# Relationships Between the FastICA Algorithm and the Rayleigh Quotient Iteration

Scott C. Douglas

Department of Electrical Engineering,
Southern Methodist University,
Dallas, Texas 75275 USA

**Abstract.** The FastICA algorithm is a popular procedure for independent component analysis and blind source separation. Recently, several of its convergence properties have been elucidated, including its average convergence performance and its finite-sample behavior. In this paper, we examine the kurtosis-based algorithm version for two-source mixtures with equal-kurtosis sources, proving that the single-unit FastICA algorithm has dynamical behavior that is identical to the Newton-based Rayleigh Quotient Iteration for finding an eigenvector of a symmetric matrix. We also derive a bound on the average inter-channel interference indicating that the initial convergence rate of FastICA is linear with a rate of (1/3). A simulation indicates its convergence performance.

## 1 Introduction

The FastICA algorithm of Hyvarinen and Oja [1] is a popular procedure for independent component analysis (ICA) and blind source separation. The technique is simple to set up, converges quickly, and provides good separation behavior in a variety of contexts. Moreover, when a fourth-moment or kurtosis-based contrast function is used within the algorithm, convergence is globally cubic about a separating solution for the linear ICA model with non-Gaussian sources [2]. The technique has become popular for a number of problems in signal analysis.

Various studies of the convergence and identification behavior of the FastICA have been made, including stationary point analyses in the two-source and $m$-source mixing cases [3,4], its average convergence performance [4,5,6], and its finite-sample behavior at convergence [7]. In this paper, we add to this knowledge about the convergence of the FastICA algorithm by studying the algorithm for mixtures of two sources with equal kurtoses, with the goal of providing additional theoretical insight into the algorithm's behavior. In this situation, we prove that

– the single-unit FastICA algorithm has dynamical behavior that is mathematically-identical to the Newton-based Rayleigh Quotient Iteration for finding a minimum eigenvalue of a symmetric matrix, and
– convergence of the average inter-channel interference is bounded above by a function that converges linearly with rate (1/3) or 4.77dB per iteration.

J. Rosca et al. (Eds.): ICA 2006, LNCS 3889, pp. 781–789, 2006.

Thus, our results support previous observations made in the ICA literature which indicated linear (exponential) convergence of FastICA with a kurtosis contrast [6], and it connects the algorithm with a well-known eigenvector search procedure. A simulation verifies the derived performance bound.

## 2 Kurtosis-Based FastICA for Two-Source Mixtures

We briefly introduce the FastICA algorithm for two-source mixtures so that notation can be defined; complete descriptions of the algorithm are in [1,2]. Let $\mathbf{s}(k) = [s_1(k)\ s_2(k)]^T$, where $s_1(k)$ and $s_2(k)$ are zero-mean, unit variance, non-Gaussian, and statistically-independent at time $k$, such that their normalized kurtoses are $\kappa_i = E\{s_i^4(k)\} - 3$, $i \in \{1,2\}$. Let $\mathbf{x}(k) = \mathbf{A}\mathbf{s}(k)$ contain a linear mixture of these sources, The single-unit FastICA procedure first determines a whitening transformation $\mathbf{P}$ such that $\mathbf{v}(k) = \mathbf{P}\mathbf{x}(k)$ contains unit-power uncorrelated signals. A weight vector $\mathbf{w}_t = [w_{1,t}\ w_{2,t}]^T$ is then adjusted such that

$$y_t(k) = \mathbf{w}_t^T \mathbf{v}(k) \tag{1}$$

is an estimate of one extracted source. The adjustment procedure is

$$\widetilde{\mathbf{w}}_t = E\{y_t^3(k)\mathbf{v}(k)\} - E\{y_t^2(k)\}\mathbf{w}_t, \qquad \mathbf{w}_{t+1} = \frac{\widetilde{\mathbf{w}}_t}{\sqrt{\widetilde{\mathbf{w}}_t^T \widetilde{\mathbf{w}}_t}} \tag{2}$$

if a kurtosis-based cost function is used. An extension allows one to use other cost functions [2]. Sampled data averages are used to compute all expectations.

