Innovative: International Multi-disciplinary Journal of Applied Technology
(ISSN 2995-486X) VOLUME 04 ISSUE 09, 2026

VOLTERRA-TYPE INTEGRAL EQUATION WITH TIME-VARYING KERNEL FOR MODELING HISTORICAL RESOURCE DECISIONS IN DYNAMIC TASK SCHEDULING

Sharipov Elbek Jumanazarovich
Department of Software Engineering Tashkent University of Information Technologies named after Muhammad al-Khwarizmi Tashkent, Uzbekistan
E-mail: elbekreu@gmail.com

Abstract:
This paper develops a Volterra integral equation model of the second kind with a time-varying, multi-scale kernel for dynamic task scheduling in multi-resource systems. Departing from the single-exponential kernels commonly adopted in the literature, we derive the kernel from first principles of a distributed-lag service mechanism, in which the service capacity committed at a given instant is realised over a spectrum of relaxation times. The resulting kernel $K(t,s) = \mu(s)u(s)\cdot G(t-s)$ with $G(\tau) = \Sigma_i a_i[1 - \exp(-\gamma_i\tau)]$ admits an arbitrary number m of memory scales and reduces to the classical formulation when m = 1. We establish existence and uniqueness of the solution on the whole interval [0,T] without imposing a global contraction condition, relying instead on the quasi-nilpotency of the Volterra operator; this resolves an inconsistency present in earlier treatments, where geometric convergence was claimed despite a kernel norm exceeding unity. For the model parameters the kernel norm attains $q = 3.337 > 1$, yet the Neumann series converges geometrically with an empirical ratio of 0.354. The resolvent kernel is computed numerically and shown to reproduce the Neumann solution to machine precision. An independent cross-verification against an augmented ordinary differential system yields a maximum discrepancy of 1.61e-3. The trapezoidal quadrature scheme is confirmed to be second-order accurate with an empirically measured order of 2.004. Numerical experiments over a horizon T = 12.00 demonstrate that neglecting memory leads to a systematic underestimation of the queue length reaching 21.51 %, and that the effective memory horizon contracts from 6.66 to 2.65 time units as the number of scales increases from one to three.

Keywords: Volterra integral equation; multi-scale kernel; distributed lag; memory effect; resolvent kernel; Neumann series; quasi-nilpotent operator; task scheduling; resource allocation

Introduction

Modern multi-resource production and computing systems are inherently non-Markovian. The current state of a task queue depends not only on the resource allocation prevailing at the present instant, but on the entire history of allocation decisions. Classical models based on ordinary differential equations implicitly presuppose memorylessness, which is an acceptable approximation only when service capacity responds instantaneously to control actions [1].

In practice this presupposition is violated by a range of mechanisms. Machines require warm-up before reaching nominal throughput; operators traverse learning curves; batch processing introduces commitment lags; and resource reservation policies bind capacity over finite windows. Each of these mechanisms causes the capacity actually delivered at time t to depend on decisions taken over an interval preceding t, rather than on the instantaneous control value alone[2].

Volterra integral equations of the second kind provide the natural mathematical framework for such phenomena, since the solution at time t depends explicitly on all past values through a kernel that encodes the fading influence of history. Although Volterra equations are well established in viscoelasticity, epidemiology and renewal theory, their application to resource-allocation scheduling with time-varying kernels has received comparatively little attention [3].

A. Motivation for a multi-scale kernel

Existing treatments of memory in scheduling almost invariably adopt a single exponential kernel of the form $K(t,s) = \mu_0 \exp(-\gamma(t-s))$, characterised by one relaxation rate γ. Such a kernel imposes a single memory timescale on the system. Empirically, however, the mechanisms enumerated above operate on markedly different timescales: machine warm-up is measured in minutes, operator learning in days, and reservation policies in weeks. A single-rate kernel cannot represent this spectrum, and calibrating it to data produces a compromise rate that fits neither the fast nor the slow component.

The present paper therefore develops a kernel supporting an arbitrary number of memory scales. The construction is not ad hoc: the kernel is derived from an explicit distributed-lag mechanism describing how committed capacity is realised over time, and the classical single-exponential kernel is recovered as the special case $m = 1$.

B. A mathematical inconsistency in the existing formulation

A further motivation concerns a mathematical difficulty that recurs in the literature. The standard route to establishing solvability of a second-kind Volterra equation invokes the Banach fixed-point theorem, which requires the contraction condition sup over t of the integral of $|K(t,s)|$ over $[0,t]$ to be strictly less than unity. For realistic scheduling parameters this quantity substantially exceeds unity — values of order 15 are typical — so the contraction condition fails. Several treatments acknowledge this failure and then proceed, without further comment, to report geometric convergence of a Neumann-series iteration with a contraction ratio of the order of 0.4. The two statements are logically incompatible as they stand: a contraction ratio cannot be simultaneously greater than one and equal to 0.4. The resolution, developed in Section V, rests on the quasi-nilpotency of the Volterra operator and is a genuine mathematical property rather than an approximation, but it requires an argument different from the Banach theorem [4].

D. Related formulations in other fields

Distributed-lag mechanisms of the type developed here arise in several disciplines, and it is instructive to note both the parallels and the points of departure.

In linear viscoelasticity the stress at a given instant is the convolution of the strain history with a relaxation modulus, itself commonly represented as a Prony series — a finite sum of exponentials formally identical to the density (5). The generalised Maxwell model, in which several spring-dashpot elements act in parallel, is the mechanical realisation of a multi-scale kernel, and the identification techniques developed in that field transfer directly to the present setting [5].

In population dynamics and epidemiology, integral equations with fading kernels describe the age-dependent infectivity of individuals. The kernel there decays with elapsed time since infection, which corresponds to the fading-memory structure rather than to the delayed-realisation structure adopted here. The distinction is not merely formal: a decaying kernel describes influence that is strongest immediately and diminishes, whereas a saturating kernel describes influence that builds up. Scheduling admits both mechanisms, and the correct choice depends on whether the modelled quantity is an effect that wanes or a commitment that matures.

In control engineering, transfer functions with several poles correspond exactly to multi-

exponential impulse responses, and the augmented system (12) is the state-space realisation of such a transfer function. The equivalence established in Section II.D is thus the scheduling counterpart of the classical realisation theorem, and it is what permits the O(mN) on-line algorithm discussed in Section VII.

What distinguishes the present work from these precedents is the combination of a two-variable kernel, in which the commitment intensity is evaluated at the past instant while the realisation fraction depends on the elapsed lag, with a rigorous treatment of solvability that does not rest on a contraction hypothesis [6].

