathcal{W}$. This class is broad as it includes Bregman divergences generated by convex potentials. The key idea is that for any such loss, convexity together with $\ell(f^\star(x), f^\star(x)) = 0$ yields the linearization

$$\ell(z, f^\star(x)) \le (z - f^\star(x))^\top \nabla \ell(z, f^\star(x)) \,,$$

where $\nabla \ell(z, f^\star(x))$ denotes the subgradient of $\ell(z, f^\star(x))$ with respect to z. This linearization plays exactly the role that the sign decomposition did in the absolute-loss case: it bounds the loss by an inner product between the prediction error and the loss derivative, and the latter is what algorithm $\mathcal{B}$ now estimates. Replacing sign with the gradient $\nabla \ell$ in Algorithm 1 produces Algorithm 2, and the decomposition of Theorem 1 carries through with essentially the same proof. The price for this generality is that the resulting bound (Theorem 5) does not adapt to the conditional variance the way Theorem 2 does. We assume throughout that $\max_{x,y \in \mathcal{W}} \|x - y\| \le B$ and $\max_{x,y \in \mathcal{W}} \|\nabla \ell(x, y)\|_\star \le L$ where $\|\cdot\|$ and $\|\cdot\|_\star$ are a dual norm pair.

Theorem 4. *For any $\mathcal{A}$ and $\mathcal{B}$, Algorithm 2 guarantees that*

$$\mathbb{E}_X[\ell(\widehat{f}(X), f^\star(X))] \le \frac{1}{n}\Big(R_n^{\mathcal{A}}(f^\star) + R_n^{\mathcal{B}}(f^\star) + \sum_{t=1}^{n}\big(\ell_t^A(f^\star) + Z_t\big)\Big),$$

where

$$Z_t = \mathbb{E}_t\left[(f_t^A(X_t) - Y_t)^\top \nabla\ell(f_t^A(X_t), f^\star(X_t))\right] - (f_t^A(X_t) - Y_t)^\top \nabla\ell(f_t^A(X_t), f^\star(X_t)).$$

The proof can be found in Appendix C. It closely mirrors the proof of Theorem 1: convexity plus $\ell(f^\star, f^\star) = 0$ replaces Jensen's inequality applied to the absolute value, the linearization plays the role of the sign decomposition, and the resulting cross terms split cleanly into the regrets of $\mathcal{A}$ and $\mathcal{B}$. Crucially, since $\mathbb{E}_t[\ell_t^A(f^\star)] = 0$ (because $\mathbb{E}_t[Y_t \mid X_t] = f^\star(X_t)$) the comparator term $\sum_{t=1}^n \ell_t^A(f^\star)$ remains a martingale difference sequence, exactly as in the absolute-loss case. Thus the reduction sidesteps irreducible noise for the same reason it did before, and Theorem 4 reduces estimation under any such loss to two online learning problems. Instantiating both algorithms with Squint+L gives the following.

Theorem 5. *Let $\mathcal{A}$ and $\mathcal{B}$ be Squint+L with prior π. For some absolute constant $C > 0$, Algorithm 2 guarantees that with probability at least $1 - \delta$,*

$$\mathbb{E}_X[\ell(\widehat{f}(X), f^\star(X))] \le CBL\sqrt{\frac{\big(\log(1/\delta) + \log\big((1 + \log(BLn))/\pi(f^\star)\big)\big)}{n}}.$$

The proof can be found in Appendix D. A few remarks on the Theorem. First, the conditions of Theorem 5 are satisfied by any Bregman divergence $D_\phi(z, y) = \phi(z) - \phi(y) - \nabla\phi(y)^\top(z - y)$ generated by a convex differentiable potential ϕ, with $\nabla\ell(z, y) = \nabla\phi(z) - \nabla\phi(y)$. The bound on L then reduces to a bound on $\|\nabla\phi\|_\star$ over $\mathcal{W}$. Theorem 5 also applies to non-Bregman losses such as l_p norms and pinball losses, where $\nabla\ell$ is interpreted as a subgradient. Second, the prior-dependent complexity $\log(1/\pi(f^\star))$ is preserved. Hence, all applications of Section 3 extend immediately to any loss in the class above. Third, the rate is $O(BL\sqrt{\big(\log(1/\delta) + \log(\log(nB)/\pi(f^\star))\big)/n})$ rather than the $O(\sqrt{\mathbb{E}_X[\sigma(X)]\big(\log(1/\delta) + \log(\log(nB)/\pi(f^\star))\big)/n})$ rate of Theorem 2: the dependence on the noise level has been replaced by the worst-case range BL. The reason is that the linearization step $\ell(z, f^\star) \le (z - f^\star)\ell'(z, f^\star)$ decouples the loss derivative from the residual $f_t^A(X_t) - Y_t$, and unlike the absolute-loss case where $\ell' = \text{sign}$ takes only two values and is bounded uniformly by 1, a general ℓ' can have nontrivial range L that propagates into the regret of $\mathcal{B}$ without any variance-adaptive cancellation. Recovering second-order rates for general losses appears to require additional structure on ℓ, such as self-bounding properties of $|\ell'|^2$ in terms of ℓ itself, see e.g. Chapter 4 of [Orabona, 2019], which we leave to future work. Another interesting approach we leave for future work are exploiting curvature assumptions on the loss which are known to yield fast rates in risk minimization such as exp-concavity, mixability, or strong convexity. We are hopeful such assumptions will lead to bounds of the form $\mathbb{E}_X[\ell(\widehat{f}(X), f^\star(X))] = O(\frac{1}{n})$.

6 Conclusion

We have presented a simple reduction of realizable regression to online learning via a two-algorithm structure, and shown that instantiating the reduction with Squint+L yields prior-dependent second-order bounds that adapt automatically to the noise level $\mathbb{E}_X[\sigma(X)]$ without any prior knowledge of it. The resulting estimator achieves the fast rate $O(B\Gamma(f^\star)/n)$ in the noiseless case and interpolates smoothly to the slow rate $O(\sqrt{\mathbb{E}_X[\sigma(X)]\Gamma(f^\star)/n})$ as the noise grows, where $\Gamma(f^\star) = \log(1/\pi(f^\star)) + \log(1 + \log(Bn)) + \log(1/\delta)$ depends on the function class only through the prior mass assigned to $f^\star$. We have also shown that the same algorithm yields an empirically verifiable bound expressed entirely in observable quantities.

