Efficient Algorithm for Nonparametric Online Prediction – Senior Thesis Extended Abstract

Haijie Gu
Advisor: John Lafferty
School of Computer Science, Carnegie Mellon University

In this abstract a new approach to online prediction using nonparametric statistics (kernel density estimation (KDE) and kernel regression (KR)) is described and analyzed. This algorithm has the computational advantage as other online algorithms with constant update cost; it also well addresses the variable bandwidth selection issue arising in the online scenario.

In theory, we proved that these online estimators achieve the same minimax rate $O(n^{-\frac{2\beta}{2\beta+1}})$, (where n is the size of the training data and β is the highest order of continuous derivative of the true function), as the standard batch estimators.

In practice, we based these single online estimators on the weighted-expert framework to select the true optimal bandwidth. Assuming the optimal estimator is among the set of experts, the combined experts estimator adapts to the true optimal risk rate – yielding global smoothness of the estimation.

The main contribution is the efficient online estimator with the proof of its minimax rate of convergence. This novel approach lays the foundation for performing more sophisticated nonparametric online predictions, for example, the multi-task online prediction. In the following sections, we will describe the background, methodology, outline of the theoretical analysis and summary.

1 Background

Nonparametric statistical methods such as kernel density estimation or regression have been well developed in the past few decades. Here, we briefly introduce kernel density estimation (KDE) and kernel regression (KR). See [3].

KDE: given $X_1, \ldots, X_n$, as i.i.d inputs, KDE estimates the density of point x as follows:

$$\hat{f}_n(x) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{h} K\left(\frac{x - X_i}{h}\right) \tag{1}$$

KR: given $(X_1, Y_1), \ldots, (X_n, Y_n)$, as i.i.d pairs, and a true function m such that $Y = m(X) + \epsilon$, where ϵ is an independent Gaussian noise with mean zero and known variance, KR estimates the regression function $\hat{m}_n$ at point x as follows:

$$\hat{m_n}(x) = \frac{\sum_{i=1}^{n} K(\frac{x - X_i}{h}) Y_i}{\sum_{i=1}^{n} K(\frac{x - X_i}{h})} \tag{2}$$

In equation 1 and 2, there are two parameters that need to be chosen: the kernel function K, and the bandwidth h. In fact the of the choice of kernel is not very significant, and here we choose K to be the Gaussian probability density function for its simplicity [3].

On the other hand, the choice of bandwidth, also known as the smoothing parameter which controls the degree of smoothness of the regression function, turns out to be critical. In the literature, it has been well studied that for standard batch kernel estimator, the optimal bandwidth $O(n^{-\frac{1}{2\beta+1}})$ is closely related to the sample size n and the degree of smoothness β of the true function. This optimal bandwidth results in the asymptotically optimal rate of convergence – $O(n^{-\frac{2\beta}{2\beta+1}})$, which is also the minimax rate for nonparametric estimation [3].

In contrast to the fixed sample size n in batch estimations, in the online scenario, predictions are made based on sequentially arriving data. With the increasing sample size, choosing the proper bandwidth become more challenging.

The first challenge is that, suppose we knew the optimal bandwidth for any given n and β, as 1 and 2 indicate, adaptively changing the bandwidth at each time step would involve recomputing the kernel weights of all previous points, which leads to a total cost of $O(n^2)$. In the paper, we developed an efficient online kernel estimator with constant update cost and a total cost of $O(n)$. Furthermore, we proved that it achieves the same optimal rate as the batch estimator.

The second challenge is that, without the assumption in the first challenge, the actual optimal bandwidth is unknown to us. All we have is the asymptotic expression $h = O(n^{-\frac{1}{2\beta+1}})$. Not only do we lack the knowledge of β, the smoothness of the true function, but we also miss the constant factor, which depends on the unknown distribution of X. Therefore, the theoretical

optimal bandwidth serves only as a guideline, and in practice, the bandwidth selection is often performed through cross-validation. To overcome this challenge while preserving a low computational cost, we use the weighted-expert algorithm. Existing results about aggregation strategies, like exponential weighted expert algorithm [1], shows that the cumulative loss of the combined estimator is at most $O(n + \log K)$ more than that of the best expert, where K is the number of the experts and the constant factor is between zero and one. Having this remark, we are able to claim that the empirical performance of our algorithm adapts to the true optimal estimator at the rate of $O(n^{-1} \log K)$ if the range of experts' bandwidth contains the optimal bandwidth.

The significance of this algorithm provides insight into the design of a nonparametric online method with computational efficiency, asymptotically optimal risk rate, and desirable empirical performance.

2 Methodology

2.1 Efficient Online Estimator

The core piece of our algorithm is the online kernel estimator extended from the traditional batch kernel estimator. Although there are many variations of kernel estimators, in the paper we focus on kernel density estimator 1, and kernel estimator for regression 2.

Unlike the batch kernel estimator, which fixes one bandwidth for all arriving data points, the online counterpart allows bandwidth h_t varying as time t. Formally, we define the online kernel density estimator at time T to be:

$$\hat{f}_T(x) = \frac{1}{T} \sum_{t=1}^{T} K_{h_t}(x, X_t) \tag{3}$$

Similarly kernel regression estimator at time T is:

$$\hat{m}_T(x) = \frac{\sum_{t=1}^{T} K_{h_t}(x, X_t) Y_t}{\sum_{t=1}^{T} K_{h_t}(x, X_t)} \tag{4}$$

where $K_{h_t}(x, X_t) = \frac{1}{h_t} K(\frac{x - X_t}{h_t})$.