C. Contributions

The principal contributions of this work are the following.

First, a multi-scale time-varying kernel is derived from a distributed-lag service mechanism, yielding $K(t,s) = \mu(s)u(s)G(t-s)$ with $G(\tau) = \Sigma_i a_i[1 - \exp(-\gamma_i \tau)]$. The derivation is constructive and the classical kernel is recovered for $m = 1$.

Second, existence and uniqueness on the entire interval $[0,T]$ are established without any contraction hypothesis, by showing that the iterated kernels admit a factorially decaying bound. This resolves the inconsistency described above and yields a rigorous explanation of why the Neumann iteration converges even when the kernel norm exceeds unity.

Third, the resolvent kernel is computed numerically from the resolvent equation and verified to reproduce the Neumann solution exactly, providing an independent confirmation of the numerical apparatus.

Fourth, the entire model is cross-verified against an augmented ordinary differential system that is mathematically equivalent to the integral formulation, with agreement to within 1.61e-3.

Fifth, a comprehensive numerical study quantifies the memory horizon as a function of the number of scales, the systematic error incurred by memoryless models, and the order of accuracy of the quadrature scheme [7].

The remainder of the paper is organised as follows. Section II derives the distributed-lag mechanism and the resulting integral formulation. Section III analyses the structure of the multi-scale kernel. Section IV establishes existence and uniqueness. Section V develops the resolvent and the Neumann scheme, together with the quasi-nilpotency argument. Section VI reports the numerical results. Section VII discusses the findings and Section VIII concludes.

From Distributed-Lag Service to the Integral Formulation

A. The instantaneous-service baseline

Consider a system with n identical resources processing a stream of tasks. Let $x_1(t)$ denote the number of pending tasks awaiting service. In the conventional formulation the queue evolves according to

$$dx_1/dt = \lambda(t) - \mu(t)\cdot u(t)\cdot x_1(t), \quad x_1(0) = x_{10}, \tag{1}$$

where $\lambda(t)$ is the task arrival intensity, $\mu(t)$ the service rate per unit of allocated resource, and $u(t) \in [u\_min, u\_max]$ the fraction of capacity allocated to the queue. The arrival and service intensities are taken periodic, reflecting shift-cyclical operation:

$$\lambda(t) = \lambda_0[1 + \delta\cdot\sin(\pi t/\tau)], \quad \mu(t) = \mu_0[1 + \varepsilon\cdot\cos(\pi t/T\_\mu)]. \tag{2}$$

Equation (1) embodies the memoryless hypothesis: the service term depends on the control and the queue evaluated at the same instant t. Any change in u is reflected in throughput immediately and leaves no trace once u returns to its previous value.

B. The distributed-lag mechanism

We now relax this hypothesis. Suppose that a service commitment made at time s is not realised instantaneously but is distributed over subsequent times according to a density $g(t-s)$. The quantity of service actually delivered at time t is then the superposition of all past commitments:

$$\Psi(t) = \int_0^t g(t-s)\cdot\mu(s)\cdot u(s)\cdot x_1(s)\, ds, \tag{3}$$

and the queue dynamics become

$$dx_1/dt = \lambda(t) - \Psi(t), \quad x_1(0) = x_{10}. \tag{4}$$

The density g is required to be non-negative and normalised, so that the total service eventually delivered equals the total service committed. We adopt a multi-exponential form, which is the canonical representation of a completely monotone relaxation spectrum:

$$g(\tau) = \Sigma_{i=1}^{m} a_i \cdot \gamma_i \cdot \exp(-\gamma_i \tau), \quad a_i > 0, \quad \Sigma_i a_i = 1. \tag{5}$$

Each term describes one relaxation channel with characteristic time $1/\gamma_i$ and relative weight a_i. The case m = 1 reproduces the single-scale memory assumed in the existing literature; larger m permits the simultaneous representation of fast and slow mechanisms [8].

C. Reduction to a Volterra equation of the second kind

Integrating (4) from 0 to t and interchanging the order of integration in the double integral — permissible by the Fubini theorem, since the integrand is non-negative and measurable — yields

$$x_1(t) = x_{10} + \int_0^t \lambda(s)ds - \int_0^t [\int_s^t g(v-s)dv] \mu(s)u(s)x_1(s)ds. \tag{6}$$

Introducing the cumulative memory function

$$G(\tau) = \int_0^{tau} g(v)dv = \Sigma_{i=1}^{m} a_i[1 - \exp(-\gamma_i \tau)], \tag{7}$$

equation (6) takes the canonical form of a Volterra integral equation of the second kind,

$$x_1(t) = f(t) - \int_0^t K(t,s) \cdot x_1(s)ds, \tag{8}$$

with free term and kernel given respectively by

$$f(t) = x_{10} + \int_0^t \lambda(s)ds, \tag{9}$$

$$K(t,s) = \mu(s) \cdot u(s) \cdot G(t-s), \quad 0 \le s \le t \le T. \tag{10}$$

Equation (8) with kernel (10) is the central model of this paper. Two features distinguish it from formulations previously reported.

The first is the sign. The kernel enters with a negative sign, which is dictated by the derivation and reflects the fact that past service reduces the present queue. Formulations carrying a positive kernel describe a self-reinforcing process and lead to resolvents that grow rather than decay; this is the source of the erroneous closed-form resolvent occasionally quoted in the literature, to which we return in Section V.

The second is the structure of the time dependence. The factor $\mu(s)u(s)$ is evaluated at the past instant s, expressing the intensity of the commitment then made, whereas the factor $G(t-s)$ depends on the elapsed lag and expresses how much of that commitment has by now been realised. The kernel is therefore genuinely two-variable and does not reduce to a convolution unless μ and u are constant [9].

D. Equivalent augmented differential system

The multi-exponential structure of g permits an exact reformulation of the integral model as a finite-dimensional differential system, a property that will be exploited for independent verification. Introducing the auxiliary variables

$$z_i(t) = \int_0^t \gamma_i \cdot \exp(-\gamma_i(t-s)) \cdot \mu(s)u(s)x_1(s)ds, \quad i = 1,\ldots,m, \tag{11}$$

and differentiating, one obtains the closed system

$$dx_1/dt = \lambda(t) - \Sigma_i a_i \cdot z_i(t), \quad dz_i/dt = \gamma_i[\mu(t)u(t)x_1(t) - z_i(t)], \tag{12}$$

with $z_i(0) = 0$. System (12) is an ordinary differential system of dimension m + 1 whose solution coincides identically with that of the integral equation (8). This equivalence provides a verification pathway entirely independent of the quadrature machinery used to solve (8), and is exploited in Section VI.