Several directions for future work suggest themselves. First, exploiting curvature assumptions as suggested in section 5 seems possible to obtain fast rates. Second, the connection to contextual bandits suggests that our bounds could yield improved instance-dependent regret guarantees in that setting, and we leave a detailed investigation of this direction to future work. Another exciting direction can be found in attempting to extend our results to convex or countably infinite $\mathcal{F}$ with a computationally efficient algorithm.

Acknowledgments and Disclosure of Funding

DvdH was supported by Netherlands Organization for Scientific Research (NWO), grant number VI.Veni.242.143 (https://doi.org/10.61686/SCGQQ55380).

References

Pierre Alquier and Karim Lounici. Pac-bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

Jean-Yves Audibert. Progressive mixture rules are deviation suboptimal. *Advances in Neural Information Processing Systems*, 2007.

Peter Bartlett, Varsha Dani, Thomas Hayes, Sham Kakade, Alexander Rakhlin, and Ambuj Tewari. High-probability regret bounds for bandit online linear optimization. In *COLT*, pages 335–342. Omnipress, 2008.

Alina Beygelzimer, John Langford, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandit algorithms with supervised learning guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 19–26, 2011.

Eugenio Clerico, Hamish Flynn, Wojciech Kotlowski, and Gergely Neu. Confidence sequences for generalized linear models via regret analysis. *arXiv preprint arXiv:2504.16555*, 2025.

Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. New York: Wiley Interscience, 1991.

Arnak Dalalyan, Yuri Ingster, and Alexandre B Tsybakov. Statistical inference in compound functional models. *Probability Theory and Related Fields*, 158(3):513–532, 2014.

Arnak S Dalalyan and Joseph Salmon. Sharp oracle inequalities for aggregation of affine estimators. *The Annals of Statistics*, pages 2327–2355, 2012.

Dylan J Foster and Akshay Krishnamurthy. Efficient first-order contextual bandits: Prediction, allocation, and triangular discrimination. *Advances in Neural Information Processing Systems*, 34: 18907–18919, 2021.

Peter D Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Johannes Kirschner, Andreas Krause, Michele Meziu, and Mojmir Mutny. Confidence estimation via sequential likelihood mixing. *arXiv preprint arXiv:2502.14689*, 2025.

Wouter M. Koolen and Tim van Erven. Second-order quantile methods for experts and combinatorial games. In *Conference on Learning Theory*, pages 1155–1175, 2015.

Yinan Li and Kwang-Sung Jun. Second-order bounds for [0, 1]-valued regression via betting loss. *arXiv preprint arXiv:2507.12584*, 2025.

Nick Littlestone and Manfred K Warmuth. The weighted majority algorithm. *Information and Computation*, 108(2):212–261, 1994.

Gábor Lugosi and Gergely Neu. Online-to-pac conversions: Generalization bounds via regret analysis. *arXiv preprint arXiv:2305.19674*, 2023.

Zakaria Mhammedi, Wouter M Koolen, and Tim van Erven. Lipschitz adaptivity with multiple learning rates in online learning. In *Conference on Learning Theory*, pages 2490–2511, 2019.

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

Francesco Orabona and Kwang-Sung Jun. Tight concentrations and confidence sequences from the regret of universal portfolio. *IEEE Transactions on Information Theory*, 70(1):436–455, 2024.

David Simchi-Levi and Yunzong Xu. Bypassing the monster: A faster and simpler optimal algorithm for contextual bandits under realizability. *Mathematics of Operations Research*, 47(3):1904–1931, 2022.

Vladimir Vapnik and Alexey Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2nd edition, 2026.

Volodimir G Vovk. Aggregating strategies. In *Proceedings of the 3rd Annual Workshop on Computational Learning Theory*, 1990.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024.

A Proof of Theorem 2

We first restate the Theorem, after which we prove it.

Theorem 2 (RESTATED). *Let $\mathcal{A}$ and $\mathcal{B}$ be Squint+L with prior π. Algorithm 1 guarantees that with probability at least $1-\delta$, for some absolute constant $C > 0$,*

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \le \frac{CB\,\Gamma(f^\star)}{n} + C\sqrt{\frac{\mathbb{E}_X[\sigma(X)]\,\Gamma(f^\star)}{n}}\,,$$

where $\Gamma(f^\star) = \log(1/\delta) + \log\big((1+\log(Bn))/\pi(f^\star)\big)$.

Proof. Denote $\widetilde{\Gamma}(f^\star) = \log((2+\log(2Bn))/\pi(f^\star)) + \frac{1}{2}$. Since $|\ell_t^A(f)| \le 2B$ and $|\ell_t^B(f)| \le 2B$, Theorem 12 gives

$$\mathcal{R}_n^{\mathcal{A}}(f^\star) \le 2\sqrt{\sum_{t=1}^n (f_t^A(X_t) - f^\star(X_t))^2\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})} + 20B(\widetilde{\Gamma}(f^\star)+2)\,,$$

$$\mathcal{R}_n^{\mathcal{B}}(f^\star) \le 2\sqrt{\sum_{t=1}^n (f_t^A(X_t) - Y_t)^2\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})} + 20B(\widetilde{\Gamma}(f^\star)+2)\,.$$

Since $(a+b)^2 \le 2a^2 + 2b^2$ we have that

$$\begin{aligned}
&2\sqrt{\sum_{t=1}^n (f_t^A(X_t) - Y_t)^2\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})} \\
&\le 2\sqrt{2\sum_{t=1}^n \big((f_t^A(X_t) - f^\star(X_t))^2 + (f^\star(X_t) - Y_t)^2\big)\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})} \\
&\le 2\sqrt{2\sum_{t=1}^n (f^\star(X_t) - Y_t)^2\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})} + \tfrac{1}{16}\sum_{t=1}^n |f_t^A(X_t) - f^\star(X_t)| \\
&\quad + 256B(1+\sqrt{2\widetilde{\Gamma}(f^\star)})^2\,,
\end{aligned}$$

where the last inequality follows from the fact that $ab \le \inf_{\gamma>0} \frac{a^2}{2\gamma} + \frac{\gamma b^2}{2}$ and that $|f_t^A(X_t) - f^\star(X_t)| \le 2B$. Therefore

$$\mathcal{R}_n^{\mathcal{B}}(f^\star) \le 2\sqrt{2\sum_{t=1}^n (f^\star(X_t) - Y_t)^2\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})} + 276B(2+\sqrt{2\widetilde{\Gamma}(f^\star)})^2 \tag{6}$$

$$+ \tfrac{1}{16}\sum_{t=1}^n |f_t^A(X_t) - f^\star(X_t)|\,. \tag{7}$$

Likewise,

$$\mathcal{R}_n^{\mathcal{A}}(f^\star) \le \tfrac{1}{16}\sum_{t=1}^n |f_t^A(X_t) - f^\star(X_t)| + 84B(2+\sqrt{2\widetilde{\Gamma}(f^\star)})^2\,.$$

