

Hierarchical Polynomial Approximation

Vincent LEFÈVRE, Jean-Michel MULLER, Serge TORRES

Arénaire, INRIA Grenoble – Rhône-Alpes / LIP, ENS-Lyon

Journées TaMaDi, Lyon, 2011-12-13

Outline

- Two-Level Polynomial Approximations
- Computation of the Coefficients of $a_j(k)$
- Computation of Consecutive Values of a Polynomial
- Error Bound on the Approximation of P_k by Q_k
- Summary for the Search for HR-Cases

Two-Level Polynomial Approximations

The problem: **After scaling, values of $f(0), f(1), f(2), \dots, f(T_I - 1)$ on a large interval I , of width T_I ?**

- On I , f is approximated by a degree- d polynomial P , with an approximation error bounded by ϵ .
- I is split into sub-intervals J_0, J_1, J_2, \dots of width T_J , where $J_k = J_{k-1} + T_J$.
- On each of the J_k , we are looking for an approximation Q_k of degree $\delta < d$.

$$P_k(X) = P(X + k \cdot T_J) = a_0(k) + a_1(k) \cdot X + a_2(k) \cdot \frac{X(X-1)}{2} + \dots + a_d(k) \cdot \frac{X(X-1)(X-2) \cdots (X-d+1)}{d!}$$

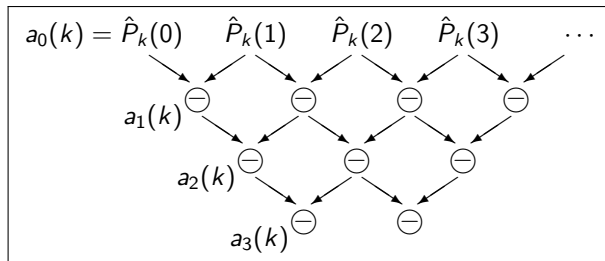
$$Q_k(X) = a_0(k) + a_1(k) \cdot X + a_2(k) \cdot \frac{X(X-1)}{2} + \dots + a_\delta(k) \cdot \frac{X(X-1)(X-2) \cdots (X-\delta+1)}{\delta!} = \sum_{i=0}^{\delta} a_i(k) \cdot \binom{X}{i}$$

Computation of the Coefficients of $a_i(k)$

The problem: **Compute the values of $a_i(k)$.**

- For each $i \in \llbracket 0, \delta \rrbracket$, $a_i(k)$ is a degree- $(d - i)$ polynomial function of k .
- We can compute the $a_i(k)$ from the $d - i$ consecutive values $P_k(0)$, $P_k(1)$, $P_k(2)$, \dots , $P_k(d - i - 1)$ for the first values of k ; the following $a_i(k)$'s are computed with the *difference table method*.
- Let $P_k(u)$ be the “true” value, and $\hat{P}_k(u)$ be the “computed” value.

For $0 \leq k \leq d - i$:



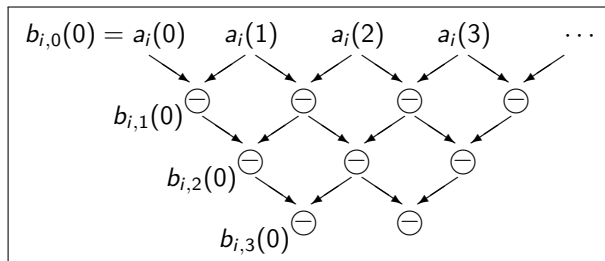
Let e be an upper bound on $|P_k(u) - \hat{P}_k(u)|$.

With exact subtractions, $a_i(k)$ will be known with an error $\leq 2^i e$.

Computation of the Coefficients of $a_i(k)$ [2]

Define values $b_{i,j}(k)$ as

$$a_i(k) = b_{i,0}(k) + b_{i,1}(k) \cdot k + b_{i,2}(k) \cdot \frac{k(k-1)}{2} + \dots + b_{i,d-i}(k) \cdot \frac{k(k-1)(k-2)\dots(k-d+i+1)}{(d-i)!} = \sum_{j=0}^{d-i} b_{i,j}(k) \cdot \binom{k}{j}$$



The term $b_{i,j}(0)$ is known with an error $\leq 2^{i+j}e$.