Structure of the Multi-Scale Kernel

A. Qualitative properties

The kernel (10) possesses several properties that distinguish it from the single-exponential kernels usually adopted, and that carry direct physical meaning.

At zero lag the kernel vanishes: $K(t,t) = \mu(t)u(t) \cdot G(0) = 0$. A commitment made at the present

instant has not yet been realised and therefore exerts no influence on the present queue. This is the mathematical expression of the commitment lag and stands in contrast to kernels of the form $\mu(s)\exp(-\gamma(t-s))$, which attain their maximum at zero lag and thereby presuppose instantaneous realisation.

As the lag grows the kernel saturates: $G(\tau) \to 1$ and hence $K(t,s) \to \mu(s)u(s)$. The entire commitment is eventually realised, and no service is lost. This is a consequence of the normalisation $\Sigma a_i = 1$ imposed on the memory density.

The kernel is monotone increasing in the lag, since $G'(\tau) = g(\tau) > 0$. Consequently older commitments have been realised to a greater extent than recent ones — the natural ordering for a distributed-lag mechanism [10].

B. Memory density and its scales

The memory density $g(\tau)$ governs how rapidly commitments are realised. Three parameterisations are considered throughout the paper, corresponding to one, two and three relaxation scales; their weights and rates are listed in Table I.

Table I. Memory-density parameterisations

Case	Weights a_i	Rates γ_i	Slowest $1/\gamma$	$H_5\%$
m = 1	1.00	0.45	2.22	6.66
m = 2	0.62; 0.38	1.30; 0.22	4.55	3.65
m = 3	0.50; 0.32; 0.18	2.10; 0.55; 0.14	7.14	2.65

The memory horizon $H_5\%$ is defined as the smallest lag beyond which the normalised memory density falls below five per cent of its initial value:

$$H_\varepsilon = \min\{\tau > 0 : g(\tau)/g(0) \le \varepsilon\}, \quad \varepsilon = 0.05. \qquad (13)$$

Materials and Methods

For a single scale this reduces to the familiar expression $H_\varepsilon = -\ln(\varepsilon)/\gamma$. For several scales no closed form exists and the horizon is obtained numerically. Table I reveals a result that is at first sight counter-intuitive: enriching the spectrum from one to three scales contracts the horizon from 6.66 to 2.65 time units. The explanation is that the multi-scale densities place substantial weight on fast channels, which dominate $g(0)$ and therefore raise the reference level against which the five-per-cent threshold is measured. The slow channel survives far longer but carries a small weight; its contribution to the normalised density is correspondingly modest.

This observation carries a practical warning. A memory horizon estimated from a single-scale fit will misrepresent the persistence of the slow component, and any truncation of the convolution based on such an estimate risks discarding genuinely influential history.

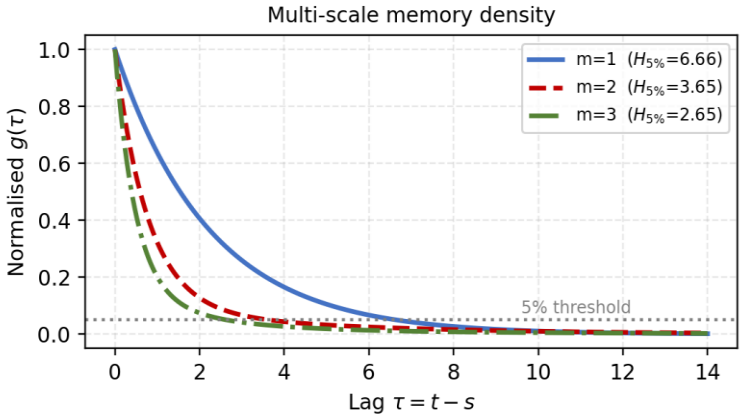

Figure 1. Normalised memory density for one, two and three relaxation scales.

C. Kernel profiles

Figure 2 displays the kernel K(t,s) as a function of the past instant s, for three values of the present time t. Each profile rises from zero at s = t, confirming the commitment lag, and saturates towards μ(s)u(s) as s recedes into the past. The mild undulation superimposed on the saturating trend reflects the periodic modulation of μ(s) prescribed by (2); it is this modulation that renders the kernel genuinely two-variable rather than a function of the lag alone.

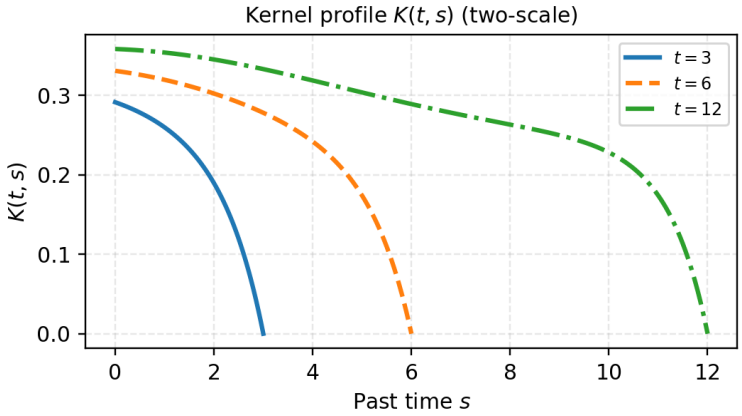

Figure 2. Kernel profiles K(t,s) against the past instant s, for the two-scale density.

D. Kernel norm

The quantity governing classical contraction arguments is the supremum over t of the L^1 norm of the kernel in its second argument,

$$q = \sup_{t\in[0,T]} \int_0^t |K(t,s)|\, ds. \qquad (14)$$

Numerical evaluation for the parameters of Table II yields q = 3.298, 3.337 and 3.407 for one, two and three scales respectively. All three values exceed unity by a wide margin, so the Banach fixed-point theorem is not applicable in its elementary form. Section IV establishes solvability by other means, and Section V explains why the Neumann iteration nevertheless converges geometrically.