For analysis, consider the behavior of FastICA in the combined coefficient vector $\mathbf{c}_t = \mathbf{A}^T\mathbf{P}^T\mathbf{w}_t$, in which case $y(k) = \mathbf{c}^T\mathbf{s}(k)$. Furthermore, for two-source mixtures, we introduce an intrinsic parametrization for $\mathbf{c}_t = [c_{1,t}\ c_{2,t}]^T$ given by

$$\mathbf{c}_t = [\cos(\theta_t)\ \ \sin(\theta_t)]^T. \tag{3}$$

Then, letting the number of data measurements tend to infinity, an equivalent expression for FastICA in this case is [4]

$$c_{1,t+1} = \frac{\kappa_1 c_{1,t}^3}{\sqrt{\kappa_1^2 c_{1,t}^6 + \kappa_2^2 c_{2,t}^6}}, \quad c_{2,t+1} = \frac{\kappa_2 c_{2,t}^3}{\sqrt{\kappa_1^2 c_{1,t}^6 + \kappa_2^2 c_{2,t}^6}}, \tag{4}$$

which can be represented even more compactly using (3) as

$$\tan(\theta_{t+1}) = \frac{\kappa_2}{\kappa_1}\tan^3(\theta_t). \tag{5}$$

When $|c_{2,t}| \leq |c_{1,t}|$, the ratio $c_{2,t}/c_{1,t} = \tan(\theta_t)$ is related to the inter-channel interference (ICI) at time $t$ as

$$ICI_t = \frac{c_{2,t}^2}{c_{1,t}^2} = \tan^2(\theta_t), \tag{6}$$

which represents a useful performance factor for the algorithm. In this paper, we shall assume equal kurtosis sources, such that $\kappa_2/\kappa_1 = 1$. Equal-magnitude kurtosis sources could be handled with additional notational changes. We shall also restrict our study to the case $|c_{2,t}| \leq |c_{1,t}|$ or $|\theta_t| \leq \pi/4$, as all other convergence regions follow from the four-fold symmetry of the parameter space.

## 3 The Kurtosis Contrast, Newton's Method, and the Rayleigh Quotient Iteration

The single-unit kurtosis-based FastICA algorithm can be derived as an *approximate* Newton's method for minimizing the cost [1,2]

$$\mathcal{J}(\mathbf{w})) = -\left|E\{y^4(k)\} - 3\left(E\{y^2(k)\}\right)^2\right| \tag{7}$$

under a unit power constraint on $y(k)$ given by $E\{y^2(k)\} = \mathbf{w}^T\mathbf{w} = \mathbf{c}^T\mathbf{c} = 1$. In the two-source case, this constraint is exactly maintained by (3). We can express the kurtosis contrast for a two source mixture with equal-kurtosis sources in the vicinity of $\theta = 0$ as

$$\mathcal{J}(\mathbf{c}) = -|\kappa|\left[c_1^4 + c_2^4\right] \quad \text{or} \quad \mathcal{J}(\theta) = -|\kappa|\left[\cos^4(\theta) + \sin^4(\theta)\right]. \tag{8}$$

Our pursuit of knowledge about the FastICA procedure can move in either a constructive or an analytical fashion, and we choose the former approach first. Ignoring our ability to represent iterative algorithms using measured data, what are some good approaches that could be used to minimie $\mathcal{J}(\theta)$ with respect to $\theta$? Clearly, Newton's method is of interest given the convexity and evenness of $\mathcal{J}(\theta)$ at $\theta = \{0, \pi/2, \pi, 3\pi/2\}$, as Newton-based methods converge cubically under such conditions. The one-dimensional gradient and Hessian of $\mathcal{J}(\theta)$ are

$$\frac{\partial \mathcal{J}(\theta)}{\partial \theta} = |\kappa|\sin(4\theta) \quad \text{and} \quad \frac{\partial^2 \mathcal{J}(\theta)}{\partial \theta^2} = 4|\kappa|\cos(4\theta), \tag{9}$$

such that Newton's method for adapting $\theta$ is

$$\theta_{t+1} = \theta_t - \frac{1}{4}\tan(4\theta_t). \tag{10}$$

Near $\theta_t = 0$, the algorithm is indeed cubically-convergent, as

$$\theta_{t+1} = -\frac{16}{3}\theta_t^3 + \mathcal{O}(\theta_t^5). \tag{11}$$

Eqn. (10) has appeared before in the analysis of the Rayleigh quotient iteration (RQI) for finding the eigenvector of a symmetric matrix [8], in which the cost function in the transformed eigenvector $\mathbf{c}_t$ and eigenvalues $\{\lambda_1, \lambda_2\}$ is

$$\overline{\mathcal{J}}(\mathbf{c}) = \lambda_1 c_1^2 + \lambda_2 c_2^2 \quad \text{or} \quad \overline{\mathcal{J}}(\phi) = \lambda_1\cos^2(\phi) + \lambda_2\sin^2(\phi), \tag{12}$$

