# A Lower Bound for a Prediction Algorithm under the Kullback-Leibler Game

**Raisa Dzhamtyrova** RDZHAMTYROVA@TURING.AC.UK
*The Alan Turing Institute, 2QR, 96 Euston Rd, Somers Town, London, NW1 2DB, UK*

**Yuri Kalnishkan** YURI.KALNISHKAN@RHUL.AC.UK
*Department of Computer Science, Royal Holloway University of London, Egham, Surrey, TW20 0EX, UK and*
*Laboratory of Advanced Combinatorics and Network Applications, Moscow Institute of Physics and Technology, Institutsky per., 9, Dolgoprudny, 41701, Russia*

**Editor:** Lars Carlsson, Zhiyuan Luo, Giovanni Cherubin and Khuong An Nguyen

## Abstract

We obtain a lower bound for an algorithm predicting finite-dimensional distributions (i.e., points from a simplex) under Kullback-Leibler loss. The bound holds w.r.t. the class of softmax linear predictors. We then show that the bound is asymptotically matched by the Bayesian universal algorithm.

**Keywords:** On-line learning, competitive prediction, loss bounds.

## 1. Introduction

We consider the on-line learning scenario with signals (on-line regression). The following events are repeated for $t = 1, 2, \ldots$. The learner sequentially reads a signal $x_t$, makes a prediction $\gamma_t$ on the basis of the signal and past observations, and then receives the true outcome $y_t$. The signals, predictions, and outcomes are drawn from spaces known from the start. The quality of the learner's predictions is assessed using a known loss function $\lambda(\gamma, y)$.

We want the learner to suffer low cumulative loss $\mathrm{Loss}(T) = \sum_{t=1}^{T} \lambda(\gamma_t, y_t)$ over $T$ steps. We approach this task within the competitive on-line prediction framework. According to this framework, no mechanism (probabilistic or other) generating the signals and outcomes is postulated. Instead we take a pool of competitor strategies and aim to build an algorithm that suffers loss not much worse than any strategy from the pool on every possible sequence of signals and outcomes.

Foster et al. (2018) consider the setup where predictions and outcomes are elements of a simplex $\Delta \in \mathbb{R}^d$ and the loss is measured by the Kullback-Leibler divergence. The framework may be regarded as soft multi-class classification; Foster et al. (2018) discuss applications to bandit multiclass learning and online multiclass boosting.

Let the signals come from $\mathbb{R}^{nd}$ and take for competitors all arrays of $d$ linear functions $\mathbb{R}^n \to \mathbb{R}$ with softmax applied on top of them to ensure we output a distribution. Foster et al. (2018) construct an algorithm for this framework with a competitive loss bound involving an extra term that grows as $5nd \ln T$ in time (we use $n$, $d$ and $T$ as in this paper). Later Dzhamtyrova and Kalnishkan

(2019) improved the bound to $\frac{nd}{2}\ln T$. This left one with the question of whether the bound can be improved further.

In this paper we obtain a lower bound for the regret of $\frac{n(d-1)}{2}\ln T - C$. This is below the regret of know algorithms of Foster et al. (2018) and Dzhamtyrova and Kalnishkan (2019). We then proceed to show that the regret is matched asymptotically by the Bayesian algorithm of Dzhamtyrova and Kalnishkan (2019). With a finer analysis, an upper bound of $\frac{n(d-1)}{2}\ln T + C$ can be obtained. Note that $n(d-1)$ can be thought of as the number of degrees of freedom of the competitor pool; the nominal dimension $nd$ is reduced by making the prediction fit the simplex.

### 1.1. Related Work

The history of algorithms competitive with large parametric classes of strategies can be traced back to universal portfolios by Cover and Ordentlich (1996), which apply in the context of investment decisions and compete against portfolio selection techniques. In that framework one is interested in maximising the wealth, but the problem can be restated in terms of losses.

One can consider outcomes and predictions from the one-dimensional interval $[0,1]$ and signals $x_t \in \mathbb{R}^n$. A natural choice of competitor strategies are then linear functions on $x_t$. Vovk (2001) and Azoury and Warmuth (2001) propose an algorithm for this framework (Vovk-Azoury-Warmuth predictor, also known as the aggregating algorithm regression) targeted at square loss $\lambda(\gamma, y) = (\gamma - y)^2$. Gammerman et al. (2004) obtain a kernelised version of the predictor. Zhdanov and Kalnishkan (2013) study similar competitive bounds for standard ridge regression.

Kakade and Ng (2005) deal with one-dimensional outcomes in the context of logarithmic loss, where a prediction $p(\cdot)$ is a function on $\mathbb{R}$ and the loss $\lambda(p, y) = -\log p(y)$ is considered. The competitors are the generalised linear models of the form $p(y \mid x, \theta) = f(y, \theta' x)$, where $\theta$ is a parameter vector and $f$ is some fixed function. Zhdanov and Vovk (2010) consider a similar set-up for square loss taking a competitor predictor $\sigma(\theta' x)$, i.e., applying a fixed function to the linear predictor.

Dzhamtyrova and Kalnishkan (2020) obtain a bound with an extra term of the order $\frac{n(d-1)}{2}\ln T$ for the Kullback-Leibler loss in the setup similar to this paper, but the competitors were logistic regression functions, where $n(d-1)$ is the explicit dimension of the pool of competitors. Dzhamtyrova and Kalnishkan (2020) also discuss some applications of the Kullback-Leibler loss in the introduction section and perform numerical experiments.

