Sensitivity to Serial Dependency of Input Processes:
A Robust Approach

Henry Lam*
Boston University

Abstract

We propose a distribution-free approach to assess the sensitivity of stochastic simulation with respect to serial dependency in the input model, using a notion of “nonparametric derivatives”. Unlike classical derivative estimators that rely on parametric models and correlation parameters, our methodology uses Pearson’s φ²-coefficient as a nonparametric measurement of dependency, and computes the sensitivity of the output with respect to an adversarial perturbation of the coefficient. The construction of our estimators hinges on an optimization formulation over the input model distribution, with constraints on its dependency structure. When appropriately scaled in terms of φ²-coefficient, the optimal values of these optimization programs admit asymptotic expansions that represent the worst-case infinitesimal change to the output among all admissible directions of model movements. Our model-free sensitivity estimators are intuitive and readily computable: they take the form of an analysis-of-variance (ANOVA) type decomposition on a “symmetrization”, or “time-averaging”, of the underlying system of interest. They can be used to conveniently assess the impact of input serial dependency, without the need to build a dependent model a priori, and work even if the input model is nonparametric. We report some encouraging numerical results in the contexts of queueing and financial risk management.

1 Introduction

Serial dependency appears naturally in many real-life processes in manufacturing and service systems. For simulation practitioners, understanding these dependencies in the input process is important yet challenging. Broadly speaking, there are two aspects to this problem. The first aspect, which consists of most of the literature, concerns the modeling and construction of computation-friendly input processes that satisfy desired correlation properties. For example, the study of copula [26], classical linear and non-linear time series [15], and more recently the autoregressive-to-anything (ARTA) scheme [3, 4] lies in this category. The second, and equally important, aspect is to assess

*Department of Mathematics and Statistics, Boston University. Email: khlam@bu.edu.
the impact of misspecification, or simply ignorance, of input dependency to the simulation outputs. In the literature, a study in the second aspect usually starts with a family of input processes, which contains parameters that control the dependency structure. Running simulations at various values of these parameters then gives measurements of the sensitivity to input dependency for the particular system of interest (examples are [24, 27]). Such type of study, though can be effective at times, is confined to parametric models, with simple (mostly linear) dependency structure, and often involves ad hoc tuning of the correlation parameters.

A more systematic, less model-dependent approach to addressing the second aspect above is practically important for a few reasons. First, fitting accurate time-series model as simulation inputs, especially for non-Gaussian marginals, is known to be challenging. Models that can fit any prescribed specification of both marginal and high-order cross-moment structure appear difficult to construct (for instance, [10] pointed out some feasibility issue for normal-to-everything (NORTA), the sister model of ARTA, and literature on fitting dependent input processes beyond correlation information is almost non-existent). Perhaps for this reason, most simulation softwares only support basic dependent inputs. With these constraints, it would therefore be useful to design schemes that can assess the sensitivity of input dependency, without having to construct complex dependent models \textit{a priori}. Such assessment will detect the adequacy of simple i.i.d. models or the need to call on more sophisticated modeling. It can also be translated into proper adjustments to simulation output intervals, without, again, the need of heavy modeling tools. Currently, the difficulties in building dependent models confine sensitivity analysis to almost exclusively correlation-type parameters, which can at best capture linear effects. Subtle changes in higher order dependency that is not parametrizable are out of the scope in sensitivity studies.

With such motivations, we attempt to build a sensitivity analysis framework for input dependency that is distribution-free and is capable of handling dependency beyond linear correlation. Our methodology is a local analysis, i.e. we take a viewpoint analogous to derivative estimation, and aim to measure the impact of small perturbation of possibly nonlinear dependency structure (in certain well-defined sense) on the system output. However, unlike classical derivative estimation, our methodology is not tied to any parametric input models, nor any parameters that control the level of dependency. The main idea, roughly, is to replace the notion of Euclidean distance in model parameters by some general statistical measures of dependency, which we shall use $\chi^2$-distance in this paper. That is, the sensitivity that we will focus on is with respect to changes in the $\chi^2$-distance of the input model.

Despite the considerable amount of literature in estimation and hypothesis testing using statistical distances like $\chi^2$, their adoption as a sensitivity tool in simulation appears to be wide open. The most closely related work is [22], which considered only the simplest case of independent inputs.
and under entropy discrepancy. In this paper, we shall substantially extend their methodology to assess model misspecification against serial dependency, which has an important broader implication: we demonstrate that by making simple tractable changes to the formulation (more discussion below), the resulting estimators in this line of analysis can be flexibly targeted at specific statistical properties, as long as they can be captured by these distances; serial dependency is, of course, one such important statistical property. This, we believe, opens the door to many model assessment schemes for features that are beyond the reach of classical derivative estimation.

To be more specific, in this paper we shall consider stationary input models that are $p$-dependent: a $p$-dependent process has the property that, conditioned on the last $p$ steps, the current state is independent of all states that are more than $p$ steps back in the past. The simplest example of a $p$-dependent process is an autoregressive process of order $p$; in this sense, the $p$-dependence definition is a natural nonlinear generalization of autoregression. In econometrics, the use of $\chi^2$-distance, or other similar types of measures such as entropy, has been used as basis for nonparametric test against such type of dependency [14, 13, 5]. The key feature that was taken advantage of is that the statistical distance between the joint distribution of consecutive states and the product of their marginals, i.e. the so-called Pearson’s $\phi^2$-coefficient in the case of $\chi^2$-distance, or mutual information in the case of Kullback-Leibler divergence, possesses asymptotic properties that lead to consistent testing procedure [20, 31]. One message in this paper is that this sort of properties can also be brought in quantifying the degree in which serial dependency can distort the output of a system.

Let us highlight another salient feature of our methodology that is distinct from classical derivative estimation: ours is robust against the worst-case misspecification of dependency structure. Intuitively, with specification only on the $\chi^2$-distance, there are typically many (in fact, infinite) directions of movement that perturbations can occur. Without knowledge on the directions of movement, we shall naturally look at the “steepest descent” direction, which is equivalent to an adversarial change in the serial dependency behavior. For this reason, an important intermediate step in our construction is to post proper optimization problems over the probability measures that generate the input process, constrained on the dependency structure measured by $\chi^2$-distance. Setting up these constraints is the key to our methodology: once this is done properly, we can show that, as the $\chi^2$-distance shrinks to zero, the optimal values of the optimization programs are asymptotically tractable in the form of Taylor-series type expansion. The coefficients in these expansions control exactly the worst-case rates of change of the system output among all admissible changes of dependency, hence giving rise to a natural robust form of “sensitivities”.

These “sensitivities” as the end-products of our analysis are conceptually intuitive and also computable. Moreover, they can be applied to a broad range of problems with little adjustment; in Section 6, we demonstrate these advantages through several numerical examples. Here, we shall
discuss the intuition surrounding the mathematical form of these sensitivity estimators, which has some connection to the statistics literature. They are expressed as the moments of certain random objects that are, in a sense, a summary of the effect of input models based on their roles in the system of interest. On a high level, these random objects comprise of two transformations of the underlying system. The first is a “symmetrization”, or “time-averaging”, of the system, which captures the frequency of appearances of the input model. A system that calls for more replications of a particular input model tends to be more sensitive; as we will see, this can be measured precisely by computing a conditional expectation of the system that involves a randomization of the time horizon. We point out that such “symmetrization” is closely related to so-called influence function in the robust statistics literature [16, 17]. The latter has been used as an assessment tool for outliers and other forms of data contamination. Our “symmetrization” and the influence function are two sides of the same coin, from the dual perspective of a stochastic system as a functional of the input random sources or a statistic as a functional of the random realization of data.

The second transformation involved in our estimators is a filtering of the marginal effect of the input process, thereby singling out the sensitivity solely with respect to serial dependency. This comprises the construction of an analysis-of-variance (ANOVA) type decomposition [6] of our symmetrization, the factors of the decomposition being, intuitively speaking, the marginal values of the input process. As we shall see, our first order sensitivity estimator takes precisely the form of the residual standard deviation after removing the “main effect” of these marginal quantities. In other words, after our symmetrization transformation, one can interpret the dependency effect from a stationary input process as the “interaction effect” across the homogeneous groups represented by the identically distributed marginal variables.

We close this introduction by discussing several related work that is used to obtain the form of the sensitivity estimators in this paper. As described, our results come from an asymptotic analysis on the optimal values of our posted optimization programs (on the space of probability measures). Such optimizations are in the spirit of distributionally robust optimization (e.g. [12, 7]) and robust control theory (e.g. [18, 30]). Broadly speaking, they are stochastic optimization problems that are subject to uncertainty of the true probabilistic model, or so-called model ambiguity [28, 8]. In particular, we derive a characterization of the optimal solutions for $\chi^2$-constrained programs similar to [29, 21], but with some notable new elements. Because of typically multiple replications in calling the input model when computing a large-scale stochastic system, the objective in our optimization is in general non-convex, which is in contrast to most robust optimization formulations (including theirs). We overcome this issue by adopting the fixed point technique in [22] that leads to asymptotically tractable optimal solution. Finally, we also note a recent work by [11], who proposed a scheme called robust Monte Carlo to compute robust bounds for estimation and
optimization problems in finance. Their results are related to ours in two ways: First, part of their work considered bivariate variables with conditional dependency, and our derivation in Section 3 bears similarity to theirs. The former, however, is not directly applicable to long sequences of stationary inputs with serial dependency structure, a setting that is studied in this paper. Second, they considered using an aggregate relative entropy measure to handle deviation due to model uncertainty, along the line of robust control theory and which leads to neat closed-form exponential changes-of-measure. As we will discuss in Section 6, in certain situations, our technique in this paper can be viewed as a more feature-targeted (and hence less conservative) alternative to their approach, when the transition structure of an input process is exploited to refine the estimate on the model misspecification effect.