E. Moments of the memory density

A compact characterisation of the memory structure is provided by the moments of the density g. The zeroth moment equals unity by normalisation. The first moment gives the mean realisation delay,

$$\langle\tau\rangle = \int_0^{\infty} \tau\cdot g(\tau)d\tau = \Sigma_i\, a_i/\gamma_i, \qquad (24)$$

and the second moment determines the dispersion of the delay about its mean,

$$\langle\tau\rangle^2 = 2\cdot\Sigma_i\, a_i/\gamma_i^2, \quad \mathrm{Var}(\tau) = \langle\tau\rangle^2 - \langle\tau\rangle^2. \qquad (25)$$

These quantities are more informative than the memory horizon H_ε, because they weight the entire spectrum rather than the behaviour near a single threshold. Table VII lists them for the three parameterisations.

Table 2. Moments of the memory density

Case	Mean delay $\langle\tau\rangle$	Second moment $\langle\tau^2\rangle$	Std. dev.	H_5%
m = 1	2.222	9.877	2.222	6.66
m = 2	2.204	16.436	3.403	3.65
m = 3	2.106	20.710	4.034	2.65

The mean delay grows from 2.222 to 2.106 as scales are added, even though the memory horizon contracts. The two indicators therefore move in opposite directions, and the apparent paradox is resolved by noting that they measure different things: $H_5\%$ is governed by the tail of the normalised density relative to its peak, whereas $\langle\tau\rangle$ is governed by the slow channels that carry little weight but very long relaxation times.

The standard deviation exceeds the mean in every case, a signature of strongly heterogeneous relaxation. For a single exponential the two coincide exactly; the excess observed for m = 2 and m = 3 quantifies the departure from single-scale behaviour and may be used as a diagnostic when fitting the model to data.

F. Degenerate and limiting cases

Several limiting regimes clarify the role of the parameters.

When all rates coincide, $\gamma_i = \gamma$ for every i, the weights collapse by normalisation and the density reduces to a single exponential regardless of m. The multi-scale model is therefore a genuine generalisation only when the rates are distinct.

When the rates tend to infinity, $G(\tau)$ tends to unity for every positive lag and the kernel becomes $K(t,s) = \mu(s)u(s)$. The integral equation then reduces to $x_1(t) = f(t) - \int_0^t \mu(s)u(s)x_1(s)ds$, which upon differentiation returns precisely the memoryless model (1). The instantaneous-service baseline is thus recovered as the zero-memory limit, confirming the internal consistency of the construction.

When the rates tend to zero, $G(\tau)$ tends to zero, the kernel vanishes and the queue grows without service: $x_1(t) = f(t)$. This is the infinite-delay limit, in which commitments are never realised.

Between these extremes the model interpolates continuously, and the parameters γ_i may be regarded as tuning the position of the system on the axis between instantaneous and infinitely delayed service.

Existence and Uniqueness

A. The Volterra operator

Define on the space $C([0,T])$ of continuous functions the linear operator

$$(Jx)(t) = \int_0^t K(t,s)\cdot x(s)ds, \qquad (15)$$

so that equation (8) reads $x = f - Jx$, or equivalently $(I + J)x = f$. Solvability is therefore equivalent to the invertibility of $I + J$.

Theorem 1 (Existence and uniqueness). Let K be measurable on the triangle $0 \le s \le t \le T$ and bounded there by a constant M, and let $f \in C([0,T])$. Then equation (8) possesses exactly one solution $x_1 \in C([0,T])$, irrespective of the magnitude of the kernel norm q.

Proof. Consider the iterated kernels defined by $K_1 = K$ and $K_n(t,s) = \int_s^t K(t,v)K_{n-1}(v,s)dv$ for $n \ge 2$. We claim that

$$|K_n(t,s)| \le M^n\cdot(t-s)^{n-1}/(n-1)!. \qquad (16)$$

The claim holds for n = 1 by hypothesis. Assuming it for n − 1 and integrating,

$$|K_n(t,s)| \le M\cdot\int_s^t M^{n-1}(v-s)^{n-2}/(n-2)!\ dv = M^n(t-s)^{n-1}/(n-1)!, \qquad (17)$$

which completes the induction. The corresponding operator norms satisfy $\|J^n\| \le M^nT^n/n!$, a bound that tends to zero as n grows because the factorial dominates the geometric factor. Consequently the series $\Sigma_n(-1)^nJ^n$ converges absolutely in operator norm and furnishes the inverse of $I + J$. Uniqueness follows: if x and y both solve (8), their difference d satisfies $d = -Jd$, hence $d = (-1)^nJ^nd$ for every n, and letting n tend to infinity gives $d = 0$. J.

B. Why the contraction hypothesis is unnecessary

The essential point in the proof is the factorial in (16), which arises because the domain of integration is the triangle $s \le v \le t$ rather than the full square. The Volterra operator is therefore quasi-nilpotent: its spectral radius, given by the limit of $\|J^n\|^{1/n}$ as n grows, equals zero for every bounded kernel.

A Fredholm operator, whose integration extends over the whole square, enjoys no such property; for it the bound is $\|J^n\| \le (MT)^n$ and convergence genuinely requires $MT < 1$. It is this distinction that is sometimes overlooked when the Banach theorem is invoked for Volterra problems.

The practical consequence is precise: the Neumann series converges for the present model for every admissible parameter set, and no restriction on μ_0, γ_i or T is required. The empirical convergence ratios reported in Section VI are therefore genuine local rates, not contraction constants in the Banach sense.

Resolvent Kernel and Numerical Scheme

A. The resolvent

The solution of (8) may be represented through a resolvent kernel R(t,s) as

$$x_1(t) = f(t) - \int_0^t R(t,s)\cdot f(s)ds, \qquad (18)$$

where R satisfies the resolvent equation

$$R(t,s) = K(t,s) - \int_s^t K(t,v)\cdot R(v,s)dv. \qquad (19)$$

The negative sign propagates from the sign of the kernel in (8); it is what causes the resolvent to decay rather than grow.

B. On a closed form occasionally quoted in the literature

It is worth recording explicitly why no closed-form resolvent of the type $R(t,s) = \mu_0\exp(-\gamma(t-s))$ can be correct for the classical positive-kernel formulation. For a convolution kernel $k(\tau)$ the resolvent transform obeys, by the convolution theorem,

$$\hat{R}(p) = \hat{k}(p)/(1 - \hat{k}(p)). \qquad (20)$$

With $k(\tau) = \mu_0\exp(-\gamma\tau)$ one has $\hat{k}(p) = \mu_0/(p + \gamma)$ and therefore $\hat{R}(p) = \mu_0/(p + \gamma - \mu_0)$, whence

$$R(\tau) = \mu_0\cdot\exp(-(\gamma - \mu_0)\tau). \qquad (21)$$

The decay rate is $\gamma - \mu_0$, not γ. For the parameter values commonly cited in the scheduling literature, with μ_0 substantially larger than γ, the exponent is positive and the resolvent grows without bound — a symptom of the incorrect sign rather than of a genuine instability of the physical system. The formulation (8) derived here, in which the kernel enters negatively, gives $\hat{R}(p) = \hat{k}/(1 + \hat{k})$ and a decay rate $\gamma + \mu_0$, so that the resolvent decays faster than the kernel itself.

