A Metric for Parametric Approximation

T. Lyche and K. Mrken

Abstract. We define a metric on the set of parametric curves that is more convenient for computational purposes than the Hausdorff metric. We then give a definition of approximation rate for parametric approximation schemes in terms of this metric, and present a simple family of odd degree parametric polynomial approximations to circle segments with approximation order twice the degree of the polynomial.

§1. Introduction

Most traditional methods for parametric curve approximation are based on approximation of functions. A parametric curve is approximated by applying standard approximation schemes for functions to the component functions of the curve. Recently, there has been increasing interest in what we call *parametric* methods, where the geometric properties of the curve are utilized for approximation purposes. One advantage of such schemes is that for polynomial approximants of fixed degree, they often provide higher approximation rates than the traditional schemes; see [1] for an early result in this direction and [3] for a good bibliography.

In order to achieve such high approximation orders, it is of fundamental importance that the error is measured suitably. In Section 2, we define a metric on the set of parametric curves which will provide us with a family of error measures suitable for our purposes. In particular, this metric leads to a natural definition of (parametric) approximation order, see [3] for a similar definition.

In earlier papers, we have studied approximation of circle segments by quadratic and cubic polynomial curves which are fourth and sixth order accurate, see [4,5]. In Section 3, we generalize this to higher degrees.

Curves and Surfaces II
P. J. Laurent, A. Le Méhauté, and L. L. Schumaker (eds.), pp. 1–9.

ISBN 0-12-XXXX.

§2. A Metric for Parametric Curves

Let $\mathbf{f} = (f_1, f_2)$ be a parametric curve in $\mathbb{R}^2$. A typical way of computing an approximation $\mathbf{p} = (p_1, p_2)$ to $\mathbf{f}$ is to let p_1 and p_2 be approximations to f_1 and f_2, respectively. The obvious advantage of this approach is that we can make use of the extensive theoretical and practical knowledge accumulated for approximation of functions. We can for instance let p_1 and p_2 be the least squares approximations to f_1 and f_2, or we can determine p_1 and p_2 by interpolating f_1 and f_2 at suitable points. In Figure 1 we have approximated the half circle $\mathbf{f}(t) = (\cos t, \sin t)$ for $t \in [0, \pi]$ by letting p_1 and p_2 be the cubic polynomials that interpolate $\cos t$ and $\sin t$ and their first derivatives at the end points $t = 0$ and $t = \pi$. This interpolant is shown with short dashes. The curve with longer dashes is a cubic interpolant to a reparametrization of the half circle; it interpolates $\mathbf{f}\big(\phi(t)\big)$ (with $\phi(0) = 0$, $\phi(\pi) = \pi$ and $\phi'(t) > 0$) and its first *two* derivatives at $t = 0$ and $t = \pi$. The lines from the approximations to the half circle connect points with the same parameter value.

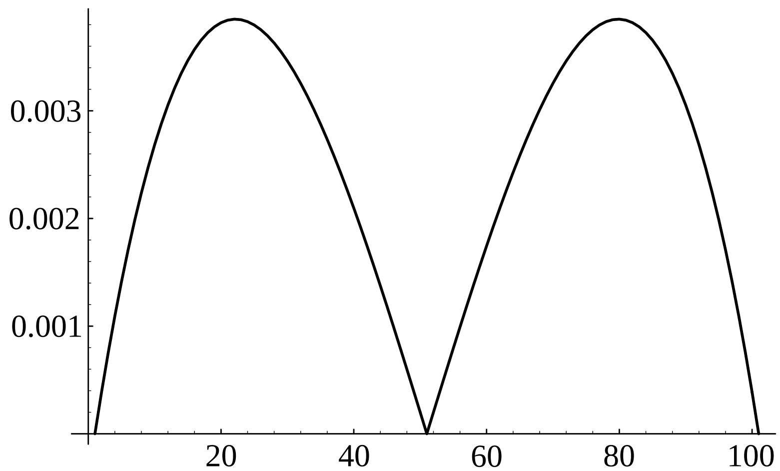

Figure 1. Two cubic Hermite interpolants approximating a half circle.

Figure 1 illustrates the fact that traditional approximation methods for curves often lead to inefficient parametrizations, which again means that we get an overestimate of the error when we compute it by comparing the values at common values of the parameter. By allowing the curve to be reparametrized before approximation, we can both get better approximations and easily compute better error estimates.