2 Problem Framework

We first introduce some terminology. For convenience, we denote a cost function $h : \mathcal{X}^T \to \mathbb{R}$, where $\mathcal{X}$ is an arbitrary (measurable) space. In stochastic simulation, the typical goal is to compute a performance measure $E[h(X_T)]$, where $X_T = (X_1, \ldots, X_T) \in \mathcal{X}^T$ is a sequence of random objects, and $T$ is the time horizon. The cost function $h$ is assumed known (i.e. capable of being evaluated by computer, but not necessarily having closed form). The random sequence $X_T$ is generated according to some input model. For concreteness, we call $P_0$ the benchmark model that the user assumes as the input model, on which the simulation of $X_T$ is based. As an example, in the queueing context, $X_T$ can denote the sequence of interarrival and service times, and the cost function can denote the average waiting times of, say, the first 100 customers, so that $T = 100$. In this case, a typical $P_0$ generates an i.i.d. sequence of $X_T$. For most part of this paper, we shall concentrate on the scenario of i.i.d. $P_0$. We provide discussion on the situation where $P_0$ is non-i.i.d. in Section 4.4.

To carry out sensitivity analysis, we define a perturbed model $P_f$ on $X_T$, and our goal is to assess the change in $E_f[h(X_T)]$ relative to $E_0[h(X_T)]$, when $P_f$ is within a neighborhood of $P_0$. As discussed in the introduction, we shall set up optimization programs in order to make an adversarial assessment. On a high level, we introduce the optimization pair

\[
\begin{align*}
\max / \min \quad & E_f[h(X_T)] \\
\text{subject to} \quad & P_f \text{ is within an } \eta \text{-neighborhood of } P_0 \\
& \{X_t\}_{t=1,\ldots,T} \text{ is a } p \text{-dependent stationary process under } P_f \\
& P_f \text{ generates the same marginal distribution as } P_0 \\
& P_f \in \mathcal{P}_0.
\end{align*}
\]

The decision variable in the maximization (and minimization) is $P_f$. The second and the third constraints restrict attention to the particular dependency structure that the user wants to inves-
tigate: the second constraint states the dependency order of interest, while the third constraint keeps the marginal distributions of $X_t$'s unchanged from $P_0$ in order to isolate the dependency effect. The last constraint aims to restrict $P_f$ to some reasonable family $\mathcal{P}_0$, which will be the set of measures that are absolutely continuous with respect to $P_0$. The first constraint confines $P_f$ to be close to $P_0$, or in other words, that the sequence $\{X_t\}_{t=1}^{T}$ behaves not far from being i.i.d. This neighborhood will be expressed as a bound in terms of $\chi^2$-distance and is parametrized by the parameter $\eta$. When $\eta = 0$, the feasible set of objective values will reduce to $E_0[h(X_T)]$, the benchmark performance measure. When $\eta > 0$, the maximum and minimum objective values under these constraints will give a robust interval on the performance measure when $P_f$ deviates from $P_0$, in terms of $\eta$.

To describe how we define $\eta$-neighborhood in the first constraint of (1), consider the setting when $p = 1$, which will serve as the main illustration of our methodology in Section 4. Recall that a $p$-dependent stationary time series has the following property: for any $t$, conditioning on $X_t, X_{t+1}, \ldots, X_{t+p-1}$, the sequence $\{\ldots, X_{t-2}, X_{t-1}\}$ and $\{X_{t+p}, X_{t+p+1}, \ldots\}$ are independent. Therefore, in the case that $p = 1$, under stationarity, it suffices to specify the distribution of two consecutive states, which completely characterizes the process. Keeping this in mind, we shall define an $\eta$-neighborhood of $P_0$ to be any $P_f$ that satisfies $\chi^2(P_f(X_{t-1}, X_t), P_0(X_{t-1}, X_t)) \leq \eta$, where $\chi^2(P_f(X_{t-1}, X_t), P_0(X_{t-1}, X_t))$ is the $\chi^2$-distance between $P_f$ and $P_0$ on the joint distribution of $(X_{t-1}, X_t)$, for any $t$, under stationarity, i.e.

$$\chi^2(P_f(X_{t-1}, X_t), P_0(X_{t-1}, X_t)) = E_0\left(\frac{dP_f(X_{t-1}, X_t)}{dP_0(X_{t-1}, X_t)} - 1\right)^2.$$  (2)

Here $\frac{dP_f(X_{t-1}, X_t)}{dP_0(X_{t-1}, X_t)}$ is the Radon-Nikodym derivative, or the likelihood ratio, between $P_f(X_{t-1}, X_t)$ and $P_0(X_{t-1}, X_t)$.

The quantity (2) can be written alternatively in terms of so-called Pearson’s $\phi^2$-coefficient. Recall that $P_0$ generates i.i.d. copies of $X_t$’s, and that under the third constraint in (1), $P_0$ and $P_f$ generate the same marginal distribution. Under this situation, the $\chi^2$-distance given by (2) is equivalent to Pearson’s $\phi^2$-coefficient on $P_f(X_{t-1}, X_t)$, defined as

$$\phi^2(P_f(X_{t-1}, X_t)) = \chi^2(P_f(X_{t-1}, X_t), P_f(X_{t-1})P_f(X_t))$$  (3)

where $P_f(X_{t-1})P_f(X_t)$ denotes the product measure of the marginal distributions of $X_{t-1}$ and $X_t$ under $P_f$. Note that $\phi^2$-coefficient in (3) precisely measures the $\chi^2$-distance between a joint distribution and its independent version. The equivalence between (2) and (3) can be easily seen by

$$\phi^2(P_f(X_{t-1}, X_t)) = \chi^2(P_f(X_{t-1}, X_t), P_f(X_{t-1})P_f(X_t)) = \chi^2(P_f(X_{t-1}, X_t), P_0(X_{t-1})P_0(X_t)) = \chi^2(P_f(X_{t-1}, X_t), P_0(X_{t-1}, X_t)).$$
With these structural setups, our main result is that, when \( \eta \) shrinks to 0, the maximum and minimum values of (1) each can be expressed as

\[
E_{f^*}[h(X_T)] = E_0[h(X_T)] + \xi_1(h, P_0)\sqrt{\eta} + \xi_2(h, P_0)\eta + \cdots
\]

where \( \xi_1(h, P_0), \xi_2(h, P_0), \ldots \) are well-defined, computable quantities in terms of the cost function \( h \) and the benchmark \( P_0 \). These quantities guide the behavior of the performance measure when \( P_f \) moves away from i.i.d. to 1-dependent models. In particular, the first order coefficient \( \xi_1(h, P_0) \) can be interpreted as the worst-(or best-) case sensitivity of the output among all \( P_f \) that is 1-dependent and generates the same marginal as \( P_0 \).

For higher order dependency, the analysis is more involved. One natural extension is to introduce more than one \( \eta \) as our “closeness” parameters, which tract the deviations of successive dependency orders. In this generalized setting, we will provide conservative bounds for the optimal values of (1) that are analogous to the right hand side of (4).

3 A Warm-Up: The Bivariate Case

To start with, we lay out our formulation (1) in the simple case when there are only two variables. This formulation will highlight the “ANOVA decomposition” feature of our sensitivity estimators as described in the introduction. The result here will also serve as a building block to facilitate the arguments in the next sections. For now, we have the cost function \( h : \mathcal{X}^2 \rightarrow \mathbb{R} \), and we shall assume that under the benchmark model \( P_0 \), the random variables \( X(\omega), Y(\omega) : \Omega \rightarrow \mathcal{X} \), are i.i.d.

We discuss more assumptions and notation. We assume that \( h \) is bounded and is non-degenerate, i.e. non-constant, under the benchmark model \( P_0 \). The latter assumption is equivalent to \( \text{Var}_{P_0}(h(X, Y)) > 0 \), where \( \text{Var}_{P_0}(\cdot) \) denotes the variance under \( P_0 \) (similarly, we shall use \( \text{sd}_{P_0}(\cdot) \) to denote the standard deviation under \( P_0 \)). For convenience, we abuse notation to denote \( P_f(x) := P_f(X \in x) \) as the marginal distribution of \( X \) under \( P_f \). The same notation is adopted for \( P_f(y), P_0(x) \) and \( P_0(y) \), and other similar instances in this paper. We always use upper case letters to denote random variables and lower case letters to denote deterministic variables. We will sometimes drop the dependence of \( (X, Y) \) on \( h \), so that we write \( E_0 h = E_0[h(X, Y)] \) and similarly for \( E_{f^*} h \). This sort of suppression of random variables as the parameters of functions such as \( h \), will be adopted repeatedly throughout the paper when no confusion arises. Moreover, we write \( E_0[h|x] = E_0[h(X, Y)|X = x] \) as the conditional expectation of \( h(X, Y) \) given \( X = x \), and similarly for \( E_0[h|y], E_{f^*}[h|x] \) and \( E_{f^*}[h|y] \) as well as other similar instances. We also write \( E_0[h|X] = E_0[h(X, Y)|X] \) interpreted as the random variable that represents the conditional expectation; similar representations for \( E_0[h|Y] \) and so forth.
Our method starts with the optimization programs

\[
\begin{align*}
\text{max} & \quad E_f[h(X,Y)] \quad \text{min} & \quad E_f[h(X,Y)] \\
\text{subject to} & \quad \phi^2(P_f(X,Y)) \leq \eta \quad \text{subject to} & \quad \phi^2(P_f(X,Y)) \leq \eta \\
& \quad P_f(x) = P_0(x) \text{ a.e.} \quad \text{and} & \quad P_f(x) = P_0(x) \text{ a.e.} \\
& \quad P_f(y) = P_0(y) \text{ a.e.} & \quad P_f(y) = P_0(y) \text{ a.e.} \\
& \quad P_f \in \mathcal{P}_0 & \quad P_f \in \mathcal{P}_0
\end{align*}
\]

(5)

(where a.e. stands for almost everywhere). These are rewrites of the generic formulation (1) under the bivariate setting, with the \( \eta \)-neighborhood defined in terms of the \( \chi^2 \)-distance between \( P_f(X,Y) \) and \( P_0(X,Y) \), which, as described in Section 2, is equivalent to \( \phi^2(P_f(X,Y)) \)). The following is a characterization of the optimal solutions of (5):