Optimality results are relatively rare in the literature. Of the work cited above, Vovk (2001) and Kakade and Ng (2005) provide lower bounds to their results. A number of results equivalent to lower bounds were obtained for the probabilistic/compression setup (Xie and Barron, 1997, 2000). The construction we use for the lower bound is essentially similar. Remark 5 discusses the difference.

It is worth noting that similar problems have been considered in the context of on-line convex optimisation of Zinkevich (2003) (see the survey by Hazan (2016)). We believe that results from that area cannot be straightforwardly applied to this case, as our settings are not $\alpha$-strongly convex.

## 2. Preliminaries

### 2.1. Games and Losses

A *game* $\mathfrak{G}$ is a triple of an *outcome space* $\Omega$, prediction space $\Gamma$, and a *loss function* $\lambda : \Gamma \times \Omega \to [0, +\infty]$.

A *prediction strategy* $\mathcal{S}$ for a game $\mathfrak{G}$ working with signals from a *signal space* $X$ is a mapping $\mathcal{S} : (X \times \Omega)^* \times X \to \Gamma$. Intuitively, $\mathcal{S}$ supplies predictions for the learner acting according to this protocol:

**for** $t = 1, 2, \ldots$ **do**
  the learner reads signal $x_t \in X$
  the learner produces $\gamma_t \in \Gamma$
  learner sees $y_t \in \Omega$
**end**

**Protocol 1:** On-line learning protocol

On a sequence $(x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$ the learner using a strategy $S$ suffers cumulative loss

$$\mathrm{Loss}_{\mathcal{S}}(T) = \sum_{t=1}^{T} \lambda(\gamma_t, y_t) = \sum_{t=1}^{T} \lambda(\mathcal{S}(x_1, y_1, \ldots, x_{t-1}, y_{t-1}, x_t), y_t) \ .$$

The index $\mathcal{S}$ will be dropped if it is clear from the context.

In this paper we consider the game where the outcome and prediction spaces are the unit $(d-1)$-simplex $\Delta = \{(p_1, p_2, \ldots, p_d) \mid p_i \geq 0 \text{ for } i = 1, 2, \ldots, d \text{ and } \sum_{i=1}^{d} p_i = 1\} \subseteq \mathbb{R}^d$. The outcomes and predictions are distributions on a finite set of $d$ elements and we use the Kullback-Leibler divergence to measure the loss. For $\gamma = (\gamma^{(1)}, \gamma^{(2)}, \ldots, \gamma^{(d)}) \in \Delta$ and $y = (y^{(1)}, y^{(2)}, \ldots, y^{(d)}) \in \Delta$ we let $\lambda(\gamma, y) = \sum_{i=1}^{d} y^{(i)} \ln \frac{y^{(i)}}{\gamma^{(i)}}$. The ambiguous expressions $0 \ln 0$ and $0 \ln \frac{0}{0}$ evaluate to 0.

The signals will be $n$-dimensional real vectors, i.e., we take $X = \mathbb{R}^n$.

### 2.2. Competitors

Our goal is to construct an algorithm performing in terms of cumulative loss nearly as well as any strategy from the following class. Take $\theta \in \mathbb{R}^{nd}$ (by default all vectors are column vectors in this paper) and partition it as $\theta' = (\theta'_1, \theta'_2, \ldots, \theta'_d)$ so that $\theta_1, \theta_2, \ldots, \theta_d \in \mathbb{R}^n$. The strategy $\mathcal{S}_\theta$ is oblivious, i.e., its prediction only depends on the current signal $x$. The strategy outputs $\xi(\theta, x) = (\xi_1(\theta, x), \xi_2(\theta, x), \ldots, \xi_d(\theta, x))'$, where $\xi_i(\theta, x)$ are the softmax functions

$$\xi_i(\theta, x) = \frac{e^{\theta'_i x}}{\sum_{j=1}^{d} e^{\theta'_j x}} \tag{1}$$

for $i = 1, 2, \ldots, d$. Clearly, $\xi(\theta, x) \in \Delta$ for all $\theta \in \mathbb{R}^{nd}$ and $x \in \mathbb{R}^n$. We will abbreviate $\mathrm{Loss}_{\mathcal{S}_\theta}(T)$ to $\mathrm{Loss}_\theta(T)$.

## 3. Loss Bounds

The following theorem provides a lower bound for any strategy $\mathcal{S}$ making predictions on the basis of signals.

**Theorem 1** *For every positive integer $n$, positive integer $d \geq 2$, and $a > 0$, there are signals $x_t$ of norm 1, $\|x_t\| = 1$, $t = 1, 2, \ldots$, a random process of outcomes $Y_t$, $t = 1, 2, \ldots$, and $C > 0$ such that for every strategy $\mathcal{S}$ we have*

$$\mathbf{E}\left(\mathrm{Loss}_{\mathcal{S}}(T) - \inf_{\theta \in \mathbb{R}^{nd}} (\mathrm{Loss}_\theta(T) + a\|\theta\|^2)\right) \geq \frac{n(d-1)}{2} \ln T - C$$

*for all* $T = 1, 2, \ldots$

The theorem is proven in Section 4. The process $Y_t$ is a mixture of i.i.d. processes with a Dirichlet prior on parameters.

The learner can match this bound with a strategy that is essentially the Bayesian mixture with the Gaussian prior. The strategy $\mathcal{B}_a$ takes a parameter $a > 0$. On step $t$, $\mathcal{B}_a$ outputs the mixture

$$\xi_t = \int_{\mathbb{R}^{nd}} \xi(\theta, x_t) p^*_{t-1}(\theta) d\theta \ , \tag{2}$$

with the density $p^*_{t-1}(\theta)$ given by

$$p^*_{t-1}(\theta) = e^{-\operatorname{Loss}_\theta(t-1)} p_0(\theta) \Big/ \int_{\mathbb{R}^{nd}} e^{-\operatorname{Loss}_\theta(t-1)} p_0(\theta) d\theta \ ,$$

where the prior density $p_0$ is Gaussian:

$$p_0(\theta) = \left(\frac{a}{\pi}\right)^{nd/2} e^{-a\|\theta\|^2} \tag{3}$$

(the norm $\|\cdot\|$ on $\mathbb{R}^m$ is the Euclidean norm).