In this section we will use the idea of reparametrization to define a suitable metric on the set of parametric curves. But let us first define precisely what we mean by parametric curves. The material to follow in the next three definitions is quite standard and is included only for the convenience of the reader.

Definition 1. *A regular parametric representation* $\mathbf{f}$ *in* $\mathbb{R}^d$ *is a mapping of a real interval* $I = [a, b]$ *into* $\mathbb{R}^d$ *such that* $\mathbf{f}'(t) \neq 0$ *for all* $t \in I$.

To generate reparametrizations of $\mathbf{f}$, we roughly compose it with increasing, real functions. To be more precise, let $I_1 = [a, b]$ and $I_2 = [c, d]$ be two real intervals. The set A_{I_1,I_2} is defined by

$$A_{I_1,I_2} = \{\phi \in C^1(I_1) \mid \phi(I_1) = I_2 \text{ and } \phi'(t) > 0 \text{ for } t \in I_1\},$$

the set of increasing functions on I_1 with continuous derivative that map I_1 onto I_2. In the following, symbols like I, I_1, I_2 will always denote real intervals. When I_1 and I_2 are irrelevant or obvious from the context we will often abbreviate A_{I_1,I_2} to A. A function $\phi \in A_{I_1,I_2}$ is often called an allowable parameter change from I_1 to I_2.

Definition 2. *A regular reparametrization of a parametric representation $\mathbf{f} : I_1 \mapsto \mathbb{R}^d$ is a regular parametric representation $\mathbf{g} : I_2 \mapsto \mathbb{R}^d$ such that $\mathbf{g}\big(\phi(t)\big) = \mathbf{f}(t)$ for some function $\phi \in A_{I_1,I_2}$.*

Definition 3. *A regular parametric curve is the equivalence class of all reparametrizations of a parametric representation $\mathbf{f}$, i.e., the set $\{\mathbf{f} \circ \phi \mid \phi \in A\}$.*

If $\mathbf{f}$ is defined on I_1, then $\phi \in A$ in Definition 3 means $\phi \in \cup_{I_2} A_{I_1,I_2}$, where the union is over all subintervals I_2 of $\mathbb{R}$. Observe however that there is no essential loss in only letting ϕ vary in A_{I_1,I_1}.

In line with common practice, we will usually identify a particular parametric representation with the curve it represents.

In order to discuss errors and approximation rates for curves, we need some way to measure the distance between two curves. An easy way to do this when the two curves $\mathbf{f}$ and $\mathbf{g}$ are parametrized over the same interval is to compare $\mathbf{g}(t)$ with $\mathbf{f}(t)$ and compute some norm of the difference between the two. As indicated above, this is not in general a particularly good approach.

We already have a metric for measuring the distance between two general sets, the Hausdorff metric. If we specialize this to curves, we find that the distance between $\mathbf{f}$ and $\mathbf{g}$ (parametrized on the intervals I_1 and I_2 respectively) is given by

$$d_H(\mathbf{f}, \mathbf{g}) = \max\Big\{\max_{t\in I_1} \min_{s\in I_2} ||\mathbf{f}(t) - \mathbf{g}(s)||, \quad \max_{s\in I_2} \min_{t\in I_1} ||\mathbf{f}(t) - \mathbf{g}(s)||\Big\},$$

with $||\cdot||$ some vector norm in $\mathbb{R}^d$. This metric is however not very appropriate for dealing with curves, and it is complicated to compute, see [2].

Let us consider the expression $\max_{t\in I_1} \min_{s\in I_2} ||\mathbf{f}(t) - \mathbf{g}(s)||$ in more detail. Fix t, and let $\phi(t)$ denote an s for which the minimum is attained; then we have

$$\max_{t\in I_1} \min_{s\in I_2} ||\mathbf{f}(t) - \mathbf{g}(s)|| = \max_{t\in I_1} \Big\|\mathbf{f}(t) - \mathbf{g}\big(\phi(t)\big)\Big\|. \tag{1}$$

The function $\phi(t)$ is in general neither monotone nor continuous even if $\mathbf{f}$ and $\mathbf{g}$ are well behaved. However, keeping in mind the definition of a parametric curve, the right hand side of (1) suggests the following metric.