**Proposition 1.** Define \( r(x,y) := r(h)(x,y) = h(x,y) - E_0[h|x] - E_0[h|y] + E_0h \). Under the condition that \( h \) is bounded and \( \text{Var}_0(r(X,Y)) > 0 \), for any small enough \( \eta > 0 \), the optimal value of the max formulation in (5) satisfies

\[
\max E_f h = E_0 h + sd_0(r(X,Y)) \sqrt{\eta}
\]

where \( sd_0(\cdot) \) is the standard deviation under \( P_0 \). On the other hand, the optimal value of the min formulation in (5) is

\[
\min E_f h = E_0 h - sd_0(r(X,Y)) \sqrt{\eta}.
\]

Regarding \( X \) and \( Y \) as the “factors” in the evaluation of \( h(X,Y) \), \( r(X,Y) \) is the residual of \( h(X,Y) \) after removing the “main effect” of \( X \) and \( Y \). To explain this interpretation, consider the following decomposition

\[
h(X,Y) = E_0 h + (E_0[h|X] - E_0h) + (E_0[h|Y] - E_0h) + \epsilon
\]

where the residual error \( \epsilon \) is exactly \( r(X,Y) \). Therefore, the magnitude of the first order terms in (6) and (7) is the residual standard deviation of this decomposition. We label \( E_0[h|X] - E_0h \) and \( E_0[h|Y] - E_0h \) as the “main effects” of \( X \) and \( Y \), as opposed to the “interaction effect” controlled by \( \epsilon \). The reason is that when \( h(X,Y) \) is separable as \( h_1(X) + h_2(Y) \) for some functions \( h_1 \) and \( h_2 \), then one can easily see that \( \epsilon \) is exactly zero. In other words, \( \epsilon \) captures any nonlinear interaction between \( X \) and \( Y \). Notably, in the case that \( h(X,Y) \) is separable, dependency does not exert any effect on the performance measure.

There is another useful interpretation of \( r(X,Y) \) as the orthogonal projection of \( h(X,Y) \) onto the closed subspace \( \mathcal{M} = \{ V(X,Y) \in \mathcal{L}_2 : E_0[V|X] = 0, E_0[V|Y] = 0 \text{ a.s.} \} \subset \mathcal{L}_2 \), where \( \mathcal{L}_2 \) is
the $L_2$-space of random variables endowed with the inner product $\langle Z_1, Z_2 \rangle = E_0[Z_1Z_2]$. Intuitively speaking, the subspace $\mathcal{M}$ represents the set defined by the marginal constraints in (5). To see that $r(X,Y)$ is a projection on $\mathcal{M}$, first note that $r(X,Y)$ satisfies $E_0[r|X] = E_0[r|Y] = 0$ a.s.. Next, let $\bar{h} := \bar{h}(X,Y) := h - r = E_0[h|X] + E_0[h|Y] - E_0h$, and consider

$$E_0[\bar{h}r] = E_0[(h - r)r] = E_0[(E_0[h|X] + E_0[h|Y] - E_0h)(h - E_0[h|X] - E_0[h|Y] + E_0h)]$$

$$= E_0(E_0[h|X])^2 + E_0(E_0[h|Y])^2 - (E_0h)^2 - E_0(E_0[h|X] + E_0[h|Y] - E_0h)^2$$

$$= 0$$

where the second-to-last equality follows by conditioning on either $X$ or $Y$ in some of the terms. This shows that $h - r$ and $r$ are orthogonal. Therefore, $h$ can be decomposed into $r \in \mathcal{M}$ and $h - r \in \mathcal{M}^\perp$, which concludes that $r$ is the orthogonal projection on $\mathcal{M}$.

We note that the optimal values of (5) in this bivariate case, as depicted in Proposition 1, behave exactly linear in $\sqrt{\eta}$, although such is not the case for our general setup later on. Also, obviously, the optimal value is increasing in $\eta$ for the maximization formulation while decreasing for minimization.

In the rest of this section, we briefly explain the argument leading to Proposition 1. To avoid redundancy, let us focus on the maximization; the minimization counterpart can be tackled by merely replacing $h$ by $-h$. First, the formulation can be rewritten in terms of the likelihood ratio $L = L(X,Y) = dP_f(X,Y)/dP_0(X,Y) = dP_f(X,Y)/dP_f(X)dP_f(Y)$:

$$\begin{align*}
\max & \quad E_0[h(X,Y)L(X,Y)] \\
\text{subject to} & \quad E_0(L - 1)^2 \leq \eta \\
& \quad E_0[L|X] = 1 \text{ a.s.} \\
& \quad E_0[L|Y] = 1 \text{ a.s.} \\
& \quad L \geq 0 \text{ a.s.}
\end{align*}$$

(8)

where the decision variable is $L$, i.e. this optimization is over measurable functions. The constraints $E_0[L|X] = 1$ and $E_0[L|Y] = 1$ come from the marginal constraints in (8). To see this, note that $P_f(X \in A) = E_0[L(X,Y); X \in A] = E_0[E_0[L|X]I(X \in A)]$ for any measurable set $A$. Thus $P_f(X \in A) = P_0(X \in A)$ for any set $A$ is equivalent to $E_0[L|X] = 1$ a.s.. Similarly for $E_0[L|Y] = 1$. Moreover, observe also that $E_0[L|X] = E_0[L|Y] = 1$ implies $E_0[L] = 1$, and so $L$ in the formulation is automatically a valid likelihood ratio (and hence, there is no need to add this extra constraint).

The formulation (8) is in a tractable form. We consider its Lagrangian

$$\min_{\alpha \geq 0} \max_{L \in \mathcal{L}} E_0[h(X,Y)L] - \alpha(E_0(L - 1)^2 - \eta)$$

(9)

where $\mathcal{L} = \{L \geq 0 \text{ a.s.} : E_0[L|X] = E_0[L|Y] = 1 \text{ a.s.}\}$. The following is the key observation in solving (8):

9
Proposition 2. Under the condition that $h$ is bounded, for any large enough $\alpha > 0$,

$$L^*(x, y) = 1 + \frac{r(x, y)}{2\alpha}$$

maximizes $E_0[h(X, Y)L] - \alpha E_0(L - 1)^2$ over $L \in \mathcal{L}$.

The $L^*$ in (10) characterizes the optimal change of measure in solving the inner maximization of (9). It is easy to verify that $L^*$ is a valid likelihood ratio that lies in $\mathcal{L}$. Moreover, it is linear in $r(x, y)$, the residual function. The proof of Proposition 2 follows from a heuristic differentiation on the inner maximization of (9), viewing it as a Euler-Lagrange equation. The resulting candidate solution is then verified by using the projection property of $r(X, Y)$. Details are provided in Appendix A.

With the dual characterization above, one can prove Proposition 1 in two main steps: First, use complementarity condition $E_0(L^* - 1)^2 = \eta$ to write $\alpha^*$, the dual optimal solution, in terms of $\eta$. Second, express the optimal value $E_0[hL^*]$ in terms of $\alpha^*$, which, by step one, can be further translated into $\eta$. This will obtain precisely (6). See again Appendix A for details.

4 Sensitivity of 1-Dependence

In this section we focus on the formulation for assessing the effect of 1-dependency of the input. Consider now a performance measure $E[h(X_T)]$, where $h : \mathcal{X}^T \rightarrow \mathbb{R}$ and $X_T = (X_1, \ldots, X_T)$ are i.i.d. under $P_0$. As discussed before, we will sometimes write $h$ as a shorthand to $h(X_T)$.

4.1 Main Results and Equivalence of Formulations

The formulation (1) in this setting is

$$\max \text{ or } \min \quad E_f[h(X_T)]$$

subject to

$$\phi^2(P_f(X_{t-1}, X_t)) \leq \eta$$

$$\{X_t : t = 1, \ldots, T\} \text{ is a 1-dependent stationary process under } P_f$$

$$P_f(x_t) = P_0(x_t) \text{ a.e.}$$

$$P_f \in \mathcal{P}_0.$$ (11)

Recall that in the case of 1-dependence, the bivariate marginal distribution on two consecutive states completely characterizes $P_f$. Hence $\phi^2(P_f(X_{t-1}, X_t))$ is the same for all $t$. Similarly, $P_f(x_t) = P_0(x_t)$ also holds for all $t$.

We shall begin by introducing the concept of “symmetrization”. Define

$$H_2(x, y) = \sum_{t=2}^{T} E_0[h(X_T) | X_{t-1} = x, X_t = y] = (T - 1)E_0[h(X_T^{X_{t-1}=x,X_t=y})].$$ (12)
Here \( h(X_T^{(X_{U-1}=x,X_U=y)}) \) denotes the cost function evaluated by fixing \( X_{U-1} = x \) and \( X_U = y \), where \( U \) is a uniformly generated random variable over \( \{2, \ldots, T\} \), and all \( X_t \)'s other than \( X_{U-1} \) and \( X_U \) are randomly generated under \( P_0 \). The function \( H_2(x,y) \) is the sum of all conditional expectations, each conditioned at a different pair of consecutive time points. Alternately, it is \((T - 1)\) times the uniform time-averaging of all these conditional expectations.

Intuitively speaking, \( H_2(x,y) \) acts as a summary of the effect of a particular consecutive inputs on the system output. This symmetrization has a close connection to so-called influence function in robust statistics [16], which uses a similar function to measure the resistance of a statistic against data contamination. Similar definition also arises in [22]. Nevertheless, the conditioning in (12) on consecutive pairs is a distinct feature when assessment is carried out for 1-dependency; this definition and its subsequent implications appear to be new (both in sensitivity analysis and in robust statistics). In fact, for higher order dependency assessment, one has to look at higher order conditioning (see Section 5). This highlights the intimate relation between the order of conditioning and the order of dependency that one is interested in assessing.

Next we define two other “symmetrizations”:

\[
H_1^{-}(x) = \sum_{t=2}^{T} E_0[h(X_T)|X_{t-1} = x] = (T - 1)E_0[h(X_T^{(X_{U-1}=x)})] 
\]

and

\[
H_1^{+}(y) = \sum_{t=2}^{T} E_0[h(X_T)|X_t = y] = (T - 1)E_0[h(X_T^{(X_U=y)})] 
\]

where, again, \( h(X_T^{(X_{U-1}=x)}) \) and \( h(X_T^{(X_U=y)}) \) are the cost function evaluated by fixing \( X_{U-1} = x \) and \( X_U = y \) respectively for a uniform random variable \( U \) over \( \{2, \ldots, T\} \), with all other \( X_t \)'s randomly generated. These two symmetrizations are the expectations of \( H_2(\cdot, \cdot) \) conditioned on one of its arguments under \( P_0 \).