The $a_i(k)$ will be computed with additions from these coefficients $b_{i,j}(0)$...

Computation of Consecutive Values of a Polynomial

Let $Q(X + k) = \sum_{i=0}^{\delta} q_i(k) \cdot \binom{X}{i}$ of degree (at most) δ , where the $q_i(k)$'s are assumed to be the exact coefficients.

The problem: **Knowing the initial coefficients $q_i(0)$ of $Q(X)$, compute the consecutive values $Q(0), Q(1), Q(2), \dots$**

$$\text{We have: } \begin{cases} q_0(k) &= q_0(k-1) &+ q_1(k-1) \\ q_1(k) &= q_1(k-1) &+ q_2(k-1) \\ &\vdots \\ q_{\delta-1}(k) &= q_{\delta-1}(k-1) &+ q_{\delta}(k-1) \end{cases}$$

q_{δ} being a constant, and $Q(k) = q_0(k)$.

The coefficients $q_i(k)$ will be represented by $\hat{q}_i(k)$ with n_i bits after the fractional point, i.e. on n_i bits as we are interested in the values modulo 1, and an initial error of one ulp: $u_i = 2^{-n_i}$. Since $q_i(k)$ depends on $q_{i+1}(k-1)$, we assume that (n_i) is increasing: $n_i \leq n_{i+1}$. And $n_{\delta} = n_{\delta-1}$ for the constant coefficient q_{δ} (using more precision would be useless).

Computation of Consecutive Values of a Polynomial [2]

Formally: $n_i \leq n_{i+1}$, $u_i = 2^{-n_i}$, $\hat{q}_i(k) \in u_i\mathbb{Z}$ or $u_i\mathbb{Z}/\mathbb{Z}$, and $|\hat{q}_i(0) - q_i(0)| \leq u_i$.

$$\text{Basic iteration: } \begin{cases} \hat{q}_0(k) &= \hat{q}_0(k-1) &+ \circ(\hat{q}_1(k-1)) \\ \hat{q}_1(k) &= \hat{q}_1(k-1) &+ \circ(\hat{q}_2(k-1)) \\ &\vdots \\ \hat{q}_{\delta-2}(k) &= \hat{q}_{\delta-2}(k-1) &+ \circ(\hat{q}_{\delta-1}(k-1)) \\ \hat{q}_{\delta-1}(k) &= \hat{q}_{\delta-1}(k-1) &+ \hat{q}_{\delta} \end{cases}$$

where $|\hat{q}_{\delta} - q_{\delta}| \leq u_{\delta-1}$ and \circ denotes the truncation of the value to the precision of the result (said otherwise, when doing an addition, the trailing bits of the more precise value are ignored).

Let $\epsilon_i(k)$ be a bound on the error on $q_i(k)$, i.e. $|\hat{q}_i(k) - q_i(k)| \leq \epsilon_i(k)$.

Initially, $\epsilon_i(0) = u_i$ for $0 \leq i \leq \delta - 1$.

We have: $\epsilon_i(k) = \epsilon_i(k-1) + \epsilon_{i+1}(k-1) + u_i$ for $0 \leq i \leq \delta - 1$,
with $\epsilon_{\delta}(0) = 0$ in order to satisfy $\epsilon_{\delta-1}(k) = \epsilon_{\delta-1}(k-1) + u_{\delta-1}$.

We can prove by induction that $\epsilon_i(k) = \sum_{j=i}^{\delta-1} u_j \cdot \binom{k+1}{j-i+1}$.

Computation of Consecutive Values of a Polynomial [3]

Proof by induction that $\epsilon_i(k) = \sum_{j=i}^{\delta-1} u_j \cdot \binom{k+1}{j-i+1}$.