Definition 4. *Let* $\mathbf{f}$ *and* $\mathbf{g}$ *be two parametric curves defined on the intervals* I_1 *and* I_2*. The parametric distance between* $\mathbf{f}$ *and* $\mathbf{g}$ *is defined by*

$$d_P(\mathbf{f},\mathbf{g}) = d_P(\mathbf{f},\mathbf{g})_{I_1} = \inf_{\phi \in A_{I_1,I_2}} \max_{t \in I_1} \big\| \mathbf{f}(t) - \mathbf{g}\big(\phi(t)\big) \big\|.$$

At a first glance this definition looks unsymmetric. But it is not too hard to see that we also have

$$d_P(\mathbf{f},\mathbf{g})_{I_1} = \inf_{\psi \in A_{I_2,I_1}} \max_{s \in I_2} \big\| \mathbf{f}\big(\psi(s)\big) - \mathbf{g}(s) \big\| = d_P(\mathbf{g},\mathbf{f})_{I_2}. \tag{2}$$

$>$From this it is clear that

$$d_P(\mathbf{f},\mathbf{g}) \ge d_H(\mathbf{f},\mathbf{g}).$$

Let us make sure that d_P really is a metric for parametric curves.

Proposition 5. *The function* d_P *is a metric on the set of parametric curves in* $\mathbb{R}^d$*.*

Proof: Observe first that d_P is independent of the particular parametric representations we have picked for $\mathbf{f}$ and $\mathbf{g}$. If $\mathbf{f}_1 = \mathbf{f} \circ \phi$ and $\mathbf{g}_1 = \mathbf{g} \circ \psi$ are two allowable reparametrizations of $\mathbf{f}$ and $\mathbf{g}$, we have

$$d_P(\mathbf{f}_1,\mathbf{g}_1) = d_P(\mathbf{f} \circ \phi, \mathbf{g} \circ \psi) = d_P(\mathbf{f}, \mathbf{g} \circ \psi \circ \phi^{-1}) = d_P(\mathbf{f},\mathbf{g}).$$

The inverse ϕ^{-1} of ϕ exists since $\phi'(t) > 0$ for all t. The last equality follows since $\psi \circ \phi^{-1} \circ \xi$ generates all allowable parameter changes when ξ varies over A.

Let us next check the axioms for metric spaces. Nonnegativity of d_P is trivial. The symmetry $d_P(\mathbf{f},\mathbf{g}) = d_P(\mathbf{g},\mathbf{f})$ follows from (2). If $\mathbf{g}$ is a reparametrization of $\mathbf{f}$, we clearly have $d_P(\mathbf{f},\mathbf{g}) = 0$. To prove the opposite, suppose that $\mathbf{f}$ and $\mathbf{g}$ are two given curves with $d_P(\mathbf{f},\mathbf{g}) = 0$. Then there is a sequence of mappings $\phi_n \in A$ such that

$$\lim_{n \to \infty} \max_t \|\mathbf{f}(t) - \mathbf{g}(\phi_n(t))\| = 0. \tag{3}$$

Then there must be a ϕ such that $\mathbf{f} = \mathbf{g} \circ \phi$. A priori, we know nothing about this ϕ; we will first show that it is unique if we require it to be continuous.

Let $\mathbf{g}(s) = \mathbf{f}(\phi(t))$ be a point that does not intersect any other points of $\mathbf{g}$. To any given $\epsilon > 0$ we can then find a $\delta_1 > 0$ such that if $\|\mathbf{g}(s) - \mathbf{g}(s_1)\| < \delta_1$, then we must have $|s - s_1| < \epsilon$, where $s_1 = \phi(t_1)$. But $\mathbf{g} \circ \phi = \mathbf{f}$ is continuous, so there is some $\delta > 0$ such that if $|t - t_1| < \delta$, then $\|\mathbf{g}(\phi(t)) - \mathbf{g}(\phi(t_1))\| < \delta_1$. From this we conclude that if $|t - t_1| < \delta$, then $|\phi(t) - \phi(t_1)| < \epsilon$, in other words ϕ is continuous at t.