Moreover, we define a residual function similar to that introduced in Section 3, but derived on our symmetrization function \( H_2(x,y) \):

\[
R(x,y) = H_2(x,y) - H_1^{-}(x) - H_1^{+}(y) + (T - 1)E_0 h. 
\]

Note that \((T - 1)E_0 h\) is merely the expectation of \( H_2(X,Y) \), with \( X \) and \( Y \) being two i.i.d. copies of \( X_t \) under \( P_0 \). This residual function will serve as a key component in our analysis of the optimal value in (11). The intuition of (15) is similar to that of \( r(X,Y) \) in the bivariate case: here, after summarizing the input effect of the model using symmetrization, we filter out the “main effect” in the symmetrization to isolate the “interaction effect” due to 1-dependency. Moreover, the projection interpretation of \( R(X,Y) \) from \( H_2(X,Y) \) onto the subspace \( \mathcal{M} = \{V(X,Y) \in \mathcal{L}_2 : E_0[V|X] = E_0[V|Y] = 1 \ a.s.\} \) also follows exactly from the bivariate case.

With the above definitions, the main result in this section is the following expansions:
Theorem 1. Assume $h$ is bounded and $\text{Var}(R(X,Y)) > 0$. The optimal value of the max formulation in (11) possesses the expansion
\[
\max E_f h = E_0 h + s d_0(R(X,Y))\sqrt{\eta} + \sum_{s,t=2,\ldots,T} \frac{E_0[R(X_{s-1}, X_s)R(X_{t-1}, X_t)h(X_T)]}{\text{Var}_0(R(X,Y))} \eta + O(\eta^{3/2})
\] (16)
as $\eta \to 0$. Similarly, the optimal value of the min formulation in (11) possesses the expansion
\[
\min E_f h = E_0 h - s d_0(R(X,Y))\sqrt{\eta} + \sum_{s,t=2,\ldots,T} \frac{E_0[R(X_{s-1}, X_s)R(X_{t-1}, X_t)h(X_T)]}{\text{Var}_0(R(X,Y))} \eta + O(\eta^{3/2})
\] (17)
as $\eta \to 0$. Here $X$ and $Y$ are two i.i.d. copies of $X_t$ under $P_0$.

In contrast to Proposition 1, the optimal values here are no longer linear in $\sqrt{\eta}$. Curvature is present in shaping the rate of change as $P_f$ moves away from $P_0$. This curvature is identical for both the max and the min formulations, and it involves the third order cross-moment between $R$ and $h$.

Theorem 1 comes up naturally from an equivalent formulation in terms of likelihood ratio. Again let us focus on the maximization formulation. It can be shown that:

Proposition 3. The max formulation in (11) is equivalent to:
\[
\max \quad E_0 \left[ h(X_T) \prod_{t=2}^T L(X_{t-1}, X_t) \right]
\]
subject to
\[
E_0(L(X, Y) - 1)^2 \leq \eta
\]
\[
E_0[L|X] = 1 \ a.s.
\]
\[
E_0[L|Y] = 1 \ a.s.
\]
\[
L \geq 0 \ a.s.
\]

where $X$ and $Y$ denote two generic i.i.d. copies under $P_0$. The decision variable is $L(x,y)$.

The new feature in (18), compared to (8), is the product form of the likelihood ratios in the objective function. Let us briefly explain how this arises. Note first that a 1-dependent process satisfies the Markov property $P_f(X_t|X_{t-1}, X_{t-2}, \ldots) = P_f(X_t|X_{t-1})$ by definition. Hence
\[
E_f[h(X_T)] = \int h(x_T)dP_f(x_1)dP_f(x_2|x_1)\cdots dP_f(x_T|x_{T-1})
\]
\[
= \int h(x_T) \frac{dP_f(x_1)}{dP_0(x_1)} \frac{dP_f(x_2|x_1)}{dP_0(x_2)} \cdots \frac{dP_f(x_T|x_{T-1})}{dP_0(x_T)} dP_0(x_1)\cdots dP_0(x_T)
\] (19)
by introducing a change of measure from $P_f$ to $P_0$. Now, by identifying $L(x_{t-1}, x_t)$ as
\[
L(x_{t-1}, x_t) = \frac{dP_f(x_{t-1}|x_t)}{dP_0(x_t)} = \frac{dP_f(x_{t-1}, x_t)}{dP_0(x_{t-1})P_0(x_t)}
\]
and using the marginal constraint that $P_f(x_1) = P_0(x_1)$, we get that (19) is equal to

$$\int h(X_T) L(x_1, x_2) \cdots L(x_{T-1}, x_T) dP_0(x_1) \cdots dP_0(x_T) = E_0 \left[ h(X_T) \prod_{t=2}^{T} L(X_{t-1}, X_t) \right]$$

which is precisely the objective function in (18). The $L$ defined this way concurrently satisfies $E_0(L - 1)^2 \leq \eta$, corresponding to the constraint $\phi^2(P_f(X_{t-1}, X_t)) \leq \eta$ in (11). The equivalence of the other constraints between formulations (11) and (18) can be routinely checked; details are provided in Appendix B.

In contrast to (8), the product form in the objective function of (18) makes the posted optimization non-convex in general. This can be overcome, nevertheless, by an asymptotic analysis as $\eta \to 0$, in which enough regularity will be present to characterize the optimal solution. In the next subsection, we shall develop a fixed point machinery to carry out such analysis. Readers who are less indulged in the mathematical details can jump directly to Section 4.4.

4.2 Fixed Point Characterization of Optimality

This section outlines the mathematical development in translating the optimization (18) to sensitivity estimates. Similar to Section 3, we will first focus on the dual formulation. Because of the product form of likelihood ratios, finding a candidate optimal change of measure for the Lagrangian will now involve the use of a heuristic “product rule”. We will see that the manifestation of this “product rule” ultimately translates to the symmetrizations defined in Section 4.1. In order to show the existence and the optimality of this candidate optimal solution, one then has to set up a suitable contraction map whose fixed point coincides with it.

4.2.1 Finding a Candidate Optimal Solution for the Lagrangian

We start with a small technical observation in re-expressing (18). Since the first constraint in (18) implies that $E_0 L^2 \leq 1 + \eta$, we can focus on any $L$ that has a bounded $L_2$-norm (where $\| \cdot \|_2 = \sqrt{E_0[\cdot^2]}$). In other words, we can consider the following formulation that is equivalent to (18) when $\eta$ is small enough:

\[
\begin{align*}
\max & \quad E_0 \left[ h(X_T) \prod_{t=2}^{T} L(X_{t-1}, X_t) \right] \\
\text{subject to} & \quad E_0(L(X, Y) - 1)^2 \leq \eta \\
& \quad E_0[L|X] = 1 \text{ a.s.} \\
& \quad E_0[L|Y] = 1 \text{ a.s.} \\
& \quad L \geq 0 \text{ a.s.} \\
& \quad E_0 L^2 \leq M
\end{align*}
\]
for some constant $M > 1$. The additional constraint $E_0L^2 ≤ M$ bounds the norm of $L$ in order to facilitate the construction of an associated contraction map. Similar to the bivariate case, we shall consider the Lagrangian

$$\min_{\alpha ≥ 0} \max_{L ∈ \mathcal{L}^+(M)} E_0[hL] - \alpha(E_0(L - 1)^2 - \eta)$$

(21)

where for convenience we denote $\mathcal{L}^+(M) = \{L ≥ 0 \ a.s.: E_0[L|X] = E_0[L|Y] = 1 \ a.s., \ E_0L^2 ≤ M\}$, and $L = \prod_{t=2}^T L(X_{t-1}, X_t)$. We will try solving the inner maximization. Relaxing the constraints $E_0[L|X] = E_0[L|Y] = 1 \ a.s.$, we have

$$\max_{L ∈ \mathcal{L}^+(M)} E_0[hL] - \alpha(E_0(L - 1)^2 - \eta) + \int (E_0[L|x] - 1)d\beta(x) + \int (E_0[L|y] - 1)d\gamma(y)$$

(22)

where $\mathcal{L}^+(M) = \{L ≥ 0 \ a.s.: E_0L^2 ≤ M\}$, and $\beta$ and $\gamma$ are signed measures with bounded variation. For any $\alpha > 0$, treat $E_0[hL] - \alpha(E_0(L - 1)^2 - \eta) + \int (E_0[L|x] - 1)d\beta(x) + \int (E_0[L|y] - 1)d\gamma(y)$ as a heuristic Euler-Lagrange equation, and differentiate with respect to $L(x,y)$. We get

$$\frac{d}{dL} \left( E_0[hL] - \alpha(E_0(L - 1)^2 - \eta) + \int (E_0[L|x] - 1)d\beta(x) + \int (E_0[L|y] - 1)d\gamma(y) \right)$$

$$= H^L(x,y) - 2\alpha L(x,y) + \beta(x) + \gamma(y)$$

(23)

where $H^L(x,y)$ is defined as

$$H^L(x,y) = \sum_{t=2}^T E_0[hL_{2:T}|X_{t-1} = x, X_t = y]$$

(24)

and $L_{2:T} = \prod_{t=2}^T L(X_{s-1}, X_s)$ is the leave-one-out product of likelihood ratios. Although this “product rule” is heuristic, it can be checked, in the case $P_0$ is discrete, that combinatorially $(d/dL(x,y))E_0[hL]$ exactly matches (24). Setting (23) to zero, we have our candidate optimal solution

$$L(x,y) = \frac{1}{2\alpha}(H^L(x,y) + \beta(x) + \gamma(y)).$$

Much like the proof of Proposition 2 (in Appendix A), it can be shown that under the condition $E_0[L|X] = E_0[L|Y] = E_0L = 1 \ a.s.$, we can rewrite $\beta(x)$ and $\gamma(y)$ to get

$$L(x,y) = 1 + \frac{R^L(x,y)}{2\alpha}$$

(25)

where

$$R^L(x,y) = H^L(x,y) - E_0[H^L|x] - E_0[H^L|y] + E_0H^L$$

(26)

is the residual function derived from $H^L$. Here $E_0[H^L|x] = E_0[H^L(X,Y)|X=x]$ and $E_0[H^L|y] = E_0[H^L(X,Y)|Y=y]$ are the expectations of $H^L(X,Y)$ under $P_0$ conditioned on each of the i.i.d. $X$ and $Y$. Note that (25) is a fixed point equation in $L$ itself. The $L$ that satisfies (25) constitutes our candidate optimal solution for the inner maximization of (21). The challenge is to verify the existence and the optimality of such an $L$, or in other words, to show that
Proposition 4. For any large enough $\alpha > 0$, the $L^*$ that satisfies (25) is an optimal solution of
\[
\max_{L \in \mathcal{L}^c(M)} E_0[hL] - \alpha E_0(L - 1)^2.
\]  
(27)

4.2.2 Construction of Contraction Operator

The goal of this subsection is to outline the arguments for proving Proposition 4. The main technical development is to show that the contraction operator that defines the fixed point in (25) possesses an ascendency property over the objective of (27), so that the fixed point indeed coincides with the optimum.

