Empirical Bernstein Stopping

Volodymyr Mnih[1], Csaba Szepesvári[1], Jean-Yves Audibert[2]

[1]Department of Computing Science, University of Alberta
[2]Certis - Ecole des Ponts, Willow - ENS / INRIA

July 8, 2008
ICML 2008

Outline

- How to design efficient stopping rules.
- Two stopping problems.
- Two inefficient algorithms.
- Two efficient algorithms.
  - Variance estimation is the key.
- Theoretical and experimental results.

Problem I

- Given two poker players decide which one is better.
  - Make them play a lot of hands.
  - After how many hands can we determine the better player?
- Many similar problems.
  - Does a weak learner have error below 0.5?
  - Is an estimated gradient close to the true gradient?

More formally

- Let $X_1, X_2, X_3, \ldots$ be *i.i.d*, bounded random variables with mean $\mu \neq 0$, variance σ^2, and range R.
  - X_i can be payoff for Player 1 for the i^{th} hand.
  - If $\mu > 0$ Player 1 is better.
  - If $\mu < 0$ Player 2 is better.

- A *stopping rule* observes X_t at time t and decides whether to stop or keep sampling. After stopping an estimate is returned.

- For the poker problem, we want a stopping rule that determines the sign of μ with high probability.

(ϵ, δ)-approximations

- Usually we also want to know by how much the stronger player is better.
- We seek a stopping rule that, given ϵ and δ, returns an estimate $\hat{\mu}_T$ satisfying

$$\mathbb{P}\left[|\hat{\mu}_T - \mu| \leq \epsilon|\mu|\right] \geq 1 - \delta.$$

- We refer to $\hat{\mu}_T$ as an (ϵ, δ)-approximation.
- The stopping rule should stop as early as possible while returning a (ϵ, δ)-approximation.

Basic Stopping Criterion

- How can we find an (ϵ, δ)-approximation?
  - Let d_t be a sequence that sums to δ.
  - Let c_t be half the width of a $1 - d_t$ confidence interval for μ, then event

    $$\mathcal{E} = \{ |\overline{X}_t - \mu| \leq c_t,\ t \in \mathbb{N}^+ \}$$

    occurs with probability at least $1 - \delta$.
  - Stop when $c_t \leq \epsilon(|\overline{X}_t| - c_t)$.
  - Return $\overline{X}_t$.
- Nonmonotonic Adaptive Sampling (Domingo et al. 1999) uses Hoeffding's inequality to define

$$c_t = R\sqrt{\frac{\log(2/d_t)}{2t}}.$$

Nonmonotonic Adaptive Sampling

- One line proof:

$$|\overline{X}_t - \mu| \le c_t < \epsilon(|\overline{X}_t| - c_t) \le \epsilon|\mu|$$

- **Theorem [Domingo et al., 1999]:** If T is the stopping time of NAS, then for X with range R, there exists a universal constant c such that

$$\mathbb{E}[T] \le c \cdot \frac{R^2}{\mu^2\epsilon^2} \cdot \left(\log\frac{1}{\delta} + \log\frac{1}{\epsilon|\mu|}\right).$$

- Where is σ^2?

Nonmonotonic Adaptive Sampling

- One line proof:

$$|\overline{X}_t - \mu| \leq c_t < \epsilon(|\overline{X}_t| - c_t) \leq \epsilon|\mu|$$

- **Theorem [Domingo et al., 1999]:** If T is the stopping time of NAS, then for X with range R, there exists a universal constant c such that

$$\mathbb{E}[T] \leq c \cdot \frac{R^2}{\mu^2\epsilon^2} \cdot \left(\log\frac{1}{\delta} + \log\frac{1}{\epsilon|\mu|}\right).$$

- Where is σ^2?

Nonmonotonic Adaptive Sampling

- One line proof:

$$|\overline{X}_t - \mu| \le c_t < \epsilon(|\overline{X}_t| - c_t) \le \epsilon|\mu|$$

- **Theorem [Domingo et al., 1999]:** If T is the stopping time of NAS, then for X with range R, there exists a universal constant c such that

$$\mathbb{E}[T] \le c \cdot \frac{R^2}{\mu^2\epsilon^2} \cdot \left(\log\frac{1}{\delta} + \log\frac{1}{\epsilon|\mu|}\right).$$

- Where is σ^2?

Nonmonotonic Adaptive Sampling

- One line proof:

$$|\overline{X}_t - \mu| \leq c_t < \epsilon(|\overline{X}_t| - c_t) \leq \epsilon|\mu|$$

- **Theorem [Domingo et al., 1999]:** If T is the stopping time of NAS, then for X with range R, there exists a universal constant c such that

$$\mathbb{E}[T] \leq c \cdot \frac{R^2}{\mu^2\epsilon^2} \cdot \left(\log\frac{1}{\delta} + \log\frac{1}{\epsilon|\mu|}\right).$$

- Where is σ^2?