The following theorem improves the result of Dzhamtyrova and Kalnishkan (2019), where a strategy equivalent to $\mathcal{B}_a$ was introduced.

**Theorem 2** *For every positive integer $n$ and positive integer $d \geq 2$ the loss of the Bayesian strategy $\mathcal{B}_a$ with $a > 0$ satisfies*

$$\begin{aligned}\operatorname{Loss}_{\mathcal{B}_a}(T) &\leq \inf_{\theta \in \mathbb{R}^{nd}} (\operatorname{Loss}_\theta(T) + a\|\theta\|^2) + \frac{d-1}{2} \ln \det \left( \frac{1}{4a} \sum_{t=1}^{T} x_t x_t' + 2I \right) + \frac{n}{2} \ln(2d-1) \\ &\leq \inf_{\theta \in \mathbb{R}^{nd}} (\operatorname{Loss}_\theta(T) + a\|\theta\|^2) + \frac{n(d-1)}{2} \ln \left( \frac{\sum_{t=1}^{T} \|x_t\|^2}{4an} + 2 \right) + \frac{n}{2} \ln(2d-1)\end{aligned}$$

*for all positive integers $T$, all signals $x_t \in \mathbb{R}^n$ and outcomes $y_t \in \Delta$, $T = 1, 2, \ldots$*

If for all $t = 1, 2, \ldots, T$ we have $\|x_t\| \leq X$, then

$$\frac{n(d-1)}{2} \ln \left( \frac{\sum_{t=1}^{T} \|x_t\|^2}{4an} + 2 \right) \leq \frac{n(d-1)}{2} \ln \left( \frac{TX^2}{4an} + 2 \right) \sim \frac{n(d-1)}{2} \ln T$$

as $T \to \infty$ for fixed $n$, $d$ and $a$. Thus the lower bound is attained asymptotically.

**Remark 3** *Practical applications of $\mathcal{B}_a$ require a way to evaluate the integral in (2). Note that we do not need to evaluate the normalising constant; upon evaluation of the integrals*

$$\int_{\mathbb{R}^{nd}} \xi_i(\theta, x_t) e^{-\operatorname{Loss}_\theta(t-1)} p_0(\theta) d\theta \ ,$$

*$i = 1, 2, \ldots, d$, one can normalise the components to sum up to 1. For small $n$ the integral can be evaluated by integrating $\xi_i(\theta, x_t) e^{-\operatorname{Loss}_\theta(t-1)}$ w.r.t. the Gaussian density. However, with the growth of $n$ this approach quickly becomes infeasible. In Dzhamtyrova and Kalnishkan (2019) MCMC was used with some success. Theoretical justification of MCMC is hard though. Theorems 1 and 3 by Roberts and Smith (1994) can be used to show that the chain converges but the speed requires further investigation.*

## 4. Proof of the Lower Bound

The outcomes $Y_t$ in our construction will take values at the set of vertexes of the simplex, $V = \{e_1, e_2, \ldots, e_d\}$.

Let $n = 1$. We take constant signals $x_1 = x_2 = \ldots = 1$. In this case the prediction output by $\mathcal{S}$ on step $t$ should be a function of the outcomes $y_1, y_2, \ldots, y_{t-1}$ observed previously.

**Lemma 4** *If there is a joint distribution of $Y_1, Y_2, \ldots, Y_t$, then the optimal on average strategy $\mathcal{B}$ outputs the conditional distribution of $Y_t \mid Y_1, Y_2, \ldots, Y_{t-1}$, i.e., $\gamma_t = \left(\gamma_t^{(1)}, \gamma_t^{(2)}, \ldots, \gamma_t^{(d)}\right)$ such that $\gamma_t^{(i)} = \Pr(Y_t = e_i \mid Y_1 = y_1, Y_2 = y_2, \ldots, Y_{t-1} = y_{t-1})$ on step $t$.*

**Proof**

The expectation of the loss on step $t$ of a prediction $\gamma_t = \gamma_t(y_1, y_2, \ldots, y_{t-1})$ can be written as

$$\mathbf{E} \sum_{i=1}^{d} Y_t^{(i)} \ln Y_t^{(i)} -$$

$$\sum_{y_1,\ldots,y_t \in V} \Pr(Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}) \sum_{i=1}^{d} \Pr(Y_t = e_i \mid Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1}) \ln \gamma_t^{(i)} \ ,$$

where the first term does not include $\gamma_t$ and the inner sum in the second term maximises on $\gamma_t^{(i)} = \Pr(Y_t = e_i \mid Y_1 = y_1, \ldots, Y_{t-1} = y_{t-1})$, $i = 1, 2, \ldots, d$, because of the information inequality (Cover and Thomas, 1991, Theorem 2.6.3). ■

Let us now describe the process $Y_t$. First, we choose $p_1, \ldots, p_d \geq 0$ and such that $\sum_{i=1}^{d} p_i = 1$ at random from the Dirichlet distribution with the parameter of $A > 0$ (to be specified later). The density of $p_1, \ldots, p_d$ is

$$p_0(p_1, \ldots, p_d) = \frac{p_1^{A-1} \ldots p_d^{A-1}}{B(A, \ldots, A)} \ . \tag{4}$$

Then $Y_t$ are independently drawn from the categorical distribution with the parameter $(p_1, \ldots, p_d)$, i.e., $\Pr(Y_t = e_i) = p_i$.