Thus, we have that

$$\begin{aligned}
\mathcal{R}_n^{\mathcal{A}}(f^\star) + \mathcal{R}_n^{\mathcal{B}}(f^\star) &\le 2\sqrt{2\sum_{t=1}^n (f^\star(X_t) - Y_t)^2\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})} + 360B(2+\sqrt{2\widetilde{\Gamma}(f^\star)})^2 \\
&\quad + \frac{1}{8}\sum_{t=1}^n |f_t^A(X_t) - f^\star(X_t)|\,.
\end{aligned}$$

Now, by Lemma 10 we have that with probability at least $1-\delta$

$$\sum_{t=1}^{n}|f_t^A(X_t)-f^\star(X_t)| \le 2\sum_{t=1}^{n}\mathbb{E}_t[|f_t^A(X_t)-f^\star(X_t)|]+4B\log(1/\delta)\,.$$

By Lemma 11 we have that with probability at least $1-\delta$

$$\sum_{t=1}^{n}(f^\star(X_t)-Y_t)^2 \le 2n\,\mathbb{E}[\sigma(X)]+4B^2\log(1/\delta)\,.$$

Thus, with probability at least $1-2\delta$

$$\begin{aligned}
&\mathcal{R}_n^{\mathcal{A}}(f^\star)+\mathcal{R}_n^{\mathcal{B}}(f^\star)\\
&\le \sqrt{16n\,\mathbb{E}[\sigma(X)]+32B^2\log(1/\delta)}\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})+360B(2+\sqrt{2\widetilde{\Gamma}(f^\star)})^2+4B\log(1/\delta)\\
&+\frac{1}{4}\sum_{t=1}^{n}\mathbb{E}_t[|f_t^A(X_t)-f^\star(X_t)|]\,,
\end{aligned}$$

where in the last inequality we used $\sqrt{a+b}\le\sqrt{a}+\sqrt{b}$.

By Lemma 9, with probability at least $1-\delta$

$$\sum_{t=1}^{n}\ell_t^A(f^\star)+Z_t \le 2\sqrt{\log(1/\delta)n6\,\mathbb{E}_X[\sigma(X)]}+\tfrac{1}{4}\sum_{t=1}^{n}\mathbb{E}_t\,|f_t^A(X_t)-f^\star(X_t)|]+64B\log(1/\delta)\,.$$

Thus, with probability at least $1-3\delta$,

$$\begin{aligned}
&\mathcal{R}_n^{\mathcal{A}}(f^\star)+\mathcal{R}_n^{\mathcal{B}}(f^\star)+\sum_{t=1}^{n}\ell_t^A(f^\star)+Z_t\\
&\le \tfrac{1}{2}\sum_{t=1}^{n}\mathbb{E}_t[|f_t^A(X_t)-f^\star(X_t)|]+\sqrt{16n\,\mathbb{E}[\sigma(X)]}\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})+360B(2+\sqrt{2\widetilde{\Gamma}(f^\star)})^2\\
&+4B\log(1/\delta)+8B\sqrt{\log(1/\delta)}\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})+2\sqrt{\log(1/\delta)n6\,\mathbb{E}_X[\sigma(X)]}+64B\log(1/\delta)\,.
\end{aligned}$$

Recall that $Z_t=\mathbb{E}_t[\big(f_t^A(X_t)-Y_t\big)\mathrm{sign}(f_t^A(X_t)-f^\star(X_t))]-\big(f_t^A(X_t)-Y_t\big)\mathrm{sign}(f_t^A(X_t)-f^\star(X_t))$. We have shown that with probability at least $1-3\delta$

$$\begin{aligned}
&\sum_{t=1}^{n}\mathbb{E}_t[|f_t^A(X_t)-f^\star(X_t)|]=\mathcal{R}_n^{\mathcal{A}}(f^\star)+\mathcal{R}_n^{\mathcal{B}}(f^\star)+\sum_{t=1}^{n}\ell_t^A(f^\star)+Z_t \qquad \text{(See also (3))}\\
&\le \tfrac{1}{2}\sum_{t=1}^{n}\mathbb{E}_t[|f_t^A(X_t)-f^\star(X_t)|]+\sqrt{16n\,\mathbb{E}[\sigma(X)]}\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})+360B(2+\sqrt{2\widetilde{\Gamma}(f^\star)})^2\\
&+4B\log(1/\delta)+8B\sqrt{\log(1/\delta)}\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})+2\sqrt{\log(1/\delta)n6\,\mathbb{E}_X[\sigma(X)]}+64B\log(1/\delta)\,.
\end{aligned}$$

By subtracting $\frac{1}{2}\sum_{t=1}^{n}\mathbb{E}_t[|f_t^A(X_t)-f^\star(X_t)|]$ from both sides and multiplying by 2 we may conclude that with probability at least $1-3\delta$

$$\begin{aligned}
&\sum_{t=1}^{n}\mathbb{E}_t[|f_t^A(X_t)-f^\star(X_t)|]=\mathcal{R}_n^{\mathcal{A}}(f^\star)+\mathcal{R}_n^{\mathcal{B}}(f^\star)+\sum_{t=1}^{n}\ell_t^A(f^\star)+Z_t\\
&\le 2\sqrt{16n\,\mathbb{E}[\sigma(X)]}\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})+720B(2+\sqrt{2\widetilde{\Gamma}(f^\star)})^2\\
&+8B\log(1/\delta)+16B\sqrt{\log(1/\delta)}\,(1+\sqrt{2\widetilde{\Gamma}(f^\star)})+4\sqrt{\log(1/\delta)n6\,\mathbb{E}_X[\sigma(X)]}+128B\log(1/\delta)\\
&\le \frac{CB\,\Gamma(f^\star)}{n}+C\sqrt{\frac{\mathbb{E}_X[\sigma(X)]\,\Gamma(f^\star)}{n}}\,, \qquad (8)
\end{aligned}$$

where the last inequality follows from choosing a sufficiently large C. Thus, by Lemma 1, for some absolute constant C we have that with probability at least $1-\delta$,

$$\mathbb{E}_X[|\widehat{f}(X)-f^\star(X)|]\le\frac{CB\,\Gamma(f^\star)}{n}+C\sqrt{\frac{\mathbb{E}_X[\sigma(X)]\,\Gamma(f^\star)}{n}}\,,$$

where $\Gamma(f^\star)=\log(1/\delta)+\log\big((1+\log(Bn))/\pi(f^\star)\big)$. □

B Proof of Theorem 3

We first restate the Theorem, after which we prove it.