The $\mathcal{AA}$ Algorithm

- Dagum, Karp, Luby, and Ross (1999).
- $\mathcal{AA}$ algorithm for X in $[0, R]$.
  - Obtain $\tilde{\mu}$, an approximation of μ.
  - Obtain $\tilde{\sigma}^2$, an approximation of σ^2.
  - Draw $c \cdot \max(\frac{\tilde{\sigma}^2}{\epsilon^2 \tilde{\mu}^2}, \frac{1}{\epsilon \tilde{\mu}}) \log(2/\delta)$ expected number of samples and return the sample mean as $\hat{\mu}$.
- For appropriate c, $\hat{\mu}$ is an (ϵ, δ)-approximation of μ.

The $\mathcal{AA}$ Algorithm - Bounds

- **Theorem [Dagum et al., 1995]:** If T is the number of samples taken by $\mathcal{AA}$, then exists $c > 0$ such that for all X

$$\mathbb{E}[T] \leq C \cdot \max\left(\frac{\sigma^2}{\epsilon^2\mu^2}, \frac{R}{\epsilon\mu}\right) \cdot \log \frac{2}{\delta}.$$

- R^2 is replaced by σ^2 and R. Much better than NAS.

The $\mathcal{AA}$ Algorithm - Bounds

- **Theorem [Dagum et al., 1995]:** If T is the number of samples taken by $\mathcal{AA}$, then exists $c > 0$ such that for all X

$$\mathbb{E}[T] \leq C \cdot \max\left(\frac{\sigma^2}{\epsilon^2\mu^2}, \frac{R}{\epsilon\mu}\right) \cdot \log\frac{2}{\delta}.$$

- R^2 is replaced by σ^2 and R. Much better than NAS.
- **Theorem [Dagum et al., 1995]:** Any stopping rule that returns an (ϵ, δ)-approximation must take a number of samples the order of

$$\max\left(\frac{\sigma^2}{\epsilon^2\mu^2}, \frac{R}{\epsilon\mu}\right) \cdot \log\frac{2}{\delta}.$$

Extending $\mathcal{AA}$

- $\mathcal{AA}$ only works for nonnegative X.
  - Cannot be applied to the poker problem.
- Can we easily extend it to the bounded case?
  - Shifting X to be nonnegative is not what we want.
  - Taking the absolute values of X is not what we want.
  - $\mathcal{AA}$ heavily relies on monotonicity of partial sums.
- No trivial extension to the bounded case seems possible.

EBStop (I)

- EBStop builds on the basic stopping criterion.
- First improvement: Construct $\{c_t\}$ with empirical Bernstein bounds.
- The empirical Bernstein bound (Audibert, Munos, Szepesvári, 2007) states that with probability at least $1 - \delta$,

$$|\overline{X}_t - \mu| \leq \overline{\sigma}_t \sqrt{\frac{2 \log(3/\delta)}{t}} + \frac{3R \log(3/\delta)}{t}.$$

- We also use $d_t = \frac{c\delta}{t^p}$ for $p > 1$.

EBStop (II)

- Second improvement: We can stop as soon as $c_t \le \epsilon|\overline{X}_t|$ instead of $c_t \le \epsilon\left(|\overline{X}_t| - c_t\right)$.
  - Let $\mathrm{LB}(t) = \max(0, |\overline{X}_t| - c_t)$ and $\mathrm{UB}(t) = |\overline{X}_t| + c_t$.
  - Stop when $(1+\epsilon)\mathrm{LB}(t) \ge (1-\epsilon)\mathrm{UB}(t)$.
  - Return $\hat{\mu} = \mathrm{sgn}(\overline{X}_t) \cdot 1/2 \cdot \left[(1+\epsilon)\mathrm{LB}(t) + (1-\epsilon)\mathrm{UB}(t)\right]$.
- The stopping conditions are equivalent and $\hat{\mu}$ is an (ϵ, δ)-approximation of μ.

EBStop - Upper bound

- **Theorem:** If T is the stopping time of EBStop, then there exists a universal constant C such that

$$\mathbb{E}[T] \le C \cdot \max\left(\frac{\sigma^2}{\epsilon^2\mu^2}, \frac{R}{\epsilon|\mu|}\right)\left(\log\frac{1}{\delta} + \log\frac{1}{\epsilon|\mu|}\right)$$

- Hence, EBStop comes close to $\mathcal{AA}$'s bound, but only needs bounded X_i.
- Can we reduce or get rid of the $\log\frac{1}{\epsilon|\mu|}$ term?

EBGStop - EBStop with Geometric Sampling

- Check the stopping condition only after $\lceil \beta^k \rceil$ samples for $k \in \mathbb{N}^+$ and some $\beta > 1$.
- Using fewer deviation bounds leads to tighter confidence intervals and earlier stopping.