We construct this contraction operator, called $\mathcal{K} : \mathcal{L}^c(M)^{T-2} \rightarrow \mathcal{L}^c(M)^{T-2}$, in a few steps. First, let us define a generalized notion of $H^L$ to cover the case where the factors in $L$ are not necessarily the same. More concretely, for any $L^{(1):T-2} := (L^{(1)}, L^{(2)}, \ldots, L^{(T-2)}) \in \mathcal{L}^c(M)^{T-2}$, let
\[
H^{L^{(1):T-2}}(x, y)
\]
\[
= E_0[hL^{(1)}(X_2, X_3)L^{(2)}(X_3, X_4)L^{(3)}(X_4, X_5) \cdots L^{(T-2)}(X_{T-1}, X_T)|X_1 = x, X_2 = y]
+ E_0[hL^{(T-2)}(X_1, X_2)L^{(1)}(X_3, X_4)L^{(2)}(X_4, X_5) \cdots L^{(T-3)}(X_{T-1}, X_T)|X_2 = x, X_3 = y]
+ E_0[hL^{(T-3)}(X_1, X_2)L^{(T-2)}(X_2, X_3)L^{(1)}(X_4, X_5) \cdots L^{(T-4)}(X_{T-1}, X_T)|X_3 = x, X_4 = y]
\]
\[
\ldots
\]
\[
+ E_0[hL^{(2)}(X_1, X_2)L^{(3)}(X_2, X_3) \cdots L^{(T-2)}(X_{T-3}, X_{T-2})L^{(1)}(X_{T-1}, X_T)|X_{T-2} = x, X_{T-1} = y]
+ E_0[hL^{(1)}(X_1, X_2)L^{(2)}(X_2, X_3) \cdots L^{(T-3)}(X_{T-3}, X_{T-2})L^{(T-2)}(X_{T-2}, X_{T-1})|X_{T-1} = x, X_T = y].
\]

In other words, $H^{L^{(1):T-2}}(x, y)$ is the sum of conditional expectations, with each summand conditioned on a different pair of $(X_{t-1}, X_t)$, and the likelihood ratio in each conditional expectation is a rolling of the sequence $L^{(1)}, L^{(2)}, \ldots$ acted on $(X_t, X_{t+1}), (X_{t+1}, X_{t+2}), \ldots$ in a roundtable manner. Note that when $L^{(k)}$’s are all identical, then $H^{L^{(1):T-2}}$ reduces to $H^L$ defined in (24) before.

With the definition above, we define a stepwise operator $K : \mathcal{L}^c(M)^{T-2} \rightarrow \mathcal{L}^c(M)$ as
\[
K(L^{(1):T-2})(x, y) = 1 + \frac{R^{L^{(1):T-2}}(x, y)}{2\alpha}
\]
(28)
where $R^{L^{(1):T-2}} = H^{L^{(1):T-2}}(x, y) - E_0[H^{L^{(1):T-2}}(x)|x] - E_0[H^{L^{(1):T-2}}(y)|y] + E_0[H^{L^{(1):T-2}}]$ is the residual of $H^{L^{(1):T-2}}(x, y)$; same as (26), we denote $E_0[H^{L^{(1):T-2}}(x)|x] = E_0[H^{L^{(1):T-2}}(X, Y)|X = x]$ and $E_0[H^{L^{(1):T-2}}(y)|y] = E_0[H^{L^{(1):T-2}}(X, Y)|Y = y]$. 

15
The operator $\mathcal{K}$ is now defined as follows. Given $L^{(1):(T-2)} = (L^{(1)}, L^{(2)}, \ldots, L^{(T-2)}) \in \mathcal{L}^c(M)^{T-2}$, define

$$
\begin{align*}
\tilde{L}^{(1)} &= K(L^{(1)}, L^{(2)}, \ldots, L^{(T-2)}) \\
\tilde{L}^{(2)} &= K(\tilde{L}^{(1)}, L^{(2)}, \ldots, L^{(T-2)}) \\
\tilde{L}^{(3)} &= K(\tilde{L}^{(1)}, \tilde{L}^{(2)}, L^{(3)}, \ldots, L^{(T-2)}) \\
& \vdots \\
\tilde{L}^{(T-2)} &= K(\tilde{L}^{(1)}, \tilde{L}^{(2)}, \ldots, \tilde{L}^{(T-1)}, L^{(T-2)}).
\end{align*}
$$

Then $\mathcal{K}(L^{(1):(T-2)}) = (\tilde{L}^{(1)}, \tilde{L}^{(2)}, \ldots, \tilde{L}^{(T-2)})$.

The $\mathcal{K}$ constructed above is a well-defined contraction operator:

**Lemma 1** (Contraction Property). For large enough $\alpha$, the operator $\mathcal{K}$ is a well-defined, closed contraction map on $\mathcal{L}^c(M)^{T-2}$, with the metric $d(L, L') = \max_{k=1,\ldots,T-2} \|L^{(k)} - L'^{(k)}\|_2$, where $L = (L^{(k)})_{k=1,\ldots,T-2}$ and $L' = (L'^{(k)})_{k=1,\ldots,T-2}$, and $\| \cdot \|_2$ is the $L_2$-norm under $P_0$. As a result $\mathcal{K}$ possesses a unique fixed point $L^* \in \mathcal{L}^c(M)^{T-2}$. Moreover, all $T-2$ components of $L^*$ are identical.

The proof of this lemma requires tedious but routine validation of all the conditions, using the martingale property of the sequential likelihood ratios defined by the products of $L^{(k)}$'s. The proof is left to Appendix B.

Next we investigate an important property of $K$, the stepwise map that defines $\mathcal{K}$: that by applying it on any $L^{(1):(T-2)} \in \mathcal{L}^c(M)^{T-2}$, the objective in (27) is always non-decreasing. To state this statement precisely, we need to define an unconditional form of $H_L^{(1):(T-2)}(x, y)$. For any $L^{(1):(T-1)} = (L^{(1)}, \ldots, L^{(T-1)}) \in \mathcal{L}^c(M)^{T-1}$, the real number $\tilde{H}_L^{(1):(T-1)} \in \mathbb{R}$ is defined as

$$
\tilde{H}_L^{(1):(T-1)} := E_0[hL^{(1)}(X_1, X_2) L^{(2)}(X_2, X_3) L^{(3)}(X_3, X_4) \cdots L^{(T-1)}(X_{T-1}, X_T)] \\
+ E_0[hL^{(T-1)}(X_1, X_2) L^{(1)}(X_2, X_3) L^{(2)}(X_3, X_4) \cdots L^{(T-2)}(X_{T-1}, X_T)] \\
+ E_0[hL^{(T-2)}(X_1, X_2) L^{(T-1)}(X_2, X_3) L^{(1)}(X_3, X_4) L^{(2)}(X_4, X_5) \cdots L^{(T-3)}(X_{T-1}, X_T)] \\
+ \cdots \\
+ E_0[hL^{(2)}(X_1, X_2) L^{(3)}(X_2, X_3) L^{(4)}(X_3, X_4) \cdots L^{(T-1)}(X_{T-2}, X_{T-1}) L^{(1)}(X_{T-1}, X_T)].
$$

Note that $\tilde{H}_L^{(1):(T-1)}$ is invariant to a shift of the indices in $L^{(1)}, \ldots, L^{(T-1)}$ (in a roundtable manner). This implies that we can express, for example, $\tilde{H}_L^{(1):(T-1)} = E_0[H_L^{(1):(T-2)}(X, Y) L^{(T-1)}(X, Y)] = E_0[H_L^{(2):(T-1)}(X, Y) L^{(1)}(X, Y)]$.
Moreover, it is more convenient to consider a \((T - 1)\)-scaling of the Lagrangian in (27), i.e.
\[
(T - 1) \times \left( E_0 \left[ h \prod_{t=2}^{T} L(X_{t-1}, X_t) \right] - \alpha E_0(L - 1)^2 \right) = \bar{H}^L - \alpha(T - 1)E_0(L - 1)^2 \tag{31}
\]
where \(\bar{H}^L\) is defined by plugging \(L = L^{(1)} = \cdots = L^{(T-1)}\) into (30). Now, we shall consider a generalized version of the objective in (31):
\[
\bar{H}^{(1):(T-1)} - \alpha \sum_{t=1}^{T-1} E_0(L^{(t)} - 1)^2
\]
as a function of \(L^{(1):(T-1)}\). Our monotonicity is on this generalized objective:

**Lemma 2 (Monotonicity).** Starting from any \(L^{(1)}, \ldots, L^{(T-2)} \in \mathcal{L}^c(M)\), consider the sequence \(L^{(k)} = K(L^{(k-T+2)}, L^{(k-T+3)}, \ldots, L^{(k-1)})\) for \(k = T - 1, T - 2, \ldots\), where \(K\) is defined in (28). The quantity
\[
\bar{H}^{(k):(k+T-2)} - \alpha \sum_{t=2}^{T-1} E_0(L^{(k+t-1)} - 1)^2
\]
is non-decreasing in \(k \geq 1\).