Newton's method for minimizing $\overline{\mathcal{J}}(\phi)$ is

$$\phi_{t+1} = \phi_t - \frac{1}{2}\tan(2\phi_t) \;\approx\; -\frac{4}{3}\phi_t^3 + \mathcal{O}(\phi_t^5), \tag{13}$$

which is identical in form to (10) for $\theta_t = 2\phi_t$. That the two costs in (8) and (12) would produce the same Newton iteration is remarkable, but our interest here is in realizable algorithms, such as FastICA for blind source separation. The following theorem, proven in the Appendix, illuminates an important relationship between FastICA and the RQI.

**Theorem 1.** *In the two-source equal-kurtosis case as the number of measurements tends to infinity, the FastICA algorithm for minimizing (8) in the intrinsic parametrization variable $\theta_t$ is*

$$\theta_{t+1} = -\theta_t + \arctan\left(\frac{1}{2}\tan(2\theta_t)\right) \;=\; \theta_t^3 + \mathcal{O}(\theta_t^5), \tag{14}$$

*which is identical in form to the Rayleigh Quotient Iteration for minimizing (12) in the intrinsic parametrization variable $\phi_t$, as given by*

$$\phi_{t+1} = \phi_t - \arctan\left(\frac{1}{2}\tan(2\phi_t)\right) \;=\; -\phi_t^3 - \mathcal{O}(\phi_t^5), \tag{15}$$

*where $\theta_t = (-1)^t\phi_t$. Moreover, both the RQI and FastICA are approximate step-and-project Newton algorithms in $\mathbf{c}_t$ employing the tangent form of the Newton update within the intrinsic parameter space $\theta_t$ or $\phi_t$.*

The above theorem states that in the two-source case, the weight vector $\mathbf{w}_t$ for the FastICA algorithm evolves identically to that for the RQI applied to a symmetric matrix $\boldsymbol{\Gamma}\mathrm{diag}\{\lambda_1, \lambda_2\}\boldsymbol{\Gamma}^T$ when $\boldsymbol{\Gamma} = \mathbf{PA}$ is orthonormal, except for the sign changes associated with the alternating update directions in the RQI. We have verified this fact numerically to the machine precision limits of MATLAB by running each algorithm its respective task for the same $\boldsymbol{\Gamma}$ separation and eigenvector matrix, respectively. That the FastICA procedure shares the same evolutionary behavior of RQI is informative, as RQI is a well-known and well-studied procedure in the numerical linear algebra community [8,9]. The RQI is considered one of the best procedures for its task due to its local cubic convergence. This link means that convergence results for RQI can potentially be applied to the FastICA algorithm, and vice versa.

To better see the geometrical relationships of the various algorithms, Figure 1 illustrates a single iteration of each algorithm in both two-dimensional $\mathbf{c}$-space as well as one-dimensional angular space. Point $O$ corresponds to the point on the unit circle at angle $\theta_t$ of the FastICA algorithm. Point $P$ is the negative of this angle at $-\theta_t$, which we will set equal to $\phi_t$ for comparison with the RQI. Vector $OD$ is the the component of the Newton update direction for minimizing the kurtosis-based cost in (8) in the tangent space at point $O$ or angle $\theta_t$. Vector $PE$ is the component of the Newton update direction for minimizing the Rayleigh quotient cost in (12) in the tangent space at angle $\phi_t$. Point $A$ is reached by the update in (10), in which the arclength $OA$ is equal to the linear distance $OD$.

