Accepted Manuscript

Tractability of batch to sequential conversion

Marcus Hutter

PII: S0304-3975(18)30275-5
DOI: https://doi.org/10.1016/j.tcs.2018.04.037
Reference: TCS 11567

To appear in: *Theoretical Computer Science*

Please cite this article in press as: M. Hutter, Tractability of batch to sequential conversion, *Theoret. Comput. Sci.* (2018), https://doi.org/10.1016/j.tcs.2018.04.037

Tractability of Batch to Sequential Conversion☆

Marcus Hutter[a]

[a] *Research School of Computer Science, Australian National University, Canberra, ACT, 0200, Australia*

Abstract

We consider the problem of converting batch estimators into a sequential predictor or estimator with small extra regret. Formally this is the problem of merging a collection of probability measures over strings of length 1,2,3,... into a single probability measure over infinite sequences. We describe various approaches and their pros and cons on various examples. As a side-result we give an elementary non-heuristic purely combinatoric derivation of Turing's famous estimator. Our main technical contribution is to determine the computational complexity of sequential estimators with good guarantees in general. We conclude with an open problem on how to derive tractable sequential from batch estimators with good guarantees in general.

Keywords: offline, online, batch, sequential, probability, estimation, prediction, time-consistency, normalization, tractable, regret, combinatorics, Bayes, Laplace, Ristad, Good-Turing.

1. Introduction

A standard problem in statistics and machine learning is to estimate or learn an in general non-i.i.d. probability distribution $q_n:\mathcal{X}^n\to[0,1]$ from a batch of data $x_1,...,x_n$. q_n might be the Bayesian mixture over a class of distributions $\mathcal{M}$, or the (penalized) maximum likelihood (ML/MAP/MDL/MML) distribution from $\mathcal{M}$, or a combinatorial probability, or an exponentiated code length, or else. This is the offline or *batch* setting. An important problem is to predict x_{n+1} from $x_1,...,x_n$ sequentially for $n=0,1,2...$, called online or *sequential* learning if the predictor improves with n. A stochastic prediction $\tilde q(x_{n+1}|x_{1:n})$ can be useful in itself (e.g. weather forecasts), or be the basis for some decision, or be used for data compression via arithmetic coding, or otherwise. We use the prediction picture, but could have equally well phrased everything in terms of log-likelihoods, or perplexity, or code-lengths, or log-loss.

The naive predictor is $\tilde q^{\text{rat}}(x_{n+1}|x_1...x_n):=q_{n+1}(x_1...x_{n+1})/q_n(x_1...x_n)$ is not properly normalized to 1 if q_n and q_{n+1} are not compatible. We could fix the problem by normalization $\tilde q^{\text{n1}}(x_{n+1}|x_1...x_n):=\tilde q^{\text{rat}}(x_{n+1}|x_1...x_n)/\sum_{x_{n+1}}\tilde q^{\text{rat}}(x_{n+1}|x_1...x_n)$, but this may result in a very poor predictor. We discuss two further schemes, $\tilde q^{\text{lim}}$ and $\tilde q^{\text{mix}}$. Both are based on extending each q_n from $\mathcal{X}^n$ to $\bar q_n:\mathcal{X}^*\to[0;1]$ by marginalizing q_n for strings shorter than n and any compatible extension for longer strings. Then $\tilde q^{\text{lim}}:=\lim_{n\to\infty}\bar q_n$, which may not exist, and $\tilde q^{\text{mix}}:=\sum_{n=0}^\infty\frac{\bar q_n}{(n+1)(n+2)}$, which has excellent performance guarantees (small regret), but a direct computation of either is prohibitive.

A major open problem is to find a computationally tractable sequential predictor $\tilde q$ with provably good performance given batch probabilities (q_n). A positive answer would benefit many applications.

Applications. (i) Being able to use a batch estimator to make stochastic predictions (e.g. weather forecasts) is of course useful. The predictive probability needs to sum to 1 which $\tilde q^{\text{n1}}$ guarantees, but the regret should also be small, which only $\tilde q^{\text{mix}}$ guarantees.

(ii) Given a parameterized class of (already) sequential estimators $\{\tilde q^\theta\}$, estimating the parameter θ from data $x_1...x_n$ (e.g. maximum likelihood) for $n=1,2,3,...$ leads to a sequence of parameters $(\hat\theta_n)$ and a sequence

☆A shorter version appeared in the proceedings of the ALT 2014 conference [Hut14].

Email address: `http://www.hutter1.net/` (Marcus Hutter)