Theorem 3 (RESTATED). *Suppose that $n \geq 4$, let $\mathcal{A}$ and $\mathcal{B}$ be Squint+L with uniform priors over $\mathcal{F}$, and for any $\delta \in (0, \frac{1}{e}]$ denote by*

$$\Gamma_\delta = 10 \log(|\mathcal{F}|(2 + \log(2Bn))) + 10 \log(1/\delta) + 2 \,.$$

Then Algorithm 1 guarantees that with probability at least $1 - (1 + \log(n))\delta$,

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \leq \frac{1}{n}\left(\sum_{t=1}^n \ell_t^A(f_t^A) + 4\sqrt{\Gamma_\delta \sum_{t=1}^n (f_t^A(X_t) - Y_t)^2} + 4B\Gamma_\delta\right) = \mathcal{R}_n \,. \tag{5}$$

Furthermore, with probability at least $1 - \delta$,

$$\mathcal{R}_n \leq \frac{CB\Gamma_\delta}{n} + C\sqrt{\Gamma_\delta \, \mathbb{E}_X[\sigma(X)]/n} \,,$$

for some absolute constant $C > 0$.

Proof. First, recalling the definition of $\mathcal{R}_n^{\mathcal{A}}$, by Lemma 1 we have that

$$\mathbb{E}_X[|\widehat{f}(X) - f^\star(X)|] \leq \frac{1}{n}\left(\sum_{t=1}^n \ell_t^A(f_t^A) + \mathcal{R}_n^{\mathcal{B}} + \sum_{t=1}^n Z_t\right) \tag{9}$$

Since $|\ell_t^B(f)| \leq 2B$, Theorem 12 gives

$$\mathcal{R}_n^{\mathcal{B}}(f^\star) \leq 2\sqrt{\sum_{t=1}^n (f_t^A(X_t) - Y_t)^2}\,(1 + \sqrt{2\Gamma}) + 20B(\Gamma + 2) \,, \tag{10}$$

where $\Gamma = \log(|\mathcal{F}|(2 + \log(2Bn))) + \frac{1}{2}$.

At this point it suffices to bound $\sum_{t=1}^n Z_t$ in terms of observable quantities. First, by Lemma 7, with probability at least $1 - \log(n)\delta$,

$$\sum_{t=1}^n Z_t \leq 4\sqrt{\log(1/\delta) \sum_{t=1}^n \mathbb{E}_t[(f_t^A(X_t) - Y_t)^2]} + 12B \log(1/\delta) \,,$$

where we used that $\mathbb{E}[(X - \mathbb{E}[X])^2] \leq \mathbb{E}[X^2]$. Since $(f_t^A(X_t) - Y_t)^2 \leq 4B^2$, by Lemma 6 with $\lambda = \frac{1}{8B^2}$ we have that with probability at least $1 - \delta$

$$\sum_{t=1}^n \mathbb{E}_t[(f_t^A(X_t) - Y_t)^2] \leq \sum_{t=1}^n (f_t^A(X_t) - Y_t)^2 + \tfrac{1}{2}\sum_{t=1}^n \mathbb{E}_t[(f_t^A(X_t) - Y_t)^2] + 8B^2 \log(1/\delta) \,,$$

which implies that

$$\sum_{t=1}^n \mathbb{E}_t[(f_t^A(X_t) - Y_t)^2] \leq 2\sum_{t=1}^n (f_t^A(X_t) - Y_t)^2 + 16B^2 \log(1/\delta) \tag{11}$$

Therefore, with probability at least $1 - (1 + \log(n))\delta$.

$$\sum_{t=1}^n Z_t \leq 4\sqrt{2\log(1/\delta) \sum_{t=1}^n (f_t^A(X_t) - Y_t)^2} + 28B \log(1/\delta) \,, \tag{12}$$

where we used that $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for $a, b \geq 0$.

Thus, by substituting (10) and (12) into (9) we can see that with probability at least $1-(1+\log(n))\delta$

$$\mathbb{E}_X[|\widehat{f}(X)-f^\star(X)|] \le \frac{1}{n}\left(\sum_{t=1}^n \ell_t^A(f_t^A) + 2\sqrt{\sum_{t=1}^n (f_t^A(X_t)-Y_t)^2}\,(1+\sqrt{2\Gamma}+\sqrt{8\log(1/\delta)}) \right.$$
$$\left. + 28B(\Gamma+2+\log(1/\delta))\right)$$
$$\le \frac{1}{n}\left(\sum_{t=1}^n \ell_t^A(f_t^A) + 4\sqrt{\Gamma_\delta \sum_{t=1}^n (f_t^A(X_t)-Y_t)^2 + 4B\Gamma_\delta}\right)$$

where in the last inequality we used the definition of Γ_δ and $\sqrt{a}+\sqrt{b}\le 2\sqrt{a+b}$ for $a,b\ge 0$. This completes the proof of the first statement.

We proceed with the proof of the second statement. Since $\sum_{t=1}^n \mathbb{E}_t[\ell_t^A(f^\star)]=0$ and $|\ell_t^A(f^\star)|\le 2B$, by using Lemma 6 with $\lambda=\min\{\frac{1}{2B},\sqrt{\frac{\log(1/\delta)}{n\,\mathbb{E}_X[\sigma(X)]}}\}$, we have that with probability at least $1-\delta$

$$\sum_{t=1}^n \ell_t^A(f^\star) \le 2\sqrt{n\,\mathbb{E}_X[\sigma(X)]\log(1/\delta)} + 2B\log(1/\delta)\,.$$