The following proof highlights the crucial role that symmetrization plays to guarantee ascendency, by reducing the problem into the bivariate situation studied in Section 3.

**Proof of Lemma 2.** Consider
\[
\bar{H}^{(k):(k+T-2)} - \alpha \sum_{t=1}^{T-1} E_0(L^{(k+t-1)} - 1)^2
\]

\[
= E_0[H^{L^{(k+1):(k+T-2)}(X,Y)}L^{(k)}(X,Y)] - \alpha E_0(L^{(k)} - 1)^2 - \alpha \sum_{t=2}^{T-1} E_0(L^{(k+t-1)} - 1)^2
\]

by the symmetric construction of \(H^{L^{(k):(k+T-2)}}\)
\[
\leq E_0[H^{L^{(k+1):(k+T-2)}(X,Y)L^{(k+T-1)}(X,Y)}] - \alpha E_0(L^{(k+T-1)} - 1)^2 - \alpha \sum_{t=2}^{T-1} E_0(L^{(k+t-1)} - 1)^2
\]

by using Proposition 2, treating the cost function as \(H^{L^{(k+1):(k+T-2)}}(x, y)\),

and recalling the definition of \(K\) in (28)
\[
= \bar{H}^{(k+1):(k+T-1)} - \alpha \sum_{t=1}^{T-1} E_0(L^{(k+t)} - 1)^2
\]

by the symmetric construction of \(H^{L^{(k+1):(k+T-1)}}\).

This concludes the ascendency of \(\bar{H}^{L^{(k):(k+T-2)}} - \alpha \sum_{t=1}^{T-1} E_0(L^{(k+t-1)} - 1)^2\). □
The final step is to conclude the convergence of the scaled objective value in (31) along the dynamic sequence defined by the iteration of $K$ to that evaluated under the fixed point. This can be seen by first noting that convergence to the fixed point associated with the operator $K$ immediately implies componentwise convergence, i.e. the sequence $L^{(k)} = K(L^{(k-T+2)}, L^{(k-T+3)}, \ldots, L^{(k-1)})$ for $k = T-1, T-2, \ldots$ defined in Lemma 2 converges to $L^*$, the identical component of the fixed point $L^*$ of $K$. Then, by invoking standard convergence tools (see Appendix B), we have the following:

**Lemma 3** (Convergence). Define the same sequence $L^{(k)}$ as in Lemma 2. We have

$$\tilde{H}L^{(k):\{k+T-2\}} - \alpha \sum_{t=1}^{T-1} E_0(L^{(k+t-1)} - 1)^2 \to \tilde{H}L^* - \alpha (T-1)E_0(L^*-1)^2$$

(32)

as $k \to \infty$.

With these lemmas in hand, the proof of Proposition 4 is immediate:

**Proof of Proposition 4.** For any $L \in \mathcal{L}^c(M)$, denote $L^{(1)} = L^{(2)} = \cdots = L^{(T-2)} = L$ and define the sequence $L^{(k)}$ for $k \geq T-1$ as in Lemma 2. By Lemmas 2 and 3 we conclude Proposition 4. \qed

### 4.3 Asymptotic Expansion

As we have characterized the dual objective value (27), we turn our attention to connecting the relation between the primal optimal value $E_f h$ and $\eta$. We shall use the following theorem from [25] (see also [22]) that relates the dual solution to the primal without assumptions on convexity:

**Theorem 2** (Adapted from [25], Chapter 8 Theorem 1). Suppose one can find $\alpha^* > 0$ and $L^* \in \mathcal{L}^c(M)$ such that

$$E_0[hL] - \alpha^*E_0(L - 1)^2 \leq E_0[hL^*] - \alpha^*E_0(L^* - 1)^2$$

(33)

for any $L \in \mathcal{L}^c(M)$. Then $L^*$ solves

$$\max E_0[hL]$$

subject to

$$E_0(L - 1)^2 \leq E_0(L^* - 1)^2$$

$L \in \mathcal{L}^c(M)$.

Using Theorem 2, we can show Theorem 1 in a few steps. First, observe that from Proposition 4 we have obtained $L^*$, in terms of $\alpha^*$, that satisfies (33) for any large enough $\alpha^*$. In view of Theorem 2, we shall set $\eta = E_0(L^* - 1)^2$, the complementarity condition, and show that for any small $\eta$ we can find a large enough $\alpha^*$ and its corresponding $L^*$ from Proposition 4 that satisfy this condition. This would allow us to invoke Theorem 2 to characterize the optimal solution of (20) when $\eta$ is small. After that, we shall invert the relation of $\alpha^*$ in terms of $\eta$ using $\eta = E_0(L^* - 1)^2$. We then express the optimal objective value $E_0[hL^*]$ in terms of $\alpha^*$ and in turn $\eta$. This will lead to Theorem 1. We provide the details in Appendix A.
4.4 Some Extensions

We close this section by discussing two directions of extensions.

Extension to non-i.i.d. benchmark. When the benchmark is non-i.i.d. yet is 1-dependent, we can formulate optimizations similar to (11) that can assess the sensitivity as dependency moves away from the benchmark within the 1-dependence structure. In terms of formulation, the first constraint in (11) can no longer be expressed in terms of $\phi^2$-coefficient, but rather has to be kept in the $\chi^2$-form $\chi^2(P_f(X_{t-1}, X_t), P_0(X_{t-1}, X_t)) \leq \eta$. Correspondingly, the likelihood ratio in the equivalent formulation (18) is

$$L(X_1, X_2) = \frac{dP_f(X_1, X_2)}{dP_0(X_1, X_2)} = \frac{dP_f(X_2|X_1)}{dP_0(X_2|X_1)}.$$ 

We shall discuss the analog of Theorem 1 under this 1-dependent benchmark formulation. The theorem holds with the coefficients in the expansions (16) and (17) evaluated under the corresponding 1-dependent $P_0$, whereby $X$ and $Y$ in the theorem form a consecutive pair of states under $P_0$. This is, however, subject to one substantial change in the computation of the quantity $R(x, y)$. More specifically, $R(X, Y)$ is still the projection of the symmetrization function $H_2(X, Y)$, defined in (12) and under the 1-dependent benchmark measure $P_0$, onto the subspace $\mathcal{M} = \{V(X, Y) \in L_2 : E_0[V|X] = 0, E_0[V|Y] = 0 \text{ a.s.}\}$. However, $R(X, Y)$ no longer satisfies (15), i.e. it is not the residual of the associated ANOVA type decomposition after removing the “main effect” of $X$ and $Y$. The reason is that $X$ and $Y$, representing a consecutive pair of states under $P_0$, are not independent. This then leads to the phenomenon that the projection on $\mathcal{M}$ deviates from the “interaction effect” under the ANOVA decomposition.

We note that Propositions 1 and 4, together with all the supporting lemmas and hence Theorem 1 hold as long as $R(X, Y)$ is the projection of $H_2(X, Y)$ onto $\mathcal{M}$ (and $r(X, Y)$ is the projection of $h(X, Y)$ onto $\mathcal{M}$), since their proofs in essence only rely on this property. The computation of this projection, however, is not straightforward once the ANOVA interpretation is lost. To get some hints, note that $\mathcal{M} = \mathcal{M}^X \cap \mathcal{M}^Y$, where $\mathcal{M}^X = \{V(X, Y) \in L_2 : E_0[V|X] = 0 \text{ a.s.}\}$ and $\mathcal{M}^Y = \{V(X, Y) \in L_2 : E_0[V|Y] = 0 \text{ a.s.}\}$. Consequently, for a projection on the intersection of two subsets, Anderson-Duffin formula (see, for example, [9]) implies that

$$R(X, Y) = 2P^X(P^X + P^Y)^\dagger P^Y H_2(X, Y)$$

where $P^X$ is the projection onto $M^X$, $P^Y$ is the projection onto $M^Y$, and $\dagger$ denotes the Moore-Penrose pseudoinverse. We do not know yet if the computation of (34) is feasible; its investigation will be left for future work.

Auxiliary random sources. Our results still hold with minor modification when there are auxiliary random sources in the model other than $X_T$. Suppose that the performance measure is $E[h(X_T, Y)]$, 

19
where $Y$ is another random object that is potentially dependent on $X_T$, and our goal is to assess the effect of serial dependency of $X_T$ on the performance measure. All the results in this section hold trivially with $h(X_T)$ replaced by $E[h(X_T, Y)|X_T]$, where the expectation is with respect to only $Y$, which is fixed in our assessment.

5 General Higher Order Dependency Assessment

We now generalize our result to higher-order dependency assessment. Here we choose to adopt the formulation in which the number of “closeness” parameters equals the dependency order, for two reasons. The first is analytical tractability: we will see that there is a natural extension of the results in Section 4, by using the building blocks that we have already developed. Secondly, from a practical viewpoint, the involved statistical distances in our constraint will be in a form that is estimable from data.