If $\mathbf{g}$ is not simple, e.g., it has one loop so that $\mathbf{g}(s_1) = \mathbf{g}(s_2)$ for $s_1 < s_2$, we conclude from the above that ϕ must be continuous in each open subinterval where it is simple. Since $\mathbf{g} \circ \phi = \mathbf{f}$, there must also be t_1 and t_2 such that $\mathbf{f}(t_1) = \mathbf{f}(t_2)$ and

$$\phi\big((a, t_1)\big) = (a, s_1), \qquad \phi\big((t_1, t_2)\big) = (s_1, s_2), \qquad \phi\big((t_2, b)\big) = (s_2, b).$$

Since $\mathbf{f}(t_1) = \mathbf{f}(t_2) = \mathbf{g}(s_1) = \mathbf{g}(s_2)$, we see that the convergence in (3) is not sufficiently strong for ϕ to distinguish the values s_1 and s_2. But then we can *choose* ϕ such that $\phi(t_1) = s_1$ and $\phi(t_2) = s_2$ which makes it continuous. The case where $\mathbf{g}$ has several loops is similar.

It remains to prove that ϕ has a continuous derivative. Recall from elementary analysis that if $\lim x_n y_n = z$ and $\lim x_n = x$, then if $x \neq 0$ the sequence (y_n) is also convergent and converges to z/x. Consider now the expression

$$\begin{aligned} \mathbf{f}'(t) &= \lim_{h \to 0} \frac{\mathbf{g}(\phi(t+h)) - \mathbf{g}(\phi(t))}{h} \\ &= \lim_{h \to 0} \frac{\mathbf{g}(\phi(t+h)) - \mathbf{g}(\phi(t))}{\phi(t+h) - \phi(t)} \cdot \frac{\phi(t+h) - \phi(t)}{h}. \end{aligned} \tag{4}$$

We know that the first quotient on the right converges to $\mathbf{g}'(\phi(t))$ which we have assumed is nonzero for all t. Therefore we can apply the result quoted above to a nonzero component of (4) to conclude that $\lim_{h \to 0}(\phi(t+h) - \phi(t))/h = \phi'(t)$ exists. From this we also obtain that ϕ' must be continuous since it is the quotient of two continuous functions. We therefore have $\mathbf{f} = \mathbf{g} \circ \phi$ for a unique $\phi \in A$.

To prove the triangle inequality, let $\mathbf{f}$, $\mathbf{g}$ and $\mathbf{h}$ be parametric curves, fix ψ and ϕ in A and consider the inequality

$$\max_t \|\mathbf{f}(t) - \mathbf{g}(\phi(t))\| \leq \max_t \|\mathbf{f}(t) - \mathbf{h}(\psi(t))\| + \max_t \|\mathbf{h}(\psi(t)) - \mathbf{g}(\phi(t))\|.$$

Since the left hand side is independent of ψ, for any given $\epsilon > 0$, we can find a $\psi_0 \in A$ so that

$$\|\mathbf{f}(t) - \mathbf{g}(\phi(t))\| \leq d_P(\mathbf{f}, \mathbf{h}) + \|\mathbf{h}(\psi_0(t)) - \mathbf{g}(\phi(t))\| + \epsilon.$$

Taking the inf over ϕ, we can find a $\phi_0 \in A$ such that

$$\|\mathbf{f}(t) - \mathbf{g}(\phi_0(t))\| \leq d_P(\mathbf{f}, \mathbf{h}) + d_P(\mathbf{h}, \mathbf{g}) + 2\epsilon.$$

Since $d_P(\mathbf{f}, \mathbf{g})$ is less than or equal to the left hand side and ϵ is arbitrary, the triangle inequality follows. ∎

Note that we have

$$d_P(\mathbf{f}, \mathbf{g}) \leq \delta_{\phi,\psi}(\mathbf{f}, \mathbf{g}) := \max_t \|\mathbf{f}(\psi(t)) - \mathbf{g}(\phi(t))\|$$

for arbitrary ϕ and ψ in A. In practice, one therefore often estimates $d_P(\mathbf{f}, \mathbf{g})$ by giving some ψ and ϕ and then using the above inequality.

Our main interest is in approximation of arbitrary curves with polynomial curves. The parametric metric d_P is then a convenient tool for measuring the error in the approximation. Of particular interest in approximation theory is the approximation order of an approximation scheme.