Let us work out the prediction of the optimal strategy $\mathcal{B}$ when the outcomes $y_1, y_2, \ldots, y_{t-1}$ have been observed by step $t$ and the number of $e_i$s among them is $n_t^{(i)}$, $i = 1, 2, \ldots, d$. These are random variable in our framework. By the Bayes theorem the density $p(p_1, \ldots, p_d \mid y_1, \ldots, y_{t-1})$ of the posterior distribution of $p_1, \ldots, p_d$ is proportional to $p_1^{n_t^{(1)}+A-1} \ldots p_d^{n_t^{(d)}+A-1}$ and thus it is the density of the Dirichlet distribution with the parameters $(n_t^{(1)} + A, \ldots, n_t^{(d)} + A)$. The prediction of the optimal learner evaluates to

$$\gamma_t^{(i)} = \Pr(Y_t = e_i \mid y_1, \ldots, y_{t-1}) =$$

$$\int_{\Delta} \Pr(Y_t = e_i \mid y_1, \ldots, y_{t-1}, p_1, \ldots, p_d) p(p_1, \ldots, p_d \mid y_1, \ldots, y_{t-1}) dp_1 \ldots dp_d =$$

$$\int_{\Delta} p_i \frac{p_1^{n_t^{(1)}+A-1} \ldots p_d^{n_t^{(d)}+A-1}}{B(n_t^{(1)} + A, \ldots, n_t^{(d)} + A)} dp_1 \ldots dp_d = \frac{n_t^{(i)} + A}{t - 1 + Ad}$$

(we made use of conditional independence of $Y_1, \dots, Y_t$ given $p_1, \dots, p_d$).

The expectation of the loss of $\mathcal{B}$ equals

$$\mathbf{E}\,\mathrm{Loss}_{\mathcal{B}}(T) = \mathbf{E}[\mathbf{E}(\mathrm{Loss}_{\mathcal{B}}(T) \mid p_1, \dots, p_d)] \ ,$$

where the inner conditional expectation equals

$$\mathbf{E}(\mathrm{Loss}_{\mathcal{B}}(T) \mid p_1, \dots, p_d) = -\sum_{t=1}^{T}\sum_{i=1}^{d} p_i \mathbf{E} \ln \frac{n_t^{(i)} + A}{t - 1 + Ad}$$

(we will be assuming fixed $p_1, \dots, p_d$ and dropping conditioning on them for brevity). The expectation in this formula can be expanded as

$$-\mathbf{E} \ln \frac{n_t^{(i)} + A}{t - 1 + Ad} = -\ln \frac{p_i(t-1) + A}{t - 1 + Ad} - \mathbf{E} \ln \left(1 + \frac{n_t^{(i)} - p_i(t-1)}{p_i(t-1) + A}\right) \ .$$

Since the arguments under the logarithm in the first term sum to 1 over $i$, they can be interpreted as probabilities and therefore

$$-\sum_{i=1}^{d} p_i \ln \frac{p_i(t-1) + A}{t - 1 + Ad} \geq -\sum_{i=1}^{d} p_i \ln p_i$$

due to the information inequality (Cover and Thomas, 1991, Theorem 2.6.3).

Let us bound the second term from below. We will denote the fraction

$$\frac{n_t^{(i)} - p_i(t-1)}{p_i(t-1) + A}$$

by $Z_{i,t}$ and use the inequality $\ln(1 + Z_{i,t}) \leq Z_{i,t} - Z_{i,t}^2/2 + Z_{i,t}^3/3$. The expectation $\mathbf{E}Z_{i,t}$ vanishes and the expectation $\mathbf{E}Z_{i,t}^3$ can be bounded from above as follows. We have $\mathbf{E}Z_{i,t} = 0$; for $t = 2, 3, \dots$ let $\mathcal{E}$ denote the event $|n_t^{(i)} - p_i(t-1)| \geq (t-1)^{0.5+\varepsilon}$. By Hoeffding's inequality $\Pr(\mathcal{E}) \leq 2e^{-2(t-1)^{2\varepsilon}}$ (Hoeffding, 1963, Theorem 2). Since $n_t^{(i)}, p_i(t-1) \in [0, t-1]$, we get

$$\mathbf{E}|Z_{i,t}|^3 = \mathbf{E}|Z_{i,t}|^3 \mathbb{1}_{\mathcal{E}} + \mathbf{E}|Z_{i,t}|^3(1 - \mathbb{1}_{\mathcal{E}}) \leq 2\frac{(t-1)^3}{A^3} e^{-2(t-1)^{2\varepsilon}} + \frac{1}{p_i^3 (t-1)^{1.5-3\varepsilon}} \ .$$

For small $\varepsilon > 0$, both the terms on the right-hand side sum over $t$ to finite numbers so that $\sum_{t=1}^{\infty}\sum_{i=1}^{d} p_i \mathbf{E}|Z_{i,t}|^3 = O\left(\sum_{i=1}^{d} p_i^{-2} + 1\right)$.