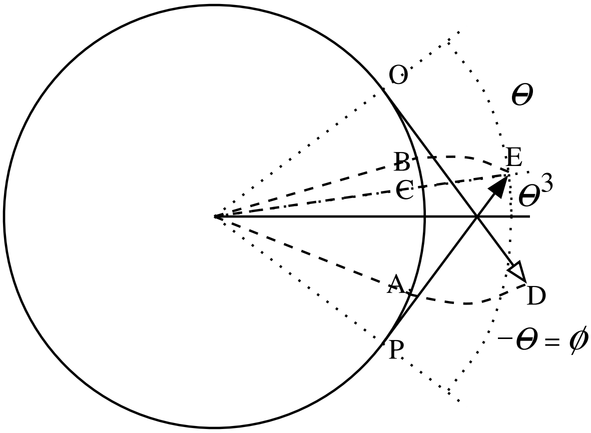

**Fig. 1.** Geometry of the FastICA algorithm and the Rayleigh Quotient Iteration - see text for label descriptions

Point $B$ is reached by the update in (13), in which the arclength $PB$ is equal to the linear distance $PE$. Point $C$ is the result of both FastICA and RQI in their respective problems, which is obtained by projecting the point $E$ to the unit circle. In all cases, the angle magnitude is reduced in proportion to the cube of the original angle $\theta_t$ or $\phi_t$ for small angles: Point $A$ is at an angle of approximately $-\frac{16}{3}\theta_t^3$, Point $B$ is at an angle of approximately $\frac{4}{3}\theta_t^3 = -\frac{4}{3}\phi_t^3$, and point $C$ is at an angle of approximately $\theta_t^3$.

In practice, the kurtoses of the two sources are not equal; even so, the locally-cubic convergence of the FastICA algorithm is maintained. From (5),

$$\theta_{t+1} = \arctan\left(\frac{\kappa_2}{\kappa_1}\tan^3(\theta_t)\right) = \frac{\kappa_2}{\kappa_1}\theta_t^3 + \mathcal{O}(\theta_t^5). \tag{16}$$

Convergence of $\theta_t$ to zero remains cubic. The $\kappa_2/\kappa_1$ factor in (16) does not significantly alter the algorithm's local convergence behavior. In fact, if $\kappa_2/\kappa_1 < 0$, the update's oscillatory behavior about $\theta = 0$ is identical to that in the RQI.

## 4 A Bound on the Average ICI for FastICA

The FastICA algorithm appears to converge quickly in many contexts. In [6], the average behavior of the inter-channel interference (ICI) for kurtosis-based FastICA on general $m$-dimensional mixtures was observed in simulations to be exponential with rate (1/3). Recent analytical work has verified this convergence property under a range of initial conditions on $\mathbf{w}_0$ [5,6]. The goal of this section is to use the simplicity of the two-source FastICA algorithm analysis to verify this property under general initial conditions for $\mathbf{w}_0$.

When $\kappa_1 = \kappa_2$, we can use (5) to write the evolutionary equation for FastICA in a remarkably simple form for the ICI at time $t$:

$$ICI_t = ICI_{t-1}^3 \;=\; (ICI_0)^{3^t}. \tag{17}$$

This scalar evolutionary equation for the ICI is cubically-convergent *globally* so long as the saddle point $ICI_0 = 1$ occurs with zero finite probability. Convergence depends on $ICI_0$, which for our analysis is assumed to have an unknown scalar p.d.f. $p_0(u)$ over the range $0 \leq u \leq 1$.

The average ICI, denoted as $E\{ICI_t\}$, is the ensemble average of the ICI values at iteration $t$ that one would obtain by running FastICA on the same data set with different initial conditions as characterized by the p.d.f of $ICI_0$. Here, infinite data has been assumed. The following bound characterizes the value of $E\{ICI_t\}$ for weak assumptions on the p.d.f. of the initial ICI, the proof of which is in the Appendix.

**Theorem 2.** *Let* $ICI_0$ *be arbitrarily-distributed on* $[0, ICI_{max}]$ *with distribution* $p_0(u)$*, where* $0 < ICI_{max} \leq 1$*, subject to the additional condition that the probability density of* $ICI_0$ *has no point masses, or equivalently, the cumulative distribution function of* $ICI_0$ *is continuous over the interval* $[0,1]$*. Define* $K = \max_{0 \leq u \leq ICI_{max}} u p_0(u)$*. Then, an upper bound on the average ICI of the FastICA algorithm at iteration t for a linear mixture and infinite data in the two-source case is*

$$E\{ICI_t\} \leq \left(\frac{1}{3}\right)^t (ICI_{max})^{3^t} K. \tag{18}$$

Theorem 2 states that for reasonable distribution assumptions on the initial ICI, the average ICI at time $t$ is bounded by a function consisting of the product of a linear-converging term and a cubically-converging term. Cubic convergence is what the deterministic analysis in [1] describes for kurtosis-based FastICA, and it is ultimately attained under stochastic initial conditions of the separation system vector if the initial distribution of the inter-channel interference is bounded away from unity. It may take a number of iterations, however, before this cubically-converging term dominates the expression. During the initial convergence period, the bound is linear with rate (1/3), as observed in simulations in [4]. Moreover, if the uncertainty about the mixing system prevents one from bounding $ICI_0$ away from unity – a likely situation – then the bound predicts *only linear convergence.*

Finite data records prevent one from attaining $\lim_{t\to\infty} E\{ICI_t\} = 0$ in practice. Experience show that linear convergence of FastICA with rate (1/3) is typically observed from the multiple-unit kurtosis-based FastICA algorithm applied to finite-length data sets. The performance "floor" due to finite measurements prevents one from observing the eventual cubic convergence of the FastICA procedure. The above bound indicates why linear convergence behavior is observed.