To highlight our idea, consider a special case: a benchmark $P_0$ that generates i.i.d. $X_t$’s, and the assessment is on 2-dependency. Note that for any 2-dependent $P_f$, the third order marginal distribution $P_f(X_{t-2}, X_{t-2}, X_t)$ completely determines $P_f$. Our formulation is then

$$\max / \min E_f[h(X_T)]$$
subject to
$$\phi^2(P_f(X_{t-1}, X_t)) \leq \eta_1$$
$$\phi^2_2(P_f(X_{t-2}, X_{t-1}, X_t)) \leq \eta_2$$
$$\{X_t : t = 1, 2, \ldots, T\} \text{ is a 2-dependent stationary process under } P_f$$
$$P_f(x_t) = P_0(x_t) \text{ a.e.}$$
$$P_f \in P_0.$$  \(35\)

We shall state the meaning of $\phi^2$ and $\phi^2_2$ above more precisely. In particular, we attempt to extend the definition of Pearson’s $\phi^2$-coefficient to multiple variables in a natural manner. For this purpose, we first introduce some notation. Given any $P_f$ that generates a $p$-dependent sequence of $X_t$’s, denote $P_{f_k}$ as the measure that generates a $k$-dependent sequence, where $k \leq p$, such that $P_{f_k}(X_{t-k}, X_{t-k+1}, \ldots, X_t) = P_f(X_{t-k}, X_{t-k+1}, \ldots, X_t)$, i.e. the marginal distribution for any
$k$ consecutive states is unchanged. Then

$$
\phi^2(P_f(X_{t-1}, X_t)) = E_{\tilde{f}_0} \left( \frac{dP_f(X_t|X_{t-1})}{dP_{\tilde{f}_0}(X_t|X_{t-1})} - 1 \right)^2 = E_0 \left( \frac{dP_f(X_t|X_{t-1})}{dP_{\tilde{f}_0}(X_t|X_{t-1})} - 1 \right)^2
$$

$$
= E_{\tilde{f}_0} \left( \frac{dP_f(X_{t-1}, X_t)}{dP_{\tilde{f}_0}(X_{t-1}, X_t)} - 1 \right)^2 = E_0 \left( \frac{dP_f(X_{t-1}, X_t)}{dP_{\tilde{f}_0}(X_{t-1}, X_t)} - 1 \right)^2
$$

$$
= E_{\tilde{f}_0} \left( \frac{dP_f(X_{t-1}, X_t)}{dP_{\tilde{f}_0}(X_{t-1})P_{\tilde{f}_0}(X_t)} - 1 \right)^2 = E_0 \left( \frac{dP_f(X_{t-1}, X_t)}{dP_{\tilde{f}_0}(X_{t-1})P_{\tilde{f}_0}(X_t)} - 1 \right)^2
$$

is exactly Pearson’s $\phi^2$-coefficient, with three equivalent expressions and with the fact that $P_{\tilde{f}_0} = P_0$.

In a similar fashion, we define $\phi_2^2$ as

$$
\phi_2^2(P_f(X_{t-2}, X_{t-1}, X_t)) = E_{\tilde{f}_1} \left( \frac{dP_f(X_t|X_{t-2}, X_{t-1})}{dP_{\tilde{f}_1}(X_t|X_{t-2}, X_{t-1})} - 1 \right)^2
$$

$$
= E_{\tilde{f}_1} \left( \frac{dP_f(X_{t-2}, X_{t-1}, X_t)}{dP_{\tilde{f}_1}(X_{t-2}, X_{t-1}, X_t)} - 1 \right)^2
$$

$$
= E_{\tilde{f}_1} \left( \frac{dP_f(X_{t-2}, X_{t-1}, X_t)P_f(X_{t-1})}{dP_{\tilde{f}_1}(X_{t-2}, X_{t-1})P_f(X_{t-1}, X_t)} - 1 \right)^2.
$$

The equality among (36), (37) and (38) can be seen easily. Observe that by construction $P_{\tilde{f}_1}(X_{t-2}, X_{t-1}) = P_f(X_{t-2}, X_{t-1})$, and so

$$
\frac{dP_f(X_{t-2}, X_{t-1})}{dP_{\tilde{f}_1}(X_{t-2}, X_{t-1})} = \frac{dP_f(X_{t-2}, X_{t-1}, X_t)}{dP_{\tilde{f}_1}(X_{t-1}, X_t)} P_f(X_{t-2}, X_{t-1}, X_t)
$$

$$
\frac{dP_f(X_{t-2}, X_{t-1}, X_t)}{dP_{\tilde{f}_1}(X_{t-2}, X_{t-1})} = \frac{dP_f(X_{t-2}, X_{t-1}, X_t)}{dP_{\tilde{f}_1}(X_{t-2}, X_{t-1})}.
$$

This shows the equivalence between (36) and (37). Next, note that $P_{\tilde{f}_1}(X_t|X_{t-2}, X_{t-1}) = P_{\tilde{f}_1}(X_t|X_{t-1}) = P_f(X_t|X_{t-1})$, the first equality coming from 1-dependence property and the second equality coming from the equality between $P_{\tilde{f}_1}$ and $P_f$ for marginal distributions up to the second order. Therefore,

$$
E_{\tilde{f}_1} \left( \frac{dP_f(X_{t-2}, X_{t-1})}{dP_{\tilde{f}_1}(X_{t-2}, X_{t-1})} - 1 \right)^2 = E_{\tilde{f}_1} \left( \frac{dP_f(X_{t-2}, X_{t-1})}{dP_f(X_{t-1})} - 1 \right)^2
$$

$$
= E_{\tilde{f}_1} \left( \frac{dP_f(X_{t-2}, X_{t-1}, X_t)P_f(X_{t-1})}{dP_{\tilde{f}_1}(X_{t-2}, X_{t-1})P_f(X_{t-1}, X_t)} - 1 \right)^2
$$

which shows the equivalence between (36) and (38).

To summarize, $\phi_2^2(P_f(X_{t-2}, X_{t-1}, X_t))$ is the $\chi^2$-distance between the conditional probabilities $P_f(X_t|X_{t-2}, X_{t-1})$ and $P_{\tilde{f}_1}(X_t|X_{t-2}, X_{t-1})$, or equivalently the distance between the marginal
probabilities \( P_f(X_{t-2}, X_{t-1}, X_t) \) and \( P_{\tilde{f}}(X_{t-2}, X_{t-1}, X_t) \). Lastly, one can alternately view \( \phi^2_2 \) using the augmented state \( (X_{t-1}, X_t) \in \mathcal{X}^2 \):

\[
\phi^2_2(P_f(X_{t-2}, X_{t-1}, X_t)) = E_{\tilde{f}} \left( \frac{dP_f((X_{t-1}, X_t)|(X_{t-2}, X_{t-1}))}{dP_{\tilde{f}}((X_{t-1}, X_t)|(X_{t-2}, X_{t-1}))} - 1 \right)^2. \tag{39}
\]

The augmented state representation is especially useful in extending the results from Section 4.

**Example 1.** To get some intuition for the relation between \( P_{\tilde{f}} \) and \( P_f \), consider the Gaussian case. Suppose \( (X_{t-2}, X_{t-1}, X_t) \) is multivariate normal with mean 0 and covariance matrix \( \Sigma \) under \( P_f \), where

\[
\Sigma = \begin{bmatrix}
1 & \rho_1 & \rho_2 \\
\rho_1 & 1 & \rho_1 \\
\rho_2 & \rho_1 & 1
\end{bmatrix}.
\]

Then under \( P_{\tilde{f}} \), \( (X_{t-2}, X_{t-1}, X_t) \) will still retain a zero-mean multivariate normal distribution, but with covariance matrix \( \tilde{\Sigma} \) being \( \Sigma \) with the \( \rho_2 \)'s eliminated, namely

\[
\tilde{\Sigma} = \begin{bmatrix}
1 & \rho_1 & 0 \\
\rho_1 & 1 & \rho_1 \\
0 & \rho_1 & 1
\end{bmatrix}.
\]

With the above definitions and the formulation (35), we have the following approximation:

**Theorem 3.** The optimal value of the max formulation in (35) is bounded above by

\[
\max E_f h \leq E_0 h + s_d(R(X,Y))\sqrt{\eta_1} + s_d(S(X,Y,Z))\sqrt{\eta_2} + o(\sqrt{\eta_1} + \sqrt{\eta_2}) \tag{40}
\]

and the min formulation is bounded below by

\[
\min E_f h \geq E_0 h - s_d(R(X,Y))\sqrt{\eta_1} - s_d(S(X,Y,Z))\sqrt{\eta_2} + o(\sqrt{\eta_1} + \sqrt{\eta_2}) \tag{41}
\]

where \( R \) is defined as in (15), i.e.

\[
R(x, y) = \sum_{t=2}^{T} E_0[h|X_{t-1} = x, X_t = y] - \sum_{t=2}^{T} E_0[h|X_{t-1} = x] - \sum_{t=2}^{T} E_0[h|X_t = y] + (T - 1)E_0 h
\]

and \( S \) is defined as the residual of symmetrization involving third order conditioning, namely

\[
S(x, y, z) = \sum_{t=3}^{T} E_0[h|X_{t-2} = x, X_{t-1} = y, X_t = z] - \sum_{t=3}^{T} E_0[h|X_{t-2} = x, X_{t-1} = y] - \sum_{t=3}^{T} E_0[h|X_{t-1} = y, X_t = z] + \sum_{t=3}^{T} E_0[h|X_{t-1} = y]. \tag{42}
\]
We shall explain how the bounds (40) and (41) are obtained, and the relaxation that we impose that leads to conservative estimates instead of tight asymptotic equality as in Section 4. As before, we focus on the maximization formulation. Suppose first that (35) possesses an optimal value and an optimal solution (if not, then a proper \( \epsilon \)-argument will work), and let \( Z^* \) be its optimal value.

Next, introduce the formulation

\[
\text{max} \quad E_f[h(X_T)] \\
\text{subject to} \quad E_0 \left( \frac{dP_f(X_{t-1}, X_t)}{dP_0(X_{t-1}, X_t)} - 1 \right)^2 \leq \eta_1 \\
\{X_t : t = 1, \ldots, T\} \text{ is a 1-dependent stationary process under } P_f \\
P_f(x_t) = P_0(x_t) \text{ a.e.} \\
P_f \in \mathcal{P}_0,
\]

i.e. the formulation (11) with \( \eta \) replaced by \( \eta_1 \). Note that (43) is a subproblem to (35), in the sense that any feasible solution of (43) is feasible in (35). This is because \( \phi_2^2(P_f(X_{t-2}, X_{t-1}, X_t)) = 0 \) for any 1-dependent measure \( P_f \). Now let \( Z' \) be the optimal value of (43), and consider the decomposition

\[
Z^* - E_0 h = (Z^* - Z') + (Z' - E_0 h) \tag{44}
\]

We know that \( Z' - E_0 h \leq sd_0(R(X,Y))\sqrt{\eta_1} + O(\eta_1) \) by Theorem 1. On the other hand, we argue that \( Z^* - Z' \leq Z^* - E_{\tilde{f}_1} \) where \( P_{\tilde{f}_1} \) is the measure that corresponds to the 1-dependent counterpart of \( P_{f^*} \), the optimal measure of (35). Note that \( P_{f^*} \) certainly satisfies all constraints in (43), inherited from the properties of \( P_{f^*} \). We then must have \( Z' \geq E_{\tilde{f}_1} h \) by the definition of \( Z' \) and the feasibility of \( P_{\tilde{f}_1} \) in (43), and so \( Z^* - Z' \leq Z^* - E_{\tilde{f}_1} h \).