Thus with probability at least $1-\delta$, by using the guarantee that squint+L gives us (Theorem 12) we can see that

$$\sum_{t=1}^n \ell_t^A(f_t^A) = \sum_{t=1}^n \ell_t^A(f^\star) + \mathcal{R}_n^{\mathcal{A}}$$
$$\le \sum_{t=1}^n \ell_t^A(f^\star) + 2\sqrt{\sum_{t=1}^n (f_t^A(X_t)-f^\star(X_t))^2}\,(1+\sqrt{2\Gamma}) + 20B(\Gamma+2)$$
$$\le 2\sqrt{n\,\mathbb{E}_X[\sigma(X)]\log(1/\delta)} + 2B\log(1/\delta) + 2\sqrt{\sum_{t=1}^n (f_t^A(X_t)-f^\star(X_t))^2}\,(1+\sqrt{2\Gamma}) + 20B(\Gamma+2)$$
$$\le 2\sqrt{n\,\mathbb{E}_X[\sigma(X)]\log(1/\delta)} + 2B\log(1/\delta) + \frac{1}{2}\sum_{t=1}^n |f_t^A(X_t)-f^\star(X_t)| + 4B(1+\sqrt{2\Gamma})^2 + 20B(\Gamma+2)\,,$$

where the last inequality follows from $ab\le \frac{a^2}{2\gamma}+\frac{\gamma b^2}{2}$ for $\gamma>0$ and $|f_t^A(X_t)-f^\star(X_t)|\le 2B$.

Next, since $(a+b)^2\le 2a^2+2b^2$, we have that

$$4\sqrt{\Gamma_\delta\sum_{t=1}^n (f_t^A(X_t)-Y_t)^2} \le 4\sqrt{2\Gamma_\delta\sum_{t=1}^n \left((f^\star(X_t)-Y_t)^2+(f_t^A(X_t)-f^\star(X_t))^2\right)}$$
$$\le 4\sqrt{2\Gamma_\delta\sum_{t=1}^n (f^\star(X_t)-Y_t)^2} + 4\sqrt{2\Gamma_\delta\sum_{t=1}^n (f_t^A(X_t)-f^\star(X_t))^2}$$
$$\le 4\sqrt{2\Gamma_\delta\sum_{t=1}^n (f^\star(X_t)-Y_t)^2} + 32B\Gamma_\delta + \frac{1}{2}\sum_{t=1}^n |f_t^A(X_t)-f^\star(X_t)|\,,$$

where the last inequality follows from $ab\le \frac{a^2}{2\gamma}+\frac{\gamma b^2}{2}$ for $\gamma>0$ and $|f_t^A(X_t)-f^\star(X_t)|\le 2B$. Now, by Lemma 10 we have that with probability at least $1-\delta$

$$\sum_{t=1}^n |f_t^A(X_t)-f^\star(X_t)| \le 2\sum_{t=1}^n \mathbb{E}_t[|f_t^A(X_t)-f^\star(X_t)|] + 4B\log(1/\delta)\,.$$

By Lemma 11 we have that with probability at least $1-\delta$

$$\sum_{t=1}^{n}(f^\star(X_t)-Y_t)^2 \le 2n\,\mathbb{E}[\sigma(X)] + 4B^2\log(1/\delta)\,.$$

Thus, with probability at least $1-3\delta$

$$\begin{aligned}
&\sum_{t=1}^{n}\ell_t^A(f_t^A) + 4\sqrt{\Gamma_\delta\sum_{t=1}^{n}(f_t^A(X_t)-Y_t)^2} \\
&\le 2\sqrt{n\,\mathbb{E}_X[\sigma(X)]\log(1/\delta)} + 2B\log(1/\delta) + \sum_{t=1}^{n}|f_t^A(X_t)-f^\star(X_t)| + 6B(1+\sqrt{2\Gamma})^2 \\
&\quad + 20B(\Gamma+2) + 4\sqrt{2\Gamma_\delta\sum_{t=1}^{n}(f^\star(X_t)-Y_t)^2} + 32B\Gamma_\delta \\
&\le 2\sqrt{n\,\mathbb{E}_X[\sigma(X)]\log(1/\delta)} + 6B\log(1/\delta) + 2\sum_{t=1}^{n}\mathbb{E}_t[|f_t^A(X_t)-f^\star(X_t)|] + 6B(1+\sqrt{2\Gamma})^2 \\
&\quad + 20B(\Gamma+2) + 4\sqrt{4\Gamma_\delta n\,\mathbb{E}[\sigma(X)]} + 4B\sqrt{8\Gamma_\delta\log(1/\delta)} + 32B\Gamma_\delta \\
&\le 2\sqrt{n\,\mathbb{E}_X[\sigma(X)]\log(1/\delta)} + 22B\log(1/\delta) + 2\sum_{t=1}^{n}\mathbb{E}_t[|f_t^A(X_t)-f^\star(X_t)|] + 6B(1+\sqrt{2\Gamma})^2 \\
&\quad + 20B(\Gamma+2) + 4\sqrt{4\Gamma_\delta n\,\mathbb{E}[\sigma(X)]} + 48B\Gamma_\delta && (ab\le \tfrac{1}{2}a^2+\tfrac{1}{2}b^2) \\
&\le 2\sqrt{n\,\mathbb{E}_X[\sigma(X)](4\log(1/\delta)+64\Gamma_\delta)} + 22B\log(1/\delta) + 2\sum_{t=1}^{n}\mathbb{E}_t[|f_t^A(X_t)-f^\star(X_t)|] \\
&\quad + 6B(1+\sqrt{2\Gamma})^2 + 20B(\Gamma+2) + 48B\Gamma_\delta && (\sqrt{a}+\sqrt{b}\le 2\sqrt{a+b}) \\
&\le C'\sqrt{\Gamma_\delta n\,\mathbb{E}_X[\sigma(X)]} + C'B\Gamma_\delta + 2\sum_{t=1}^{n}\mathbb{E}_t[|f_t^A(X_t)-f^\star(X_t)|]\,,
\end{aligned}$$

where the last inequality follows from taking a sufficiently large C' and the definition of Γ_δ. To complete the proof of the second statement it suffices to use (8). □

C Proof of Theorem 4

We first restate the Theorem, after which we prove it.