To verify the above behavior, the following simulations were carried out. The FastICA procedure was applied to 10000 different realizations of $N = 1000$ snapshots of mixtures of Unif-$[-\sqrt{3}, \sqrt{3}]$-distributed sources. The initial coefficient

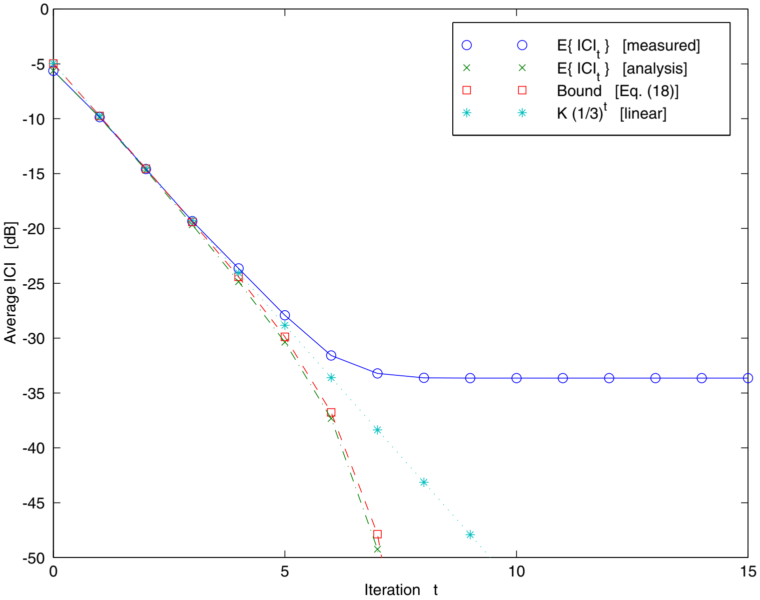

**Fig. 2.** Convergence of $E\{ICI_t\}$ from simulations and from predictions

vector $\mathbf{w}_0$ for each realization was randomly and uniformly-selected from an arbitrary point on the unit circle satisfying $ICI_0 < ICI_{max} = 0.999$. Ensemble averages were then used to compute $E\{ICI_t\}$ for comparison with the average behavior as predicted by the bound in (17), where $K \approx 1/\pi$ [5]. As a check, the random values of $ICI_0$ were used to compute an estimate $E\{\widehat{ICI}_t\}$ of $E\{ICI_t\}$ using ensemble averages of (17), which assumes infinite data ($N \rightarrow \infty$).

Figure 2 shows the evolutions of the various measures of inter-channel interference, in which the bound in (18) is seen to accurately predict both $E\{ICI_t\}$ and $E\{\widehat{ICI}_t\}$ for small $t$. Eqn. (18) closely follows the average behavior of (17) for larger $t$, but the measured $E\{ICI_t\}$ continues to converge linearly with rate (1/3) until performance limits due to prewhitening and finite-sample effects are reached. In essence, the FastICA algorithm does not achieve cubic convergence on average despite having cubic convergence in a deterministic setting.

## 5 Conclusions

In this paper, we illustrate an important connection between the popular FastICA algorithm for independent component analysis and the Rayleigh Quotient Iteration in numerical linear algebra. We also derive a bound on the evolution of the average inter-channel interference for the FastICA algorithm for equal-kurtosis two-source mixtures which predicts linear convergence of the algorithm initially. Simulations show that the average ICI in FastICA typically converges linearly despite having cubic convergence in a deterministic setting.