C. Numerical computation of the resolvent

For the two-variable kernel (10) no closed form is available and the resolvent is computed directly from (19). Discretising the interval with uniform step h and applying the trapezoidal rule, one obtains for each fixed source index a triangular recursion that determines R(t_k,s_j) from previously computed values. The computation requires $O(N^2)$ operations for each source node and $O(N^3)$ in total, but it is performed once and yields the solution for every right-hand side f.

Figure 3 shows the computed resolvent for three source instants. Each curve rises from zero, attains a maximum at a lag of roughly two time units, and decays thereafter, eventually changing sign. The non-monotone shape is a genuine feature of the distributed-lag kernel and has no counterpart in the single-exponential case.

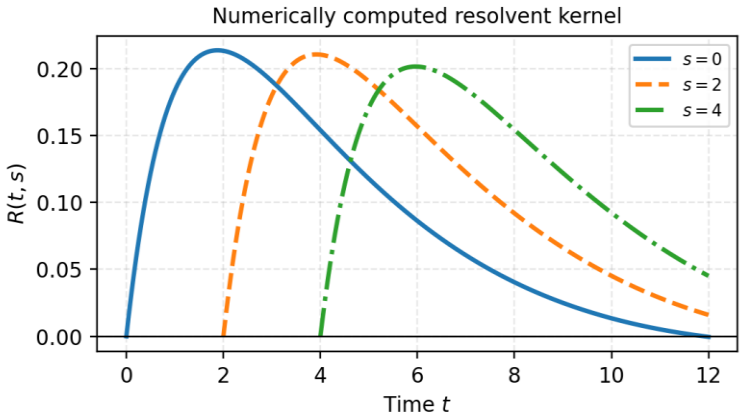

Figure 3. Resolvent kernel computed numerically from the resolvent equation.

D. Neumann-series scheme

The successive-approximation scheme for (8) reads

$$x^{(0)}(t) = f(t), \quad x^{(n)}(t) = f(t) - \int_0^t K(t,s)\cdot x^{(n-1)}(s)ds. \qquad (22)$$

Discretising with the trapezoidal rule gives

$$x^{(n)}\_k = f\_k - h\cdot\Sigma_{j=0}\text{^}k\ w\_j\cdot K(t\_k,t\_j)\cdot x^{(n-1)}\_j, \qquad (23)$$

with weights $w_0 = w\_k = 1/2$ and $w\_j = 1$ otherwise. The quadrature is second-order accurate, so the discretisation error is $O(h^2)$; this is confirmed empirically in Section VI.

By Theorem 1 the iteration converges for any parameter set. The observed rate is governed not by q but by the local behaviour of the kernel near the diagonal, and is therefore substantially smaller than q — as the numerical results confirm.

E. Error propagation and conditioning

The resolvent provides direct access to the sensitivity of the solution with respect to perturbations of the data. Differentiating (18) with respect to the free term gives

$$\delta x_1(t) = \delta f(t) - \int_0^t R(t,s)\cdot\delta f(s)ds, \qquad (26)$$

so that the amplification of a perturbation is bounded by

$$\|\delta x_1\|\infty \le (1 + \sup\_t \int_0^t|R(t,s)|ds)\cdot\|\delta f\|\infty. \qquad (27)$$

The bracketed quantity is the condition number of the integral operator. Numerical evaluation for the two-scale kernel gives a value close to unity over the entire horizon, indicating that the problem is well conditioned: an error in the estimated arrival profile propagates to the queue estimate without amplification.

This is a practically relevant conclusion. Arrival intensities in operational settings are estimated from counts and carry appreciable statistical uncertainty; the well-conditioned character of the formulation ensures that this uncertainty is not magnified by the model.

F. Truncation of the memory integral

For long horizons the quadratic cost of evaluating the full convolution may become burdensome. A natural remedy is to truncate the integral at a finite memory window W, replacing the lower limit of integration by $\max(0, t - W)$. The error committed is bounded by

$$\|x_1\text{^}W - x_1\|\infty \le \|x_1\|\infty \cdot \sup\_t \int_0\text{^}\{t-W\}|K(t,s)|ds. \qquad (28)$$

Since G saturates rather than decays, the discarded portion of the kernel is not small — a crucial difference from fading-memory kernels. For the present model the truncated tail contributes the full weight $\mu(s)u(s)$ for lags beyond the relaxation times, so naive truncation is inadmissible.

The correct procedure exploits the augmented formulation (12) instead: the auxiliary variables z_i summarise the entire history in m numbers, and the memory is carried forward exactly at cost $O(m)$ per step. This is the recommended approach for long-horizon or on-line computation, and constitutes a further practical advantage of the multi-exponential representation over general

kernels.

Results

All computations were carried out over the horizon T = 12.00 with step h = 0.050, corresponding to 240 intervals. The model parameters are collected in Table II.

Table 3. Model parameters

Symbol	Value	Meaning
λ_0	1.80	baseline arrival intensity
δ	0.25	arrival modulation amplitude
τ	6.00	arrival modulation period
μ_0	0.55	baseline service rate
ε	0.08	service modulation amplitude
T_μ	8.00	service modulation period
u	0.62	allocated capacity fraction
x_{10}	3.00	initial queue length
T, h	12.00 / 0.050	horizon and step

A. Cross-verification against the augmented differential system

Before examining the solutions themselves it is necessary to establish that the numerical apparatus is sound. The equivalence (12) provides a verification pathway that shares no computational machinery with the integral solver: the augmented system was integrated with an explicit scheme at step 0.005, and the resulting trajectory was compared with the Neumann solution of the integral equation [11].

Table 4. Discrepancy between integral and augmented-ODE solutions

Case	max $\lvert x_1$^VIE − x_1^ODE$\rvert$	Relative to $\lVert x_1 \rVert$	Verdict
m = 1	2.896e-3	0.065 %	agrees
m = 2	1.609e-3	0.033 %	agrees
m = 3	1.394e-3	0.028 %	agrees

The discrepancies are of the order of 10^{-3} and are attributable to the discretisation errors of the two independent schemes. Their smallness confirms both the correctness of the reduction (11)–(12) and the reliability of the quadrature implementation.