This is true for $k = 0$ and for $i = \delta$, and

$$\begin{aligned}\epsilon_i(k-1) + \epsilon_{i+1}(k-1) + u_i &= u_i + \sum_{j=i+1}^{\delta-1} u_j \cdot \binom{k}{j-i} + \sum_{j=i}^{\delta-1} u_j \cdot \binom{k}{j-i+1} \\ &= \sum_{j=i}^{\delta-1} u_j \cdot \left[\binom{k}{j-i} + \binom{k}{j-i+1} \right] \\ &= \sum_{j=i}^{\delta-1} u_j \cdot \binom{k+1}{j-i+1} = \epsilon_i(k)\end{aligned}$$

Computation of Consecutive Values of a Polynomial [4]

If ℓ denotes the length of the interval (the number of values), i.e. $0 \leq k < \ell$, then the error is bounded by:

$$\max_{0 \leq k < \ell} \epsilon_0(k) \leq \sum_{i=0}^{\delta-1} u_i \cdot \binom{\ell}{i+1} = \sum_{i=0}^{\delta-1} 2^{-n_i} \cdot \binom{\ell}{i+1}.$$

The values of n_i will be determined so that this error bound is less than some given error bound E . This can be done by determining n_0 , then n_1 , then n_2 , and so on, each time by taking the smallest n_i (multiple of the word size) such that $2^{-n_i} \cdot \binom{\ell}{i+1}$ is less than a fraction of the remaining error bound.

Note: if $n_{i+1} = n_i$, one could have taken $\epsilon'_i(k) = \epsilon_{i+1}(k)$, but the gain would probably be low (since $\epsilon_{i+1}(k) \sim k \cdot u_{i+1} = k \cdot u_i$ at least) and the formulas would probably be much more complicated.

Error Bound on the Approximation of P_k by Q_k

On the interval J_k : $P_k(m) - Q_k(m) = \sum_{i=\delta+1}^d a_i(k) \cdot \binom{m}{i}$

with $a_i(k) = \Delta^i P(kT_J) = \sum_{j=i}^d a_j(0) \cdot \binom{kT_J}{j-i}$.

Thus

$$P_k(m) - Q_k(m) = \sum_{j=\delta+1}^d a_j(0) \sum_{i=\delta+1}^j \binom{kT_J}{j-i} \binom{m}{i}$$

with $0 \leq m \leq T_J - 1$ and $0 \leq kT_J \leq T_I - T_J$. This gives the following error bound on the approximation of P_k by Q_k :

$$\sum_{j=\delta+1}^d |a_j(0)| \sum_{i=\delta+1}^j \binom{T_I - T_J}{j-i} \binom{T_J - 1}{i}$$

but one can get better dynamical bounds by considering the sign of $a_j(0)$.

Summary for the Search for HR-Cases

- 1 Determine the exponent range of f ; f will be scaled by a power of the radix, so that we are interested in the values modulo 1: $\tilde{f} = \beta^s \cdot f$.
- 2 Determine the admissible final error bound ε_0 on \tilde{f} .
- 3 Assume we have a degree- d polynomial approximation $P(m)$ to $\tilde{f}(x)$ with an error bound ε_f , where $x = x_0 + m \cdot d_x$ (the evaluation error of the polynomial is not taken into account here).
- 4 $P_k(X) = P(X + kT_J)$ will be approximated by degree- δ polynomials Q_k . The coefficients $a_i(k)$ of Q_k are in fact polynomials in k of degree $d - i$, where $0 \leq i \leq \delta$. The initial coefficients $b_{i,j}(0)$ of these polynomials $a_i(k)$ in the binomial base are computed with an error bounded by $2^{i+j}e$, where e is a bound on the evaluation error of the polynomial. For the ulp error condition, we want $2^{i+j}e \leq 2^{-n_j-1}$, thus $e \leq 2^{-i-j-n_j-1}$, where the values of the n_j 's can be determined so that the error on $a_i(k)$ is less than some given bound:

$$\sum_{j=0}^{d-i-1} 2^{-n_j} \cdot \binom{T_I/T_J}{j+1} \leq E_i$$