The term $Z_{i,t}^2/2$ takes more detailed investigation. For $t = 2, 3, \dots$ we have

$$\mathbf{E}\left(\frac{n_t^{(i)} - p_i(t-1)}{p_i(t-1) + A}\right)^2 = \frac{(t-1)p_i(1-p_i)}{(p_i(t-1) + A)^2} =$$

$$\frac{1 - p_i}{p_i(t-1)\left(1 + \frac{A}{p_i(t-1)}\right)^2} \geq \frac{1 - p_i}{p_i(t-1)}\left(1 - \frac{2A}{p_i(t-1)}\right)$$

(we calculated the variance of the sum of conditionally independent terms $Y_1 + \ldots + Y_{t-1}$ and used the inequality $1/(1+x)^2 \geq 1 - 2x$). The second term in the last expression again sums to a finite number:

$$\sum_{t=2}^{\infty} \sum_{i=1}^{d} p_i \frac{2A(1-p_i)}{p_i^2 (t-1)^2} = O\left( \sum_{i=1}^{d} p_i^{-1} \right) \quad .$$

Putting all these terms together, we get

$$\mathbf{E}(\mathrm{Loss}_{\mathcal{B}}(T) \mid p_1, \ldots, p_d) \geq -T \sum_{i=1}^{d} p_i \ln p_i + \frac{d-1}{2} \sum_{t=2}^{T} \frac{1}{t-1} - \tilde{C} \left( \sum_{i=1}^{d} \frac{1}{p_i^2} + 1 \right)$$

for some constant $\tilde{C}$. The second term can be bounded from below:

$$\sum_{t=2}^{T} \frac{1}{t-1} \geq \ln T - \hat{C}$$

for some constant $\hat{C}$.

It remains to integrate this over the simplex with the prior density $p_0$ from (4). By the properties of the Dirichlet distribution, $p_1^{\alpha_1} \ldots p_d^{\alpha_d}$ integrates to a positive number as long as all $\alpha_i > -1$. Thus for $A > 2$ all terms integrate. We thus bound $\mathbf{E}\,\mathrm{Loss}_{\mathcal{B}}(T)$ from below.

Now take $p_1, \ldots, p_t$ and consider a strategy that "knows" the probabilities. If all $p_i > 0$, we can take $\theta = (\ln p_1, \ldots, \ln p_d)$. This strategy outputs the prediction $\gamma_t = (p_1, \ldots, p_d)$ and suffers loss

$$\mathbf{E}(\mathrm{Loss}_{\theta}(T) \mid p_1, \ldots, p_d) = -T \sum_{i=1}^{d} p_i \ln p_i \quad .$$

(If some $p_i$ are zero, we can take the minimax $\theta = (0, 0, \ldots, 0)$; it will lead to finite loss on every step and will not affect the integral as the probability of $p_i = 0$ is zero.) The norm of $\theta$ can be (crudely) bounded as

$$\|\theta\|^2 = \sum_{i=1}^{d} (\ln p_i)^2 \leq \sum_{i=1}^{d} \frac{1}{p_i}$$

and arguing as above it integrates to a finite number for $A > 2$. This gives us an upper bound on $\mathbf{E} \inf_{\theta \in \mathbb{R}^{nd}} (\mathrm{Loss}_{\theta}(T) + a\|\theta\|^2)$ and completes the case of $n = 1$.

For $n > 1$ we generate $n$ arrays of probabilities $(p_1, \ldots, p_d)$ independently. On step $t$ such that $t \mod n = m - 1$ we use the $m$th array for generating $Y_t$ and take $x_t = e_n$. By considering $t$s with each particular $m$ we will be getting regrets of the order $\ln(T/n) = \ln T - \ln n$, and additive $\ln n$ does not matter in our argument.

**Remark 5** *The construction of the proof is similar to those of Xie and Barron (1997) and Xie and Barron (2000), where the regret is bounded from below in a similar setup. We also use a Dirichlet prior and the main term $\frac{d-1}{2} \ln T$ is the same. The difference is in the regret term $\|\theta\|^2$, which is important in our approach. Low loss $\mathrm{Loss}_{\theta}(T)$ can be achieved by a predictor with large $\|\theta\|$, and the norm can grow with $T$. Thus we believe our lower bound does not follow straightforwardly from earlier work.*

## 5. Loss Bounds and the Bayesian Algorithm

Vovk (2001, Lemma 4) shows that Kullback-Leibler loss is mixable. If follows by induction on time that

$$\mathrm{Loss}_{\mathcal{B}}(T) \leq -\ln \int_{\mathbb{R}^{nd}} e^{-\mathrm{Loss}_\theta(T)} p_0(\theta) d\theta = -\ln \left[ \left(\frac{a}{\pi}\right)^{\frac{nd}{2}} \int_{\mathbb{R}^{nd}} e^{-Q(\theta)} d\theta \right] \ ,$$

where $Q(\theta) = \sum_{t=1}^{T} \sum_{i=1}^{d} y_t^{(i)} \ln \frac{y_t^{(i)}}{\xi_i(\theta, x_t)} + a\|\theta\|^2$ and $p_0$ is given by (3).

Our plan is to reduce the integral over $\mathbb{R}^{nd}$ to an integral over $\mathbb{R}^{n(d-1)}$ by substitution. Let

$$\begin{aligned} \varphi_1 &= \theta_1 - \theta_d \\ &\vdots \\ \varphi_{d-1} &= \theta_{d-1} - \theta_d \\ \varphi_d &= \theta_d \end{aligned}$$

and $\varphi' = (\varphi'_1, \varphi'_2, \ldots, \varphi'_d)$. Clearly, this is a linear bijection and its Jacobian evaluates to 1. We get

$$\int_{\mathbb{R}^{nd}} e^{-Q(\theta)} d\theta = \int_{\mathbb{R}^{nd}} e^{-\widetilde{Q}(\varphi)} d\varphi \ ,$$

where $\widetilde{Q}(\varphi) = Q(\theta(\varphi))$ and $\theta(\varphi)$ is the inverse transformation giving $\theta$ from $\phi$.