Theorem 4 (RESTATED). *For any $\mathcal{A}$ and $\mathcal{B}$, Algorithm 2 guarantees that*

$$\mathbb{E}_X[\ell(\widehat{f}(X),f^\star(X))] \le \frac{1}{n}\Big(R_n^{\mathcal{A}}(f^\star)+R_n^{\mathcal{B}}(f^\star)+\sum_{t=1}^{n}\left(\ell_t^A(f^\star)+Z_t\right)\Big)\,,$$

where

$$Z_t = \mathbb{E}_t\left[(f_t^A(X_t)-Y_t)^\top\nabla\ell(f_t^A(X_t),f^\star(X_t))\right] - (f_t^A(X_t)-Y_t)^\top\nabla\ell(f_t^A(X_t),f^\star(X_t))\,.$$

Proof. By convexity of ℓ in its first argument and Jensen's inequality,

$$\begin{aligned}
\mathbb{E}_X[\ell(\widehat{f}(X),f^\star(X))] &\le \mathbb{E}_X\left[\frac{1}{n}\sum_{t=1}^{n}\ell(f_t^A(X),f^\star(X))\right] \\
&= \frac{1}{n}\sum_{t=1}^{n}\mathbb{E}_X[\ell(f_t^A(X),f^\star(X))]\,.
\end{aligned}$$

Next, since $\ell(\cdot, f^\star(X))$ is convex and $\ell(f^\star(X), f^\star(X)) = 0$, by convexity we have $\ell(z, f^\star(X)) \le (z - f^\star(X))^\top \nabla \ell(z, f^\star(X))$ for all z. Applying this with $z = f_t^A(X)$ gives

$$\begin{aligned}\mathbb{E}_X[\ell(f_t^A(X), f^\star(X))] &\le \mathbb{E}_X\left[(f_t^A(X) - f^\star(X))^\top \nabla \ell(f_t^A(X), f^\star(X))\right] \\ &= \mathbb{E}_{X,Y}\left[(f_t^A(X) - Y)^\top \nabla \ell(f_t^A(X), f^\star(X))\right],\end{aligned}$$

where the last equality uses $\mathbb{E}[Y \mid X] = f^\star(X)$. Summing over t and converting the expectation over (X, Y) to the observed (X_t, Y_t) via the martingale difference sequence $\{Z_t\}$, we obtain

$$\begin{aligned}&\frac{1}{n}\sum_{t=1}^n \mathbb{E}_t\left[(f_t^A(X_t) - Y_t)^\top \nabla \ell(f_t^A(X_t), f^\star(X_t))\right] \\ &= \frac{1}{n}\sum_{t=1}^n \left((f_t^A(X_t) - Y_t)^\top \nabla \ell(f_t^A(X_t), f^\star(X_t)) + Z_t\right).\end{aligned}$$

It remains to relate $(f_t^A(X_t) - Y_t)^\top \nabla \ell(f_t^A(X_t), f^\star(X_t))$ to the regret terms. We have

$$\begin{aligned}&\sum_{t=1}^n (f_t^A(X_t) - Y_t)^\top \nabla \ell(f_t^A(X_t), f^\star(X_t)) \\ &= \sum_{t=1}^n (f_t^A(X_t) - Y_t)^\top \Big(\sum_{f\in\mathcal{F}} p_t^B(f) \nabla \ell(f_t^A(X_t), f(X_t))\Big) \\ &\quad + \sum_{t=1}^n (f_t^A(X_t) - Y_t)^\top \Big(\nabla \ell(f_t^A(X_t), f^\star(X_t)) - \sum_{f\in\mathcal{F}} p_t^B(f) \nabla \ell(f_t^A(X_t), f(X_t))\Big) \\ &= \sum_{t=1}^n \ell_t^A(f_t^A) + \sum_{t=1}^n \Big(\sum_{f\in\mathcal{F}} p_t^B(f)\ell_t^B(f) - \ell_t^B(f^\star)\Big) \\ &= \sum_{t=1}^n \ell_t^A(f^\star) + R_n^{\mathcal{A}}(f^\star) + R_n^{\mathcal{B}}(f^\star),\end{aligned}$$

where in the third equality we identified $\ell_t^A(f_t^A) = (f_t^A(X_t) - Y_t)^\top \Big(\sum_{f\in\mathcal{F}} p_t^B(f)\nabla\ell(f_t^A(X_t), f(X_t))\Big)$ and used

$$\begin{aligned}&(f_t^A(X_t) - Y_t)^\top \Big(\nabla \ell(f_t^A(X_t), f^\star(X_t)) - \sum_{f\in\mathcal{F}} p_t^B(f) \nabla \ell(f_t^A(X_t), f(X_t))\Big) \\ &= \sum_{f\in\mathcal{F}} p_t^B(f)\ell_t^B(f) - \ell_t^B(f^\star),\end{aligned}$$

and in the last equality the definitions of $R_n^{\mathcal{A}}(f^\star)$ and $R_n^{\mathcal{B}}(f^\star)$. Combining everything concludes the proof. □

D Proof of Theorem 5

We first restate the Theorem, after which we prove it.

Theorem 5 (RESTATED). *Let $\mathcal{A}$ and $\mathcal{B}$ be Squint+L with prior π. For some absolute constant $C > 0$, Algorithm 2 guarantees that with probability at least $1 - \delta$,*

$$\mathbb{E}_X[\ell(\widehat{f}(X), f^\star(X))] \le CBL\sqrt{\frac{\left(\log(1/\delta) + \log\left((1 + \log(BLn))/\pi(f^\star)\right)\right)}{n}}.$$

Proof. Denote by $\Gamma(f^\star) = \log((2 + \log(2BLn))/\pi(f^\star)) + 1$. By Hölder's inequality and the assumptions on $\nabla\ell$ and $\mathcal{W}$,

$$\begin{aligned}|\ell_t^A(f)| &\le \|f(X_t) - Y_t\| \Big\| \sum_{f'\in\mathcal{F}} p_t^B(f') \nabla \ell(f_t^A(X_t), f'(X_t))\Big\|_\star \\ &\le \|f(X_t) - Y_t\| \sum_{f'\in\mathcal{F}} p_t^B(f') \|\nabla \ell(f_t^A(X_t), f'(X_t))\|_\star \le BL,\end{aligned}$$

where the second inequality uses the triangle inequality for $\|\cdot\|_\star$. Likewise,

$$|\ell_t^B(f)| \le \|f_t^A(X_t) - Y_t\| \, \|\nabla \ell(f_t^A(X_t), f(X_t))\|_\star \le BL\,,$$

for all $f \in \mathcal{F}$ and $t = 1, \dots, n$. Theorem 12 with range parameter $2BL$ gives

$$\mathcal{R}_n^{\mathcal{A}}(f^\star) \le 2\sqrt{\sum_{t=1}^n (\ell_t^A(f_t^A) - \ell_t^A(f^\star))^2 \, (1 + \sqrt{2\Gamma(f^\star)})} + 20BL(\Gamma(f^\star) + 2)\,.$$

Since $|\ell_t^A(f_t^A) - \ell_t^A(f^\star)| \le 2BL$, we have $\sum_{t=1}^n (\ell_t^A(f_t^A) - \ell_t^A(f^\star))^2 \le 4B^2L^2n$, and thus

$$\mathcal{R}_n^{\mathcal{A}}(f^\star) \le 4BL\sqrt{n}\,(1 + \sqrt{2\Gamma(f^\star)}) + 20BL(\Gamma(f^\star) + 2) \le C_0\, BL\sqrt{n\Gamma(f^\star)} \tag{13}$$

for some absolute constant $C_0 > 0$, where the last inequality uses $1 + \sqrt{\Gamma(f^\star)} \le 2\sqrt{\Gamma(f^\star)}$ for $\Gamma(f^\star) \ge 1$. The identical bound holds for $\mathcal{R}_n^{\mathcal{B}}(f^\star)$.