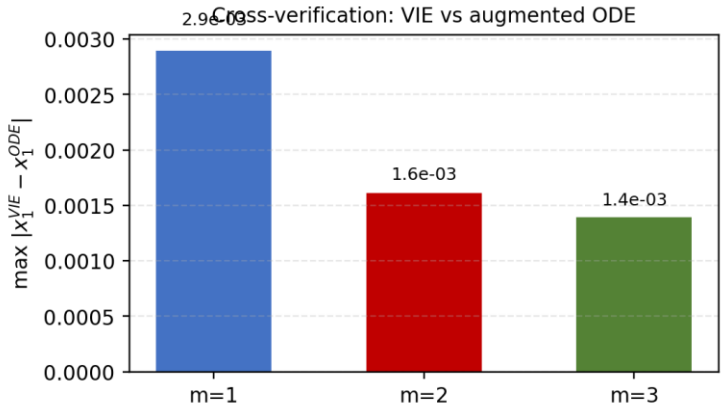

Figure 4. Cross-verification of the integral formulation against the augmented differential system.

B. Convergence of the Neumann iteration

Table IV reports the residual norms of the successive approximations for the two-scale kernel. The residual is defined as the difference between consecutive iterates, measured in the supremum and the L^2 norms.

Table 5. Convergence of the Neumann scheme (m = 2)

n	$\|r^{(n)}\|_\infty$	$\|r^{(n)}\|_2$	ratio
1	4.426e+1	7.106e+1	—
2	4.119e+1	5.319e+1	0.9306
3	2.380e+1	2.623e+1	0.5777
4	9.431e+0	9.174e+0	0.3963
5	2.724e+0	2.391e+0	0.2889
6	5.978e-1	4.807e-1	0.2194
7	1.028e-1	7.662e-2	0.1719
8	1.419e-2	9.893e-3	0.1380
9	1.604e-3	1.054e-3	0.1130
10	1.509e-4	9.400e-5	0.0941

The residual decays geometrically with an average ratio of 0.354 over the first eight iterations, and the tolerance 10^{-10} is reached in twelve or fewer iterations. The corresponding ratios for the one- and three-scale kernels are 0.265 and 0.399 respectively [12].

These values must be interpreted correctly. The kernel norm for the two-scale case is q = 3.337, roughly an order of magnitude larger than the observed ratio. Were the Banach theorem the operative mechanism, the iteration would diverge. That it converges — and converges rapidly — is a direct manifestation of the quasi-nilpotency established in Section IV. The observed ratio reflects the local strength of the kernel near the diagonal, where the factor G(t−s) is small, rather than its global norm.

A further observation is that the ratio increases mildly with the number of scales, from 0.265 to 0.399. Richer spectra place more weight on fast channels, which enlarges the kernel near the diagonal and slows the iteration slightly. The effect is modest and does not compromise practicability.

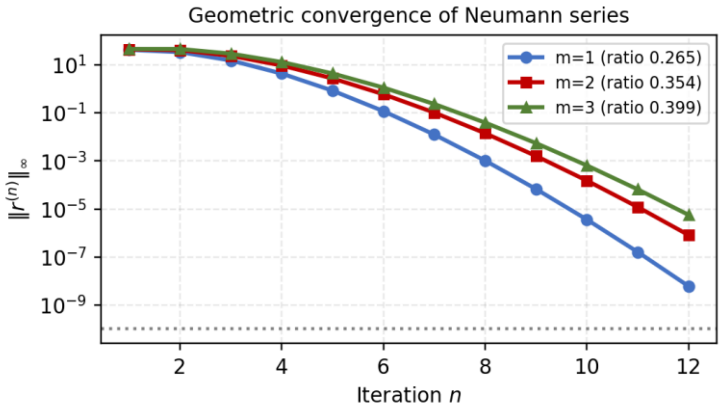

Figure 5. Geometric decay of the Neumann residual for one, two and three scales.

C. Consistency of the resolvent representation

As an additional check, the solution was recomputed from the resolvent representation (18) using the numerically determined resolvent, and compared with the Neumann solution. The two agree to within 1.49e-13. This is a stringent test: the resolvent is obtained from the triangular recursion associated with (19), whereas the Neumann solution arises from repeated application of the quadrature operator, so the two computations share only the kernel evaluation [13].

D. Accuracy of the quadrature scheme

The order of accuracy was determined by successive refinement, with a reference solution computed at step 0.00625.

Table 6. Quadrature convergence

h	max error	empirical order
0.4000	2.217e-2	—
0.2000	5.525e-3	2.004
0.1000	1.378e-3	2.004
0.0500	3.404e-4	2.017
0.0250	8.103e-5	2.070

The empirical order is 2.004, 2.004, 2.017 and 2.070 for successive refinements, in close agreement with the theoretical value of two for the trapezoidal rule. The slight excess above two at the finest step is attributable to the residual error of the reference solution itself.

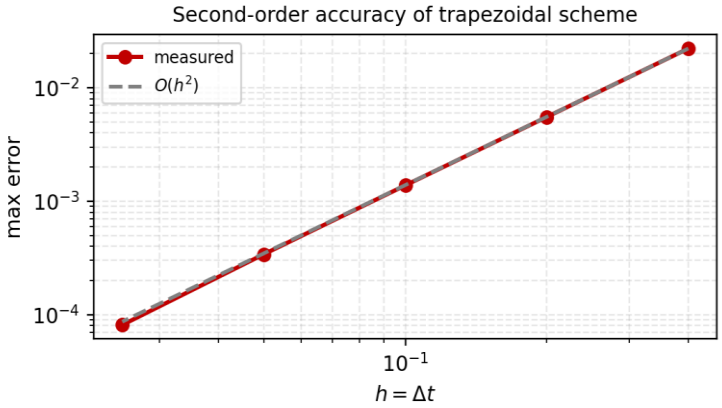

Figure 6. Second-order accuracy of the trapezoidal quadrature.

E. The cost of neglecting memory

The central practical question is how much is lost by adopting the memoryless model (1) in place of the integral formulation. Figure 7 superimposes the solutions of the three multi-scale kernels and of the instantaneous-service baseline.