Definition 6. *Let S_h be an approximation scheme that to each parametric curve $\mathbf{f} \in \mathbb{R}^d$ defined on an interval $[a, b]$ and for each $h < b - a$ assigns an approximation $S_h(\mathbf{f})$ to the part of $\mathbf{f}$ defined on $I_h[a, a+h]$. Then S_h is said to have approximation order m if the inequality*

$$d_P(\mathbf{f}, S_h(\mathbf{f}))_{I_h} \le Ch^m,$$

holds for some constant C independent of h.

To establish that a scheme has approximation order m, it will be sufficient to show that

$$\max_t \|\mathbf{f}(\phi(t)) - S_h(\mathbf{f})(t)\| \le Ch^m,$$

for some C independent of ϕ and h, see also [3].

§3. Parametric Approximation of Circle Segments

In two earlier papers, we have studied approximation of circle segments by parametric quadratic and cubic polynomials. In particular, it was shown in [5] that there is a quadratic polynomial curve that approximates a segment of angular width α with error proportional to α^4. In [4] it was shown that if we use cubic polynomials, then there are approximations that give an error which is proportional to α^6. The following result introduces a high order approximation to circle segments for any odd degree n. Its relation to the metric of the previous section is discussed below.

Lemma 7. *Let n be a positive, odd integer and define the two functions $x_n(t)$ and $y_n(t)$ by*

$$\begin{aligned} x_n(t) &= 2\sum_{i=1}^{(n-1)/2} (-1)^{i-1} t^{2i-1} + (-1)^{(n-1)/2} t^n \\ y_n(t) &= 1 - 2\sum_{i=1}^{(n-1)/2} (-1)^{i-1} t^{2i}. \end{aligned} \tag{5}$$

Then the relation

$$x_n(t)^2 + y_n(t)^2 = 1 + t^{2n} \tag{6}$$

holds for all t in $\mathbb{R}$.

Proof: Summing the geometric series in (5), we find

$$x_n(t) = x_0(t) - (-1)^{(n-1)/2} t^n y_0(t) \quad \text{and} \quad y_n(t) = y_0(t) + (-1)^{(n-1)/2} t^n x_0(t),$$

where

$$x_0(t) = \frac{2t}{1+t^2} \qquad \text{and} \qquad y_0(t) = \frac{1-t^2}{1+t^2}$$

are standard rational parametrizations of circle segments so that $x_0(t)^2 + y_0(t)^2 = 1$. Squaring the expressions for $x_n(t)$ and $y_n(t)$ we obtain (6). ■

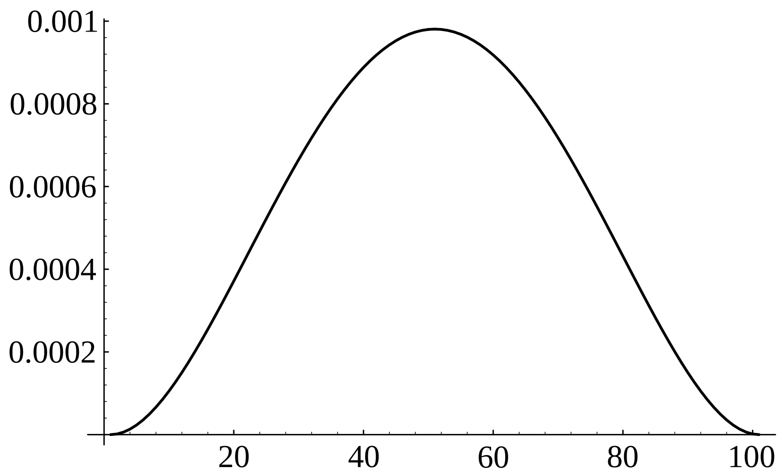

Figure 2. Polynomial curve approximation to a circle.

As is evident from the lemma, the approximation $\big(x_n(t), y_n(t)\big)$ approximates a circle segment with approximation order $2n$.

Theorem 8. *Let* $\mathbf{r}(s) = (\sin s, \cos s)$ *be a parametrization of the circle, and let* $\mathbf{p}_n(t) = (x_n(t), y_n(t))$ *denote the polynomial curve of Lemma 7. Then* $\mathbf{p}_n$ *provides a circle approximation with approximation order* $2n$ *in a neighbourhood of* $(0, 1)$.