We will apply a trick due to Zhdanov and Vovk (2010) to bound $\widetilde{Q}$ from above. Let $\widetilde{Q}$ achieve its global minimum at $\varphi_0$. Applying Taylor decomposition with Lagrange's remainder at $\theta_0$ we get

$$\widetilde{Q}(\varphi) = \widetilde{Q}(\varphi_0) + \frac{1}{2}(\varphi - \varphi_0)' H_{\widetilde{Q}}(\psi(\varphi))(\varphi - \varphi_0) \ ,$$

where $H_{\widetilde{Q}}$ is the Hessian of $\widetilde{Q}$ and $\psi(\varphi)$ is a convex combination of $\varphi_0$ and $\varphi$. Since

$$\widetilde{Q}(\varphi_0) = \inf_{\varphi} \widetilde{Q}(\varphi) = \inf_{\theta} Q(\theta) = \inf_{\theta \in \mathbb{R}^{nd}} (\mathrm{Loss}_\theta(T) + a\|\theta\|^2) \ ,$$

we get

$$\mathrm{Loss}_{\mathcal{B}}(T) \leq \inf_{\theta \in \mathbb{R}^{nd}} (\mathrm{Loss}_\theta(T) + a\|\theta\|^2) - \ln \left[ \left(\frac{a}{\pi}\right)^{\frac{nd}{2}} \int_{\mathbb{R}^{nd}} e^{-\frac{1}{2}\varphi' H_{\widetilde{Q}}(\psi(\varphi+\varphi_0))\varphi} d\varphi \right] \ .$$

We will now bound the second differential $\varphi' H_{\widetilde{Q}}(\psi(\varphi + \varphi_0))\varphi$ from above.

One can write

$$\xi_i(\theta, x) = \frac{e^{\theta'_i x}}{\sum_{j=1}^{d} e^{\theta'_i x}} = \frac{e^{(\theta'_i - \theta'_d)x}}{\sum_{j=1}^{d} e^{(\theta'_j - \theta'_d)x}} = \frac{e^{\varphi'_i x}}{\sum_{j=1}^{d-1} e^{\varphi'_j x} + 1}$$

for $i = 1, 2, \ldots, d-1$ and

$$\xi_d(\theta, x) = \frac{e^{\theta'_d x}}{\sum_{j=1}^{d} e^{\theta'_i x}} = \frac{1}{\sum_{j=1}^{d-1} e^{\varphi'_j x} + 1} \ .$$

Therefore the functions $\zeta_i(\varphi, x) = \xi_i(\theta(\varphi), x)$ do not depend on $\varphi_d$. In

$$\widetilde{Q}(\varphi) = \sum_{t=1}^{T} \sum_{i=1}^{d} y_t^{(i)} \ln \frac{y_t^{(i)}}{\zeta_i(\varphi, x_t)} + a\|\theta(\varphi)\|^2$$

only the last term depends on $\varphi_d$.

One can check by direct evaluation that

$$\frac{\partial \zeta_i(\varphi, x)}{\partial \varphi_k} = \zeta_i(\varphi, x)(\delta_{i,k} - \zeta_k(\varphi, x))x' \ ,$$

for $i = 1, 2, \ldots, d$ and $k = 1, 2, \ldots, d-1$, where $\delta_{i,k}$ is the Kronecker delta. The derivatives w.r.t. $\varphi_d$ equal to 0. Therefore

$$\frac{\partial}{\partial \varphi_k} \sum_{t=1}^{T} \sum_{i=1}^{d} y_t^{(i)} \ln \frac{y_t^{(i)}}{\zeta_i(\varphi, x_t)} = -\sum_{t=1}^{T} (y_t^{(k)} - \zeta_k) x_t' \ ,$$
$$\frac{\partial^2}{\partial \varphi_k \partial \varphi_m} \sum_{t=1}^{T} \sum_{i=1}^{d} y_t^{(i)} \ln \frac{y_t^{(i)}}{\zeta_i(\varphi, x_t)} = \sum_{t=1}^{T} \zeta_k(\varphi, x_t)(\delta_{m,k} - \zeta_m(\varphi, x_t)) x_t x_t'$$

for $k, m = 1, 2, \ldots, d-1$. If $k = d$ or $m = d$, the derivative vanishes.

Letting $Z_t = (\zeta_1(\varphi, x_t), \zeta_2(\varphi, x_t), \ldots, \zeta_d(\varphi, x_t))'$, we can write the matrix of second derivatives as $\sum_{t=1}^{T} (\mathrm{diag}(Z_t) - Z_t Z_t') \otimes x_t x_t'$, where $\mathrm{diag}(Z_t)$ is the diagonal matrix having elements of $Z_t$ on the diagonal.

We now need to bound the corresponding quadratic form from above uniformly over $\varphi$.

**Lemma 6** *Let $P = (p_1, p_2, \ldots, p_m)'$, where $p_i \geq 0$ for $i = 1, 2, \ldots, m$ and $\sum_{i=1}^{m} p_i = 1$. Then*

$$0 \leq \sum_{i=1}^{m} p_i x_i^2 - \left( \sum_{i=1}^{m} p_i x_i \right)^2 \leq \frac{1}{2} \sum_{i=1}^{m} x_i^2$$

*for all $x_1, x_2, \ldots, x_m \in \mathbb{R}$.*

The lemma is proven in Appendix A.