Since $\mathbb{E}_t[\ell_t^A(f^\star) + Z_t] = 0$ and, by Hölder's inequality,

$$\begin{aligned}|\ell_t^A(f^\star) + Z_t| &\le |\ell_t^A(f^\star)| + |(f_t^A(X_t) - Y_t)^\top \nabla \ell(f_t^A(X_t), f^\star(X_t))| \\ &\quad + |\,\mathbb{E}_t[(f_t^A(X_t) - Y_t)^\top \nabla \ell(f_t^A(X_t), f^\star(X_t))]| \le 3BL\,,\end{aligned}$$

by Lemma 8 with $R = 3BL$, with probability at least $1 - \delta$,

$$\sum_{t=1}^n (\ell_t^A(f^\star) + Z_t) \le 3BL\sqrt{2n\log(1/\delta)}\,. \tag{14}$$

By Theorem 4, (13), and (14), with probability at least $1 - \delta$,

$$\begin{aligned}\mathbb{E}_X[\ell(\widehat{f}(X), f^\star(X))] &\le \frac{1}{n}\Big(\mathcal{R}_n^{\mathcal{A}}(f^\star) + \mathcal{R}_n^{\mathcal{B}}(f^\star) + \sum_{t=1}^n (\ell_t^A(f^\star) + Z_t)\Big) \\ &\le \frac{2C_0\, BL\sqrt{n\Gamma(f^\star)} + 3BL\sqrt{2n\log(1/\delta)}}{n} \\ &\le CBL\sqrt{\frac{(\Gamma(f^\star) + \log(1/\delta))}{n}}\end{aligned}$$

for some absolute constant $C > 0$, where the last inequality combines the two square-root terms using $\sqrt{a} + \sqrt{b} \le \sqrt{2}\sqrt{a+b}$. Using $\Gamma(f^\star) = O(\log((1 + \log(BLn))/\pi(f^\star)))$ completes the proof. □

E Concentration Inequalities

We need the following concentration inequality for martingales whose proof can be found in [Beygelzimer et al., 2011, Theorem 1].

Lemma 6. *Let $X_1, \dots, X_T$ be a martingale difference sequence adapted to a filtration $(\mathcal{F}_i)_{i \le T}$. That is, in particular, $\mathbb{E}_t[X_t] = 0$. Suppose that $|X_t| \le R$ almost surely. Then for any $\delta \in (0,1)$, $\lambda \in [0, 1/R]$, with probability at least $1 - \delta$, it holds that*

$$\sum_{t=1}^n X_t \le \lambda(e-2)\sum_{t=1}^n \mathbb{E}_t[X_t^2] + \frac{\log(1/\delta)}{\lambda}\,.$$

We also need the following concentration inequality due to Bartlett et al. [2008].

Lemma 7. *Suppose $X_1, \dots, X_T$ is a martingale difference sequence with $|X_t| \le b$. Let*

$$\mathrm{Var}_t X_t = \mathrm{Var}(X_t \mid X_1, \dots, X_{t-1}),$$

and let $V = \sum_{t=1}^T \mathrm{Var}_t X_t$ be the sum of conditional variances. Then, for any $\delta < 1/e$ and $T \ge 4$,

$$\Pr\left(\sum_{t=1}^T X_t > 2\max\Big\{2\sqrt{V},\, b\sqrt{\log(1/\delta)}\Big\}\sqrt{\log(1/\delta)}\right) \le \log(T)\,\delta.$$

Lemma 8 (Azuma-Hoeffding inequality)**.** *Let $X_1, \dots, X_n$ be a martingale difference sequence adapted to a filtration $(\mathcal{F}_t)_{t\le n}$, i.e. $\mathbb{E}_t[X_t] = 0$. Suppose that $|X_t| \le R$ almost surely. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sum_{t=1}^{n} X_t \le R\sqrt{2n\log(1/\delta)}\,.$$

Lemma 9. *With probability at least $1-\delta$,*

$$\sum_{t=1}^{n} \ell_t^A(f^\star) + Z_t \le 2\sqrt{6\log(1/\delta) n\, \mathbb{E}_X[\sigma(X)]} + \tfrac{1}{4}\sum_{t=1}^{n} \mathbb{E}_t\left|f_t^A(X_t) - f^\star(X_t)\right| + 64B\log(1/\delta)\,.$$

Proof. By assumption on $f(X)$ and Y we have that

$$\ell_t^A(f^\star) + Z_t \le 6B$$

and

$$\begin{aligned}\mathbb{E}_t[\left(\ell_t^A(f^\star) + Z_t\right)^2] &\le \mathbb{E}_t[2(f^\star(X_t) - Y_t)^2 + 2(f_t^A(X_t) - Y_t)^2] \\ &\le \mathbb{E}_t[6(f^\star(X_t) - Y_t)^2 + 4(f_t^A(X_t) - f^\star(X_t))^2]\,,\end{aligned}$$

where the last inequality is due to $(a-b)^2 \le 2(a-c)^2 + 2(c-b)^2$. Thus, by Lemma 6 we have that for any $\lambda \in [0, \frac{1}{6B}]$

$$\begin{aligned}&\sum_{t=1}^{n}\left(\ell_t^A(f^\star) + Z_t\right) \le \lambda\sum_{t=1}^{n}\mathbb{E}_t[6(f^\star(X_t) - Y_t)^2 + 4(f_t^A(X_t) - f^\star(X_t))^2] + \frac{1}{\lambda}\log(1/\delta) \\ &\le \lambda\sum_{t=1}^{n}\mathbb{E}_t[6(f^\star(X_t) - Y_t)^2 + 16B|f_t^A(X_t) - f^\star(X_t)|] + \frac{1}{\lambda}\log(1/\delta) \\ &\le 2\sqrt{\log(1/\delta)\sum_{t=1}^{n}\mathbb{E}_t[6(f^\star(X_t) - Y_t)^2]} + \sum_{t=1}^{n}\tfrac{1}{4}\,\mathbb{E}_t\left|f_t^A(X_t) - f^\star(X_t)\right| + 64B\log(1/\delta)\,,\end{aligned}$$

where the last inequality follows from picking $\lambda = \min\{\frac{1}{64B}, \sqrt{\frac{\log(1/\delta)}{\sum_{t=1}^{n}\mathbb{E}_t[6(f^\star(X_t)-Y_t)^2]}}\}$.