Proof: The mapping given by $s(t) = \arctan\big(x_n(t)/y_n(t)\big)$ satisfies $s(0) = 0$ and $s'(0) = 1$. Therefore, s maps some small interval $I_h = [-h, h]$ one-to-one and onto some small interval $[-\alpha, \alpha]$. Note that then $\mathbf{p}_n(t)$ and $\mathbf{r}(s)$ lie on the same ray from the origin. We will compare $\mathbf{p}_n$ and $\mathbf{r}$ by comparing $\mathbf{p}_n(t)$ with $\mathbf{r}(s(t))$. Since those two vectors are parallel, we find

$$\begin{aligned}
||\mathbf{p}_n(t) - \mathbf{r}(s(t))|| &= \big|\, ||\mathbf{p}_n(t)|| - ||\mathbf{r}(s(t))||\, \big| \\
&= \Big|(x_n(t)^2 + y_n(t)^2 - 1\Big| / \Big(\sqrt{x_n(t)^2 + y_n(t)^2} + 1\Big) \\
&\le \Big|(x_n(t)^2 + y_n(t)^2 - 1\Big|.
\end{aligned}$$

For $t \in [-h, h]$ we therefore have

$$\begin{aligned}
d_P(\mathbf{r}, \mathbf{p}_n)_{I_h} &\le \max_{t \in I_h} \big\|\mathbf{r}(s(t)) - \mathbf{p}_n(t)\big\| \\
&\le \max_{t \in I_h} \big|x_n(t)^2 + y_n(t)^2 - 1\big| \\
&\le h^{2n},
\end{aligned}$$

by Lemma 7. $\blacksquare$

Since $|x_n(t)^2 + y_n(t)^2 - 1| \le 2^{-2n}$ for $t \le 1/2$, it is clear that for such values of t we have better and better approximations to a part of the circle as n increases. One can ask the question about how large a segment of the circle this represents. It can be seen that $x_n(t)$ is increasing and $y_n(t)$ is decreasing for $t \leq 1/2$. Using the formulas for $x_n(t)$ and $y_n(t)$ in the proof of Lemma 7 we also see that for $|t| < 1$

$$\lim_{n\to\infty} \frac{x_n(t)}{y_n(t)} = \frac{2t}{1-t^2}.$$

For $t = 1/2$ the right hand side of this expression is equal to $4/3$. Thus the curve $(x_n(t), y_n(t))$ will be a good approximation to a segment of a circle of angular width α_n with $\lim_{n\to\infty} \tan\alpha_n = 4/3$.

Finding approximations of even order n is not so simple. For general even degree n there does not appear to be any simple polynomials x_n and y_n such that $x_n(t)^2 + y_n(t)^2 - 1 = \alpha t^{2n}$ for a suitable constant α. A solution for $n = 6$ is shown in Figure 2. Here $x(t) = 2t - (3-\sqrt{3})t^3 + 2(2-\sqrt{3})^2 t^5$, $y(t) = 1 - 2t^2 + (4-2\sqrt{3})t^4 - (26-15\sqrt{3})t^6$, and $x(t)^2 + y(t)^2 - 1 = (1351 - 780\sqrt{3})t^{12} \approx 0.00037t^{12}$. The curve shown corresponds to $|t| \leq 1.8$ and we have an approximation to a full circle with error 0.02.

References

1. de Boor, C., K. Höllig, and M. Sabin, High accuracy geometric Hermite interpolation, Computer-Aided Geom. Design **4** (1988), 269–278.
2. Degen, W. L. F., Best approximations of parametric curves by splines, in *Mathematical Methods in Computer Aided Geometric Design II*, T. Lyche and L. L. Schumaker (eds.), Academic Press, Boston, 1992, 171–184.
3. Degen, W. L. F., High accurate rational approximation of parametric curves, Computer-Aided Geom. Design **10** (1993), 293–313.
4. Dokken, T., M. Dæhlen, T. Lyche, and K. Mørken, Good approximation of circles by curvature-continuous Bézier curves, Computer-Aided Geom. Design **7** (1990), 33–41.
5. Mørken, K., Best approximation of circles segments by quadratic Bézier curves, in *Curves and Surfaces*, P. J. Laurent, A. Le Méhautè, and L. L. Schumaker, (eds.), Academic Press, New York, 1991, 331–336.

Acknowledgements. The first author was supported in part by the Research Council of Norway through project STP28402 and SINTEF SI.

T. Lyche and K. Mørken
Dept. of Informatics, University of Oslo
P. B. 1080, Blindern
0316 Oslo, Norway