## References

1. A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Computation,* vol. 9, no. 7, pp. 1483-1492, Oct. 1997.
2. A. Hyvarinen, J. Karhunen, and E. Oja, *Independent Component Analysis* (New York: Wiley, 2001).
3. E. Oja, "Convergence of the symmetrical FastICA algorithm," *Proc. 9th Int. Conf. Neural Inform. Processing,* Singapore, vol. 3, pp. 1368-1372, Nov. 2002.
4. S.C. Douglas, "On the convergence behavior of the FastICA algorithm," *Proc. Fourth Symp. Indep. Compon. Anal. Blind Signal Separation,* Kyoto, Japan, pp. 409-414, Apr. 2003.
5. S.C. Douglas, "A statistical convergence analysis of the FastICA algorithm for two-source mixtures," *Proc. 39th Asilomar Conf. Signals, Syst., Comput.*, Pacific Grove, CA, Oct. 2005.
6. S.C. Douglas, Z. Yuan, and E. Oja, "Average convergence behavior of the FastICA algorithm for blind source separation," to appear in *Proc. 6th Int. Conf. Indep. Compon. Anal. Blind Source Separation,* Charleston, SC, Mar. 2006.
7. Z. Koldovsky, P. Tichavsky, and E. Oja, "Cramer-Rao lower bound for linear independent component analysis," *Proc. IEEE Int. Conf. Acoust., Speech, Signal Processing,* Philadelphia, PA, vol. 3, pp. 581-584, Mar. 2005.
8. A. Edelman, T.A. Arias, and S.T. Smith, "The geometry of algorithms with orthogonality constraints," *SIAM J. Matrix Anal., Appl.* vol. 20, pp 303-353, Apr. 1999.
9. G.H. Golub and C.F. Van Loan, *Matrix Computations,* 3rd. ed. (Baltimore: Johns Hopkins, 1996).

## Appendix

*Proof of Theorem 1.* From (5), we can write the update of the single-unit FastICA procedure applied to mixtures of two equal-kurtosis sources in the variable $\theta_t$ as

$$\theta_{t+1} = \arctan\left(\tan^3(\theta_t)\right). \tag{19}$$

Consider an expression for $\tan^3(\theta)$ of the form

$$\tan^3(\theta) = \tan(\alpha - \theta). \tag{20}$$

Applying the tan difference formula, we obtain the relationship

$$\tan^3(\theta) = \frac{\tan(\alpha) - \tan(\theta)}{1 + \tan(\alpha)\tan(\theta)}, \tag{21}$$

$$\text{or} \qquad \tan(\alpha)(1 - \tan^2(\theta)) = \tan(\theta). \tag{22}$$

Assume first that $\theta \neq \pi/4$, in which case

$$\tan(\alpha) = \frac{\tan(\theta)}{1 - \tan^2(\theta)} = \frac{1}{2}\tan(2\theta). \tag{23}$$

Now, considering the case that $\theta = \pi/4$ and the solution for $\tan(\alpha)$ in (23) , the left-hand-side of (22) can be evaluated using L'Hopital's Rule as

$$\lim_{\theta \to \pi/4} \frac{1}{2} \tan(2\theta)(1 - \tan^2(\theta)) = \lim_{\theta \to \pi/4} \frac{1}{2} \sin^2(2\theta)[\tan^3(\theta) + \tan(\theta)] \; = \; 1. \quad (24)$$

Thus, we have $\alpha = \arctan(0.5 \tan(2\theta))$ for all $\theta$, such that

$$\tan^3(\theta) = \tan\left(\arctan\left(\frac{1}{2}\tan(2\theta)\right) - \theta\right). \quad (25)$$

Setting $\theta = \theta_t$ and taking the arc-tangent of both sides of (25), the result follows.

*Proof of Theorem 2.* Let $p_0(u)$ denote the p.d.f. of $ICI_0$. Then, we have

$$E\{ICI_t\} = \int_0^{ICI_{max}} u^{3^t} p_0(u) du \; = \; \int_0^{ICI_{max}} u^{3^t - 1} \cdot u p_0(u) du. \quad (26)$$

Using the Holder inequality, we have

$$E\{ICI_t\} \le \left(\int_0^{ICI_{max}} u^{r(3^t-1)}\right)^{\frac{1}{r}} \left(\int_0^{ICI_{max}} u^s p_0^s(u) du\right)^{\frac{1}{s}} \quad (27)$$

$$\le \left(\frac{1}{r(3^t-1)+1}\right)^{\frac{1}{r}} (ICI_{max})^{3^t - 1 + \frac{1}{r}} \left(\int_0^{ICI_{max}} u^s p_0^s(u) du\right)^{\frac{1}{s}} \quad (28)$$

where $1/r + 1/s = 1$. Letting $s \to \infty$ and $r \to 1$, we have the inequality in (18).