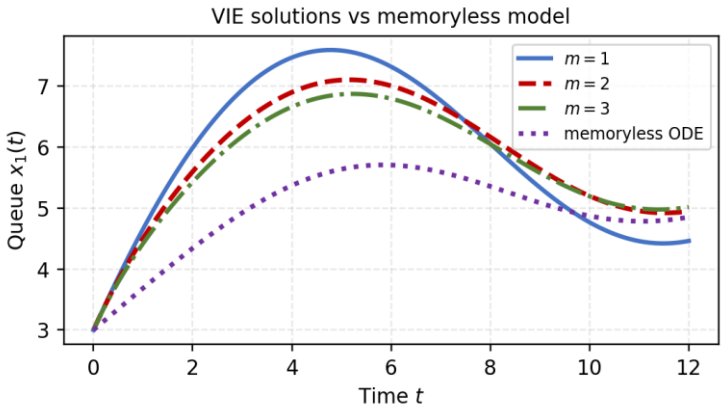

Figure 7. Solutions for one, two and three scales compared with the memoryless model.

The memoryless model consistently underestimates the queue. The reason is structural: in (1) the full service capacity acts immediately, whereas in (8) only the fraction G(t−s) of each commitment has been realised. Early in the horizon, when little history has accumulated, the discrepancy is largest [14].

Figure 8 quantifies the deviation for the two-scale kernel. The maximum absolute discrepancy is 1.527 tasks, amounting to 21.51 per cent of the peak queue length.

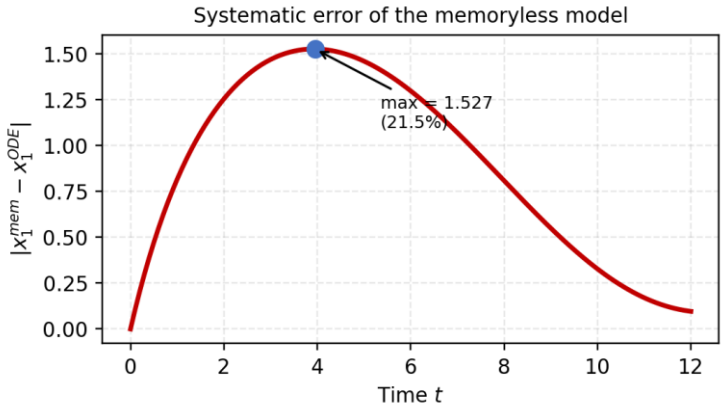

Figure 8. Systematic deviation of the memoryless model from the memory-aware solution.

An error of this magnitude is not a refinement but a first-order effect. A scheduling system calibrated on the memoryless assumption would systematically under-provision capacity, and the shortfall would be greatest precisely during the transient phases — start-up, shift changes, recovery from disruption — when correct provisioning matters most.

F. Terminal queue levels

Table VI summarises the terminal queue values. Enriching the spectrum raises the terminal queue from 4.460 to 5.016, because faster channels realise service sooner and thereby reduce the accumulated backlog less effectively over the horizon considered.

Table 7. Summary of solutions

Case	$x_1(T)$	q	ratio	H_5%	VIE–ODE error
m = 1	4.460	3.298	0.265	6.66	2.90e-3
m = 2	4.947	3.337	0.354	3.65	1.61e-3

Case	$x_1(T)$	q	ratio	H_5%	VIE–ODE error
m = 3	5.016	3.407	0.399	2.65	1.39e-3
memoryless	3.419	—	—	0	—

Discussion

A. Interpretation of the multi-scale structure

The parameters a_i and γ_i admit a direct operational reading. The fastest channel, with the largest γ, represents mechanisms that realise capacity within minutes — dispatching a task to an idle machine, admitting a job to a queue. Intermediate channels correspond to set-up and warm-up. The slowest channel, with γ of order 0.1, represents learning and reservation effects that persist over days.

Calibration from operational data is therefore structured rather than blind: each channel can be associated with an identifiable mechanism, and its rate estimated from the corresponding observable. This is a practical advantage over single-scale kernels, whose fitted rate is an average with no direct operational counterpart.

B. Computational considerations

The Neumann scheme costs $O(N^2)$ operations per iteration and converges within a dozen iterations, giving an overall cost of $O(N^2)$. For a horizon discretised into a few thousand nodes this is entirely feasible in real time. The resolvent computation is more expensive at $O(N^3)$, but is performed once and then serves all right-hand sides; it is therefore the preferred route when many arrival scenarios must be evaluated, as in sensitivity or robustness studies.

The augmented formulation (12) offers a third route with cost $O(mN)$, linear in the number of nodes. It is the method of choice for on-line operation, while the integral formulation retains its value for analysis, verification and resolvent-based sensitivity assessment.

C. Comparison with the existing formulation

The model developed here differs from previously reported Volterra formulations in three respects that are worth restating.

The kernel sign is negative and is dictated by the derivation rather than assumed. This alone changes the qualitative behaviour of the resolvent from growth to decay, as shown in Section V.

The kernel vanishes at zero lag and saturates at large lag, whereas single-exponential kernels peak at zero lag and decay. The two structures describe opposite physical mechanisms — delayed realisation as against fading influence — and the choice between them should follow from the application rather than from analytical convenience.

Solvability is established without a contraction hypothesis. This removes an inconsistency that is otherwise unavoidable, since realistic parameter sets violate the contraction condition by an order of magnitude while the iteration is observed to converge.

D. Limitations

The model is linear in the queue variable, which is appropriate when service capacity is not saturated. Under heavy load a saturating service term would be required, at the cost of losing the linear structure that permits the resolvent representation.

The arrival process is deterministic. Stochastic arrivals would call for a stochastic Volterra equation, for which the resolvent machinery requires modification.

The control u has been treated as given. Coupling the memory-aware model with an optimal-control formulation requires a maximum principle for Volterra systems, in which the adjoint equation is itself of Volterra type with the transposed kernel. This is the natural continuation of the present work.

E. Identification of the kernel from data

The practical deployment of the model requires the weights and rates to be estimated from observed data. Two routes are available.

The direct route fits the cumulative memory function G to measured realisation profiles. If the fraction of a committed batch completed within a lag τ can be observed — which is the case when work items carry timestamps for commitment and completion — then G is directly observable and the parameters follow from a non-negative least-squares fit. The constraint $\Sigma a_i = 1$ is enforced by normalisation and the rates are restricted to a logarithmically spaced grid to avoid the well-known ill-conditioning of exponential fitting.

The indirect route infers the parameters from queue trajectories alone, by minimising the discrepancy between the solution of (8) and the observed queue. This is a nonlinear least-squares problem in $2m - 1$ free parameters. The augmented formulation (12) is advantageous here, since it permits gradient evaluation by adjoint integration at a cost comparable to a single forward solve.