In both case, the estimator for next time step can be easily updated by the following rule:

$$\hat{m}_{T+1}(x) = \frac{Nom(\hat{m}_T(x)) + K_{h_{T+1}}(x, X_{T+1}) Y_{T+1}}{Dnom(\hat{m}_T(x)) + K_{h_{T+1}}(x, X_{T+1})} \tag{5}$$

Notice that the new bandwidth h_{T+1} is only applied to the latest data points (X_{T+1}, Y_{T+1}), leaving all previous points using "old" bandwidth. Although we choose not to update the kernel weights for previous data points, the accuracy is not sacrificed much. In the risk analysis section, we will show that by letting $h_t = O(t^{-\frac{1}{2\beta+1}})$, these online estimators achieve the same optimal rate as the batch estimators.

In terms of the implementation detail, we need to keep the nominator and the denominator for each possible value x in an discretized input space; this can be implemented by binning the input space. Hence, for fixed resolution and size of the experts, these online estimators make online prediction at the cost of $O(1)$ for runtime, and also $O(1)$ for space. The impact of large constant induced by high dimension and resolution can be soothed by parallel computing over the grid of the input space.

2.2 Weighted Expert

Although we have asymptotic result for the step-wise optimal bandwidth h_t, the true optimal bandwidth h_t remains unknown for each time step t. In order to adapt to a (near) optimal estimation, we employ a set of experts who make their predictions using different belief on the "optimal" bandwidth h_k^t varying with the time t, where k is the index of the expert.

Suppose there are K experts in an expert space ξ, and at each time step t, experts make their predictions: $\{\hat{m}_{E,t} : E \in \xi\}$, we make our own prediction $\hat{m}_t$ based on the experts' predictions. After the true label Y_t is revealed, each expert incurs an instant loss based on some loss function $l : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, where $\mathcal{Y}$ is the outcome space. The regret on expert k is defined in 6 as the difference of the cumulative loss between our predictor and that of expert k. The cumulative loss of the predictor and expert E at time t are denoted as $\hat{L}_t$ and $L_{k,t}$ respectively.

$$\begin{aligned} R_{E,t} &= \sum_{i=1}^{t} l(\hat{m}_i(x_i), y_i) - \sum_{i=1}^{t} l(\hat{m}_{i,E}(x_i), y_i) \\ &= \hat{L}_t - L_{E,t} \end{aligned} \tag{6}$$

If we know that one of the experts biases on the optimal bandwidth, by minimizing the regret on the best expert, we adapt our risk to the near optimal risk. The exponential weighting procedure described below is known to minimize 6 at rate $O(1 + n^{-1} \log K)$. [1]

2.3 Algorithm

Let the expert space $\xi = \{1 \dots K\}$.

Define $\hat{\mathbf{m}}_{\xi,\mathbf{t}}$, $\mathbf{w_t}$, $\mathbf{R_t}$ to be the K-length vector of experts' predictions, weights, and loss at the time t respectively, $t = 1, \dots, T$.

Algorithm 1 Online Expert Kernel Regression

for $k \in \xi$ **do**
  $w_{k,1} \leftarrow \frac{1}{K}$
end for
for $t = 1$ to T **do**
  $\hat{m}_t \leftarrow \mathbf{w_t} \cdot \hat{\mathbf{m}}_{\xi,\mathbf{t}}$
  $\mathbf{R_t} \leftarrow \mathbf{R_{t-1}} + l(\hat{m}_t, y_t) - l(\hat{m}_{\xi,t}, y_t)$
  $\mathbf{w_{t+1}} \leftarrow \frac{\exp R_t}{||\exp R_t||}$
end for

In the above algorithm, the expert's estimator $\hat{\mathbf{m}}_{\xi,\mathbf{t}}$ is the online kernel estimator described in (4) with an expert-dependent step-wise bandwidth function taken from a bandwidth space. In our algorithm, at time t, the expert k chooses its bandwidth to be $h_t^k = c_k t^{-\frac{1}{2\beta_k+1}}$, where c_k and β_k are the belief of expert k on the constant and smoothness parameter in the bandwidth.

3 Risk Analysis

The general technique to show that the online estimator achieves the same optimal rate is done by performing bias-variance risk decomposition [3] of the online estimator, and bound it using integral approximation. The proof section of the paper is organized as follows: we first prove the KDE case when $\beta = 2$, and then extend it to KR (local constant regression) when $\beta = 2$. For $\beta > 2$, we prove the result for local polynomial regression, which cancels out all lower order terms in the Taylor's expansion. See [2] for details of local polynomial regression.

At the end of the proof section, we have reached the conclusion that our online kernel estimators have the asymptotically rate $O(n^{-\frac{2\beta}{2\beta+1}})$, which is the same as the minimax rate of the batch estimators.

4 Summary

We discussed the challenges of nonparametric online prediction, and proposed efficient online estimation algorithms which achieve the same asymptotically optimal rate as their batch counterparts. Combining the weighted expert framework, the resulting estimator adapts to the true optimal risk at a rate of $O(1 + n^{-1} \log K)$. The performance is evaluated through simulation on various regression functions.

References

[1] Nicolo Cesa-Bianchi and Gabor Lugosi. *PREDICTION LEARNING AND GAMES*, pages 15–17. Cambridge University Press.

[2] Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*, pages 57–105. Chapman and Hall/CRC.

[3] Larry Wasserman. *All of Nonparametric Statistics*, pages 54–57. Springer.