□

Lemma 10. *We have that with probability at least $1-\delta$*

$$\sum_{t=1}^{n}|f_t^A(X_t) - f^\star(X_t)| \le 2\sum_{t=1}^{n}\mathbb{E}_t[|f_t^A(X_t) - f^\star(X_t)|] + 4B\log(1/\delta)\,.$$

Proof. The proof follows from using that $|f_t^A(X_t) - f^\star(X_t)| \le 2B$ and Lemma 6 with $\lambda = \frac{1}{2B}$. □

Lemma 11. *With probability at least $1-\delta$*

$$\sum_{t=1}^{n}(f^\star(X_t) - Y_t)^2 \le 2n\,\mathbb{E}_X[\sigma(X)] + 4B^2\log(1/\delta)\,.$$

Proof. Since $(f^\star(X_t) - Y_t)^2 \le 4B^2$, by Lemma 6 with $\lambda = \frac{1}{4B^2}$, with probability at least $1-\delta$

$$\sum_{t=1}^{n}(f^\star(X_t) - Y_t)^2 \le 2\sum_{t=1}^{n}\mathbb{E}_t[(f^\star(X_t) - Y_t)^2] + 4B^2\log(1/\delta)\,.$$

□

F Squint+L and Its Regret Guarantee

Squint+L, introduced by Mhammedi et al. [2019], is an online learning algorithm for prediction with expert advice that achieves second-order regret bounds — that is, bounds that scale with the cumulative squared losses of the comparator rather than the worst-case range of the losses. It extends the Squint algorithm of Koolen and Van Erven [2015] to handle losses with unknown loss ranges.

In our setting, at each round t, algorithm $\mathcal{A}$ receives the loss function $\ell_t^A : \mathcal{F} \to \mathbb{R}$ defined by

$$\ell_t^A(f) = (f(X_t) - Y_t) \sum_{f' \in \mathcal{F}} p_t^B(f') \text{sign}(f_t^A(X_t) - f'(X_t)) \,,$$

and algorithm $\mathcal{B}$ receives the loss function $\ell_t^B : \mathcal{F} \to \mathbb{R}$ defined by

$$\ell_t^B(f) = -(f_t^A(X_t) - Y_t) \text{sign}(f_t^A(X_t) - f(X_t)) \,.$$

Since $|f(x)| \le B$ and $|Y_t| \le B$ for all $f \in \mathcal{F}$ and t, both loss functions satisfy $|\ell_t^A(f)| \le 2B$ and $|\ell_t^B(f)| \le 2B$ for all $f \in \mathcal{F}$ and $t = 1, \dots, n$. The following theorem records the regret guarantee of Squint+L in this setting.

Theorem 12 (Squint+L regret bound, Mhammedi et al. [2019, Theorem 3])**.** *Let $\mathcal{A}$ and $\mathcal{B}$ be instances of Squint+L with prior π over $\mathcal{F}$ and range parameter $2B$. For any $f^\star \in \mathcal{F}$, define*

$$\Gamma(f^\star) = \log\left(\frac{2 + \log(2Bn)}{\pi(f^\star)}\right) + \frac{1}{2} \,.$$

Then $\mathcal{A}$ guarantees

$$\mathcal{R}_n^{\mathcal{A}}(f^\star) \le 2\sqrt{\sum_{t=1}^n (\ell_t^A(f_t^A) - \ell_t^A(f^\star))^2 \,(1 + \sqrt{2\Gamma(f^\star)})} + 20B(\Gamma(f^\star) + 2) \,,$$

and $\mathcal{B}$ guarantees

$$\mathcal{R}_n^{\mathcal{B}}(f^\star) \le 2\sqrt{\sum_{t=1}^n (\ell_t^B(f^\star))^2 \,(1 + \sqrt{2\Gamma(f^\star)})} + 20B(\Gamma(f^\star) + 2) \,.$$

In particular, since $|\ell_t^A(f_t^A) - \ell_t^A(f^\star)| \le |f_t^A(X_t) - f^\star(X_t)|$ and $|\ell_t^B(f^\star)| \le |f_t^A(X_t) - Y_t|$, these simplify to

$$\mathcal{R}_n^{\mathcal{A}}(f^\star) \le 2\sqrt{\sum_{t=1}^n (f_t^A(X_t) - f^\star(X_t))^2 \,(1 + \sqrt{2\Gamma(f^\star)})} + 20B(\Gamma(f^\star) + 2) \,,$$

$$\mathcal{R}_n^{\mathcal{B}}(f^\star) \le 2\sqrt{\sum_{t=1}^n (f_t^A(X_t) - Y_t)^2 \,(1 + \sqrt{2\Gamma(f^\star)})} + 20B(\Gamma(f^\star) + 2) \,.$$

The second-order nature of the Squint+L bound is what enables the variance-dependent rates in Theorems 2 and 3. In particular, the term $\sqrt{\sum_{t=1}^n (f_t^A(X_t) - Y_t)^2}$ in the regret bound of $\mathcal{B}$ is small when the noise $f^\star(X_t) - Y_t$ is small, even if n is large, which ultimately leads to faster rates when $\mathbb{E}_X[\sigma(X)]$ is small.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The claims made in the introduction and abstract are backed by theorems.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We clearly state the assumptions under which the results hold and the results are limited to hold under these assumptions. From the proofs it becomes clear that slight deviations from realizability are acceptable. Computational complexity is (briefly) discussed.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: For each theorem the assumptions and proofs are provided.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: The paper does not contain experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: the paper does not include experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: the paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: the paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: the paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification:

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [No]

    Justification: The work is mainly theoretical and does not have any direct societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [No]

    Justification: The work is mainly theoretical and as such does not require safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification:

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification:

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.