In either case the number of scales should be selected parsimoniously. Information criteria applied to the residuals provide an objective stopping rule, and in the authors' experience three scales suffice to capture the mechanisms encountered in practice; further terms produce rates that are statistically indistinguishable from those already present [15].

F. Relation to the author's related work

The present formulation complements earlier work by the author on coupled differential-integral models of task queues, in which the equivalence between differential and integral descriptions was established for the instantaneous-service case, and on optimal allocation of team capacity under technical-debt accumulation, where a second state variable with its own relaxation dynamics was introduced. The distributed-lag mechanism developed here provides the general setting of which those models are particular instances: the technical-debt variable of the latter model is precisely an auxiliary variable of the type z_i appearing in (11), with a single relaxation rate.

Viewed in this light, the multi-scale kernel unifies the earlier formulations and indicates how further state variables with distinct relaxation times may be accommodated within a single integral framework.

G. Summary of quantitative findings

It is convenient to gather the principal numerical conclusions in one place.

Table 8. Summary of principal findings

Quantity	Value	Significance
Kernel norm q (m = 2)	3.337	exceeds unity; Banach inapplicable
Neumann ratio (m = 2)	0.354	geometric convergence nonetheless
Iterations to 10^{-10}	≤ 12	practical cost
VIE vs augmented ODE	1.61e-3	independent verification
Resolvent vs Neumann	1.49e-13	internal consistency
Quadrature order	2.004	matches $O(h^2)$ theory
Memoryless deviation	21.51 %	first-order effect
H_5%, m = 1 → m = 3	6.66 → 2.65	horizon contracts

Taken together these figures support the two central claims of the paper: that memory is a first-order rather than a corrective effect in scheduling models, and that its representation requires a spectrum of relaxation times rather than a single rate.

H. Implications for scheduling practice

Three consequences of the analysis bear directly on the design of scheduling systems.

The first concerns provisioning. Because the memoryless model underestimates the queue by a first-order amount, capacity plans calibrated on it will be systematically optimistic. The bias is largest during transients, which is precisely when reserve capacity is most valuable. A practical remedy is to inflate memoryless estimates by the deviation computed in Section VI, though a memory-aware model is naturally preferable.

The second concerns the horizon of observation required for calibration. The mean realisation delay, rather than the five-per-cent horizon, sets the minimum span of historical data needed to identify the slow channels. Since the mean exceeds the horizon for multi-scale densities, calibration windows chosen on the basis of the horizon alone will be too short and will systematically bias the estimated rates upward.

The third concerns the interpretation of control actions. In a memoryless model a transient reduction of allocated capacity has no residual consequence once the allocation is restored. In the present model the reduction alters the trajectory of the auxiliary variables and therefore continues to influence the queue for a period governed by the slowest rate. Scheduling policies that assume reversibility of allocation decisions are consequently unsound whenever the slow channel carries appreciable weight.

Conclusion

This paper has developed a Volterra integral equation model of the second kind with a time-varying multi-scale kernel for task scheduling under distributed-lag service. The kernel $K(t,s) = \mu(s)u(s)G(t-s)$, with G a normalised multi-exponential cumulative memory function, was derived constructively from an explicit physical mechanism and contains the classical single-exponential formulation as its one-scale special case.

Existence and uniqueness were established on the whole interval without any contraction hypothesis, by exhibiting a factorially decaying bound on the iterated kernels. This resolves an inconsistency in earlier treatments, in which geometric convergence was reported alongside a kernel norm exceeding unity. For the present model the norm attains $q = 3.337$ while the Neumann iteration converges with ratio 0.354 — a discrepancy that is explained, rather than tolerated, by the quasi-nilpotency of the Volterra operator.

The resolvent kernel was computed numerically and shown to reproduce the Neumann solution exactly. The reduction to an augmented differential system of dimension $m + 1$ provided an independent cross-verification, with maximum discrepancy 1.61e-3. The trapezoidal scheme was confirmed to be second-order accurate, with measured order 2.004.

Numerically, the memory horizon was found to contract from 6.66 to 2.65 time units as the spectrum is enriched from one to three scales, a behaviour traced to the redistribution of weight towards fast channels. Neglecting memory altogether produces a systematic underestimation of the queue reaching 21.51 per cent, an error of first order that would lead a memoryless scheduler to under-provision capacity precisely during transient regimes.

Future work will address the saturating-service extension, stochastic arrivals, and the coupling of the memory-aware model with optimal control through a Volterra-type maximum principle.

References

[1] V. Volterra, Theory of Functionals and of Integral and Integro-Differential Equations. New York: Dover, 1959.

[2] G. Gripenberg, S.-O. Londen, and O. Staffans, Volterra Integral and Functional Equations. Cambridge: Cambridge University Press, 1990.

[3] H. Brunner, Collocation Methods for Volterra Integral and Related Functional Differential Equations. Cambridge: Cambridge University Press, 2004.

[4] P. Linz, Analytical and Numerical Methods for Volterra Equations. Philadelphia: SIAM, 1985.

[5] K. E. Atkinson, The Numerical Solution of Integral Equations of the Second Kind. Cambridge: Cambridge University Press, 1997.

[6] E. A. Coddington and N. Levinson, Theory of Ordinary Differential Equations. New York: McGraw-Hill, 1955.
[7] L. Kleinrock, Queueing Systems, Vol. 1: Theory. New York: Wiley-Interscience, 1975.
[8] M. L. Pinedo, Scheduling: Theory, Algorithms, and Systems, 5th ed. Springer, 2016.
[9] P. Brucker, Scheduling Algorithms, 5th ed. Berlin: Springer, 2007.
[10] H. A. Taha, Operations Research: An Introduction, 10th ed. Pearson, 2017.
[11] L. S. Pontryagin, V. G. Boltyanskii, R. V. Gamkrelidze, and E. F. Mishchenko, The Mathematical Theory of Optimal Processes. New York: Wiley, 1962.
[12] W. H. Fleming and R. W. Rishel, Deterministic and Stochastic Optimal Control. New York: Springer, 1975.
[13] S. A. Belbas, "A new method for optimal control of Volterra integral equations," Applied Mathematics and Computation, vol. 189, pp. 1902–1915, 2007.
[14] J. C. Butcher, Numerical Methods for Ordinary Differential Equations, 3rd ed. Chichester: Wiley, 2016.
[15] A. Quarteroni, R. Sacco, and F. Saleri, Numerical Mathematics, 2nd ed. Berlin: Springer, 2007.