EBGStop - EBStop with Geometric Sampling

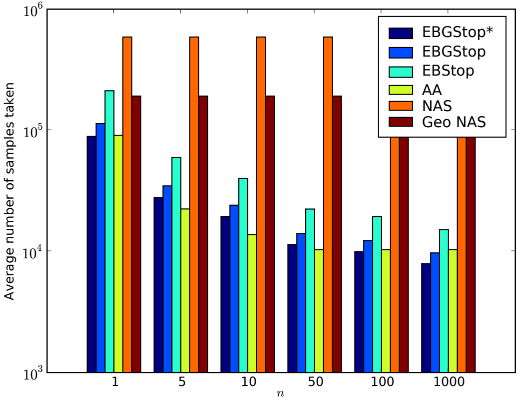
- Check the stopping condition only after $\lceil \beta^k \rceil$ samples for $k \in \mathbb{N}^+$ and some $\beta > 1$.
- Using fewer deviation bounds leads to tighter confidence intervals and earlier stopping.
- If T is the stopping time of EBGStop, then there exists a universal constant C such that

$$\mathbb{E}[T] \leq C \cdot \max\left(\frac{\sigma^2}{\epsilon^2\mu^2}, \frac{R}{\epsilon|\mu|}\right)\left(\log\frac{1}{\delta} + \log\log\frac{1}{\epsilon|\mu|}\right)$$

- We can also achieve this bound while testing after every t.

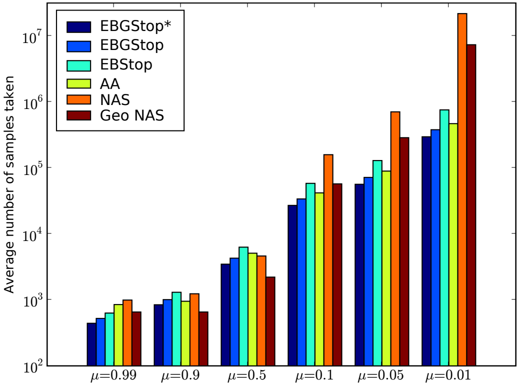
Results - Effect of Variance

- Stopping times for finding (0.01, 0.1)-approximations of averages of n Uniform(0,1) random variables.



Results - Bernoulli Random Variables

- Stopping times for finding (0.1, 0.1)-approximations of Bernoulli means.



Problem II - Picking the winner

- You have a number of predictors.
- Want to decide which one is the best quickly.
  - Test each one on a holdout set.
  - Pick the one with the highest accuracy/average reward/likelihood.
- It is possible to save a lot of time by using a stopping rule.
  - Stop evaluating a predictor as soon as it is clear that it is bad.

More formally - Racing algorithms

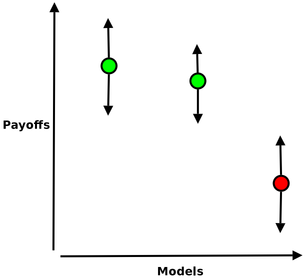
- Given: M options, N data points, confidence parameter $\delta > 0$.
- At time t
  - Receive data point D_t.
  - Can choose to compute $X_{m,t}$, the payoff for option m on D_t.
  - Can choose to discard any option(s).
  - $\{X_{m,t}\}_{t \geq 1}$ are $i.i.d$ with mean μ_m and range R.
- A racing algorithm terminates when
  - It has found the best option with probability at least $1 - \delta$, or
  - It has received all N data points.
- The goal is to keep the best option and compute much fewer than MN payoffs.

Hoeffding Races (Maron and Moore, 1994)

- At time t build $1 - \delta/MN$ confidence interval for each μ_m using Hoeffding's inequality.

$$\left[\overline{X}_{m,t} - R\sqrt{\frac{\log(2MN/\delta)}{2t}}, \overline{X}_{m,t} + R\sqrt{\frac{\log(2MN/\delta)}{2t}}\right]$$

- If the upper confidence of option j is smaller than the lower confidence of any option, discard option j.
- Illustration:



Hoeffding Races - Bounds

- **Theorem:** The number of samples taken by the Hoeffding Race algorithm is bounded from above by

$$\sum_{\mu_m < \mu_{m*}} \left\lceil \frac{8R^2 \log(2MN/\delta)}{(\mu_m - \mu_{m*})^2} \right\rceil.$$

- Dependence on R^2 and no σ^2.

Empirical Bernstein Races

- Simple improvement: Use empirical Bernstein bounds instead of Hoeffding bounds to build confidence intervals.
- **Theorem:** The number of samples taken by the EBRace algorithm is bounded from above by

$$\sum_{\mu_m < \mu_{m*}} \left\lceil \frac{8(\sigma_m + \sigma_{m*})^2 + 18R(\mu_{m*} - \mu)}{(\mu_{m*} - \mu_m)^2} \log(4MN/\delta) \right\rceil .$$

- Dependence on R^2 traded for dependence on R and σ_m^2.

RL & AI

Racing Algorithms - Results

- Comparison on the task of selecting the best k for nearest neighbor regression and classification through leave-one-out cross-validation.
- Started with 11 models/values of k.
- We show percentage of tests saved over the MN required by brute force and the number of models left.

Data set	Hoeffding	EB
SARCOS	0.0% / 11	44.9% / 4
Covertype2	14.9% / 8	29.3% / 5
Local	6.0% / 9	33.1% / 6

Conclusions

- Empirical Bernstein bounds can be used to design efficient stopping rules.
- When used in place of Hoeffding's inequality:
  - Linear dependence on R^2 in sample complexity is reduced to a linear dependence on σ^2 and R.
  - Can offer huge computational savings in practice.