**Remark 7** *The upper bound in the lemma is sharp for $m \geq 2$. Take $p_1 = p_2 = 1/2$, $p_3 = p_4 = \ldots = p_m = 0$ and $x_1 = -x_2 = x$, $x_3 = x_4 = \ldots = x_m = 0$. The upper bound turns into an equality.*

The matrix $\mathrm{diag}(Z_t) - Z_t Z_t'$ is symmetric. The lemma implies that it is positive semi-definite and $\mathrm{diag}(Z_t) - Z_t Z_t' \preccurlyeq \frac{1}{2} I_{d-1}$, where $I_{d-1}$ is the $((d-1) \times (d-1))$-unit matrix. Distributivity of the Kronecker product and the properties of the product of positive semi-definite matrices (see, e.g., the book by Horn and Johnson, 1991, 4.2.7 and 4.2.13) imply

$$\sum_{t=1}^{T} (\mathrm{diag}(Z) - Z_t Z_t') \otimes x_t x_t' \preccurlyeq \sum_{t=1}^{T} \frac{1}{2} I_{d-1} \otimes x_t x_t' = \frac{1}{2} I_{d-1} \otimes \sum_{t=1}^{T} x_t x_t' \ .$$

The term $\|\theta\|^2$ can be bounded from above as follows:

$$\|\theta\|^2 = \sum_{i=1}^{d} \|\theta_i\|^2 = \sum_{i=1}^{d-1} \|\varphi_i + \varphi_d\|^2 + \|\varphi_d\|^2 \leq 2\sum_{i=1}^{d-1} \|\varphi_i\|^2 + (2d-1)\|\varphi_d\|^2 \ .$$

**Remark 8** *One can check that this bound is sharp by considering* $\varphi_i \in \mathbb{R}$ *such that* $\varphi_1 = \varphi_2 = \ldots = \varphi_d = 1$ *and the corresponding* $\theta_1 = \theta_2 = \ldots \theta_{d-1} = 2$, $\theta_d = 1$.

Putting this all together we get

$$\frac{1}{2}\varphi' H_{\tilde{Q}}(\psi(\varphi+\varphi_0))\varphi \leq \frac{1}{2}\varphi'_{1:d-1}\left(\frac{1}{2}I_{d-1} \otimes \sum_{t=1}^{T} x_t x_t'\right)\varphi_{1:d-1} + 2a\sum_{i=1}^{d-1}\|\varphi_i\|^2 + (2d-1)a\|\varphi_d\|^2 \ ,$$

where $\varphi'_{1:d-1} = (\varphi'_1, \varphi'_2, \ldots, \varphi'_{d-1})$. Fubini's theorem imply

$$\int_{\mathbb{R}^{nd}} e^{-\frac{1}{2}\varphi' H_{\tilde{Q}}(\psi(\varphi+\varphi_0))\varphi} d\varphi \geq$$

$$\int_{\mathbb{R}^{nd}} e^{\frac{1}{2}\varphi'_{1:d-1}\left(\frac{1}{2}I_{d-1}\otimes\sum_{t=1}^{T} x_t x_t'\right)\varphi_{1:d-1} + 2a\sum_{i=1}^{d-1}\|\varphi_i\|^2 + (2d-1)a\|\varphi_d\|^2} d\varphi =$$

$$\int_{\mathbb{R}^{n(d-1)}} e^{-\varphi'_{1:d-1}\left(\frac{1}{4}I_{d-1}\otimes\sum_{t=1}^{T} x_t x_t' + 2aI_{n(d-1)}\right)\varphi_{1:d-1}} d\varphi_{1:d-1} \int_{\mathbb{R}^n} e^{-(2d-1)a\|\varphi_d\|^2} d\varphi_d =$$

$$\int_{\mathbb{R}^n} e^{-\varphi'_1\left(\sum_{t=1}^{T} x_t x_t'/4 + 2aI_n\right)\varphi_1} d\varphi_1 \ldots \int_{\mathbb{R}^n} e^{-\varphi'_{d-1}\left(\sum_{t=1}^{T} x_t x_t'/4 + 2aI_n\right)\varphi_{d-1}} d\varphi_{d-1} \cdot$$

$$\int_{\mathbb{R}^n} e^{-(2d-1)a\|\varphi_d\|^2} d\varphi_d = \left(\int_{\mathbb{R}^n} e^{-\varphi'\left(\sum_{t=1}^{T} x_t x_t'/4 + 2aI_n\right)\varphi} d\varphi\right)^{d-1} \int_{\mathbb{R}^n} e^{-(2d-1)a\|\varphi_d\|^2} d\varphi$$

(the matrix $\frac{1}{4}I \otimes \sum_{t=1}^{T} x_t x_t'$ is block diagonal with identical blocks).

We are now going to evaluate the integrals of exponents of quadratic forms (Beckenbach and Bellman, 1961, Chapter 2, Theorem 3). If an $(n \times n)$-matrix $A$ is symmetric and positive definite, then

$$\int_{\mathbb{R}^n} e^{-x'Ax} dx = \frac{\pi^{n/2}}{\sqrt{\det A}} \ .$$

We get

$$\int_{\mathbb{R}^n} e^{-(2d-1)a\|\varphi_d\|^2} d\varphi_d = \left(\frac{\pi}{a(2d-1)}\right)^{n/2} \ ,$$

$$\int_{\mathbb{R}^n} e^{-\varphi'\left(\sum_{t=1}^{T} x_t x_t'/4 + 2aI\right)\varphi} d\varphi = \frac{\pi^{n/2}}{\sqrt{\det\left(\sum_{t=1}^{T} x_t x_t'/4 + 2aI\right)}} \ .$$

Therefore

$$-\ln\left[\left(\frac{a}{\pi}\right)^{\frac{nd}{2}}\int_{\mathbb{R}^{nd}} e^{-\frac{1}{2}\varphi' H_{\widetilde{Q}}(\psi(\varphi+\varphi_0))\varphi} d\varphi\right] \le$$
$$\frac{nd}{2}\ln\frac{\pi}{a} + \frac{d-1}{2}\ln\det\left(\sum_{t=1}^{T} x_t x_t'/4 + 2aI\right) - \frac{n(d-1)}{2}\ln\pi + \frac{n}{2}\ln\frac{a(2d-1)}{\pi} =$$
$$\frac{d-1}{2}\ln\det\left(\frac{1}{4a}\sum_{t=1}^{T} x_t x_t' + 2I\right) + \frac{n}{2}\ln(2d-1)$$

It remains to bound the determinant from above. For a symmetric positive semi-definite $(m \times m)$-matrix $M$ we have

$$\det M \le \left(\frac{\operatorname{tr} M}{m}\right)^m .$$

(see, e.g., Proposition 5 by Kalnishkan (2016)).

Thus

$$\det\left(\frac{1}{4a}\sum_{t=1}^{T} x_t x_t' + 2I\right) \le \left(\frac{\sum_{t=1}^{T}\|x_t\|^2}{4an} + 2\right)^n .$$

Theorem 2 follows.

## References


K. S. Azoury and M. K. Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43:211–246, 2001.

E. F. Beckenbach and R. E. Bellman. *Inequalities*. Springer, 1961.

T. M. Cover and E. Ordentlich. Universal portfolios with side information. *IEEE Transactions on Information Theory*, 42(2), 1996.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley & Sons, 1991.

R. Dzhamtyrova and Y. Kalnishkan. Competitive online generalised linear regression with multidimensional outputs. In *2019 International Joint Conference on Neural Networks (IJCNN)*. IEEE, 2019.

R. Dzhamtyrova and Y. Kalnishkan. Universal algorithms for multinomial logistic regression under Kullback–Leibler game. *Neurocomputing*, 397:369–380, 2020.

D. J. Foster, S. Kale, H. Luo, M. Mohri, and K. Sridharan. Logistic regression: The importance of being improper. In *Conference on Learning Theory 2018*, volume 75 of *Proceedings of Machine Learning Research*, pages 167–208, 2018.

A. Gammerman, Y. Kalnishkan, and V. Vovk. On-line prediction with kernels and the complexity approximation principle. In *Uncertainty in Artificial Intelligence, Proceedings of the Twentieth Conference*, pages 170–176. AUAI Press, 2004.

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

W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

R. A. Horn and C. R. Johnson. *Topics in Matrix Analysis*. Cambridge University Press, 1991.

S. M. Kakade and A. Y. Ng. Online bounds for bayesian algorithms. In *Advances in neural information processing systems*, pages 641–648, 2005.

Y. Kalnishkan. An upper bound for aggregating algorithm for regression with changing dependencies. In *International Conference on Algorithmic Learning Theory*, pages 238–252. Springer, 2016.

Gareth O Roberts and Adrian FM Smith. Simple conditions for the convergence of the gibbs sampler and metropolis-hastings algorithms. *Stochastic processes and their applications*, 49(2):207–216, 1994.

V. Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

Q. Xie and A. R. Barron. Minimax redundancy for the class of memoryless sources. *IEEE Transactions on Information Theory*, 43(2):646–657, 1997.

Q. Xie and A. R. Barron. Asymptotic minimax regret for data compression, gambling, and prediction. *IEEE Transactions on Information Theory*, 46(2):431–445, 2000.

F. Zhdanov and Y. Kalnishkan. An identity for kernel ridge regression. *Theoretical Computer Science*, 473:157–178, 2013.

F. Zhdanov and V. Vovk. Competitive online generalized linear regression under square loss. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 531–546. Springer, 2010.

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

## Appendix A. Proof of the Variance Lemma

In this section we prove Lemma 6.

The expression $\sum_i p_i x_i^2 - (\sum_i p_i x_i)^2$ has the meaning of variance of $x_i$ w.r.t. the distribution $p_i$, so it is non-negative.

The upper bound is equivalent to

$$\frac{1}{2}\sum_i x_i^2 + \left(\sum_i p_i x_i\right)^2 - \sum_i p_i x_i^2 \geq 0 \ .$$

Opening up the brackets yields

$$\frac{1}{2}\sum_i x_i^2 + \sum_i p_i^2 x_i^2 + \sum_{i\neq j} p_i p_j x_i x_j - \sum_i p_i x_i^2 \geq 0$$

and

$$\sum_i \left(\frac{1}{2} + p_i^2 - p_i\right) x_i^2 \geq -\sum_{i\neq j} p_i p_j x_i x_j$$

The absolute value of the right-hand side can be bounded as follows:

$$\left|\sum_{i\neq j} p_i p_j x_i x_j\right| \leq \sum_{i\neq j} p_i p_j |x_i||x_j| \leq \sum_{i\neq j} p_i p_j \frac{x_i^2 + x_j^2}{2} =$$

$$\frac{1}{2}\sum_i \sum_{j,j\neq i} p_i p_j x_i^2 + \frac{1}{2}\sum_j \sum_{i,j\neq i} p_i p_j x_j^2 =$$

$$\frac{1}{2}\sum_i p_i(1-p_i)x_i^2 + \frac{1}{2}\sum_j p_j(1-p_j)x_j^2 = \sum_i p_i(1-p_i)x_i^2 \ .$$

For every $i$ we have $\frac{1}{2} + p_i^2 - p_i \geq p_i(1-p_i)$ since $\frac{1}{4} \geq p_i(1-p_i)$ for $p_i \in [0,1]$. The lemma follows.