Next, note that \( Z^* \) coincides with the optimal value of

\[
\text{max} \quad E_f[h(X_T)] \\
\text{subject to} \quad E_{\tilde{f}_1} \left( \frac{dP_f(X_{t-2}, X_{t-1}, X_t)}{dP_{\tilde{f}_1}(X_{t-2}, X_{t-1}, X_t)} - 1 \right)^2 \leq \eta_2 \\
\{X_t : t = 1, \ldots, T\} \text{ is a 2-dependent stationary process under } P_f \\
P_f(x_{t-1}, x_t) = P_{f^*}(x_{t-1}, x_t) \\
P_f \in \mathcal{P}_0.
\]

To argue this, note that any feasible solution \( P_f \) in (45) must be feasible in (43), since \( P_f \) has the same bivariate marginal as \( P_{\tilde{f}_1} \); this in turn implies feasibility in (35). So the optimal value of (45) is at most \( Z^* \). On the other hand, \( P_{f^*} \) is clearly a feasible solution of (45), and hence the optimal value of (45) is at least \( Z^* \). This leads to the coincidence of \( Z^* \) with the optimal value of (45).
Next, we note that (45) is equivalent to
\[
\max E_{f_i}[h(X_T)]
\]
subject to
\[
E_{f_i} \left( \frac{dP_t((X_{t-1},X_t)|(X_{t-2},X_{t-1}))}{dP^*_1((X_{t-1},X_t)|(X_{t-2},X_{t-1}))} - 1 \right)^2 \leq \eta_2
\]
\{X_t : t = 1, \ldots, T\} is a 2-dependent stationary process under \( P_f \)
\[
P_f(x_{t-1},X_t) = P_{f_i}^*(x_{t-1},x_t)
\]
\( P_f \in \mathcal{P}_0 \)

by using the augmented state space representation. Under this representation, the formulation (46) reduces into (11), and Theorem 1 (the non-i.i.d. benchmark version discussed in Section 4.4) implies an optimal value of \( E_{\tilde{f}_1} h + sd_{\tilde{f}_1} (S_{\tilde{f}_1} (X,Y,Z)) \sqrt{\eta_2} + O(\eta_2) \), where \( S_{\tilde{f}_1} (X,Y,Z) \) is the projection of the symmetrization in two-state augmented state space representation
\[
\tilde{H}_{f_1,2}(x,y,z) := \sum_{t=3}^{T} E_{f_1} [h|(X_{t-2},X_{t-1}) = (x,y), (X_{t-1},X_t) = (y,z)]
\]
\[
= \sum_{t=3}^{T} E_{f_1} [h|X_{t-2} = x, X_{t-1} = y, X_t = z]
\]
ono{onto}
\[
\tilde{M}_{f_1} := \left\{ V(X,Y,Z) \in \mathcal{L}_2(P_{f_1}) : E_{f_1}[V|X,Y] = E_{f_1}[V|Y,Z] = 0 \ a.s. \right\}
\]
where \((X,Y)\) and \((Y,Z)\) are two consecutive states in the augmented representation of \( P_{f_1}^* \). Therefore, from (44) we have \( Z^* - E_0 h \leq \eta_1 f_1 (S_{\tilde{f}_1} (X,Y,Z)) \sqrt{\eta_2} + O(\eta_2) + sd_0 (R(X,Y)) \sqrt{\eta_1} + O(\eta_1) \).

To conclude Theorem 3, one has to make two observations: first, the object \( S(X,Y,Z) \) defined in (42) is precisely the projection of
\[
\tilde{H}_{0,2}(x,y,z) := \sum_{t=3}^{T} E_0 [h|(X_{t-2},X_{t-1}) = (x,y), (X_{t-1},X_t) = (y,z)]
\]
\[
= \sum_{t=3}^{T} E_0 [h|X_{t-2} = x, X_{t-1} = y, X_t = z]
\]
ono{onto}
\[
\tilde{M}_0 := \left\{ V(X,Y,Z) \in \mathcal{L}_2(P_0) : E_0[V|X,Y] = E_0[V|Y,Z] = 0 \ a.s. \right\}
\]
i.e. \( S(X,Y,Z) \) is defined as \( S_{\tilde{f}_1} (X,Y,Z) \) but with the probability measure \( P_{f_1}^* \) replaced by the benchmark \( P_0 \). Second, one has to demonstrate that \( sd_{\tilde{f}_1} (S_{\tilde{f}_1} (X,Y,Z)) \rightarrow sd_0 (S(X,Y,Z)) \) as \( \eta_1 \rightarrow 0 \). These arguments involve some analysis using the projection properties, and will be laid out in detail in Appendix B.
In general, for $p$-dependence, we use the formulation

$$\max \text{ / } \min \quad E_f[h(X_T)]$$

subject to

$$\phi_k^2(P_f(X_{t-k}, X_{t-k+1}, \ldots, X_t)) \leq \eta_k, \quad k = 1, \ldots, p$$

$\{X_t : t = 1, \ldots, T\}$ is a $p$-dependent stationary process under $P_f$

(47)

$P_f(x_t) = P_0(x_t)$ a.e.

$P_f \in P_0$.

Here $\phi_k^2(P_f(X_{t-k}, X_{t-k+1}, \ldots, X_t)) = E_{f_{k-1}}((L_k - 1)^2$ where

$$L_k = \frac{dP_f(X_t|X_{t-k}, \ldots, X_{t-1})}{dP_{f_{k-1}}(X_t|X_{t-k}, \ldots, X_{t-1})} = \frac{dP_f(X_{t-k-1}, \ldots, X_t)}{dP_{f_{k-1}}(X_{t-k}, \ldots, X_t)}$$

By using essentially the same arguments for Theorem 3, we can obtain the following estimates for formulation (47):

**Theorem 4.** The optimal value of the max formulation in (47) is bounded above by

$$\max E_f h \leq E_0 h + \sum_{k=1}^p sd_0(R_k(X_{k+1}))\sqrt{\eta_k}I(k < T) + o\left(\sum_{k=1}^p \sqrt{\eta_k}\right)$$

(48)

and the min formulation is bounded below by

$$\min E_f h \geq E_0 h - \sum_{k=1}^p sd_0(R_k(X_{k+1}))\sqrt{\eta_k}I(k < T) + o\left(\sum_{k=1}^p \sqrt{\eta_k}\right)$$

(49)

where

$$R_k(x_{k+1}) = \sum_{t=k+1}^T E_0[h|X_{t-k} = x_1, X_{t-k+1} = x_2, \ldots, X_t = x_{k+1}]$$

$$- \sum_{t=k+1}^T E_0[h|X_{t-k} = x_1, X_{t-k+1} = x_2, \ldots, X_{t-1} = x_k]$$

$$- \sum_{t=k+1}^T E_0[h|X_{t-k+1} = x_2, X_{t-k+2} = x_3, \ldots, X_t = x_{k+1}]$$

$$+ \sum_{t=k+1}^T E_0[h|X_{t-k+1} = x_2, X_{t-k+2} = x_3, \ldots, X_{t-1} = x_k].$$

The indicators $I(k < T)$ in (48) and (49) reveal the trivial observation that deviation from $k - 1$ to $k$-dependence of the input process causes effect on the performance measure only when the time horizon $T$ is greater than $k.$

25
6 Numerical Implementation

In this section we discuss the numerical implementation of our method. We apply our bounds in Sections 4 and 5 to quantify the uncertainty due to input serial dependency for two examples: a queueing model and the hedging of a financial option.

6.1 Computing the First and Second Order “Nonparametric Derivatives”

Before going through the examples, we first spend this subsection to discuss how to compute the coefficients in our bounds. Consider the first order coefficient \( sd_0(R(X,Y)) \) in Theorem 1, where \( R(x,y) \), with fixed \( x \) and \( y \), can be expressed as the expectation

\[
R(x,y) = \sum_{t=2}^{T} E_0 \left[ h(X_{t-1}^{X_{t-1}=x,X_t=y}) - h(X_{t}^{X_{t-1}=y}) - h(X_t^{X_{t-1}=y}) + h(X_T) \right]
\]

(50)

Here we use the notation similar as before that \( h(X_{t-1}^{X_{t-1}=x,X_t=y}) \) denotes the cost function with \( X_{t-1} \) fixed at \( x \) and \( X_t \) fixed at \( y \), but other \( X_t \)'s are randomly generated etc. Hence, from either (50) and (51), we see that \( Var_0(R(X,Y)) \) is in the form of the variance of conditional expectation. One consistent unbiased procedure for simulating such object is proposed in [32]. Their idea is to use nested simulation, by treating the variance of conditional expectation as the between-group variance in ANOVA study, and deriving unbiased formula through manipulating the sum-of-squares. We shall adapt their method to our setting. For this purpose we need to introduce some notation. We let \( K \) be the number of “outer” samples, and \( n \) be the number of “inner” samples. The first step is to generate \( K \) copies of i.i.d. couple \((X,Y)\) under \( P_0 \) (the outer simulation). Say \((x_1,y_1), \ldots, (x_K,y_K)\) are the \( K \) realizations. Then for each \((x_k,y_k)\), conditional on its value, generate \( n \) copies of

\[
Z_{kj} = \sum_{t=2}^{T} \left( h(X_{t-1}^{X_{t-1}=x_k,X_t=y_k}) - h(X_{t}^{X_{t-1}=y}) - h(X_t^{X_{t-1}=y}) + h(X_T) \right)
\]

(52)

where the involved \( X_t \)'s are simulated using new replications each time they show up, except the ones that are fixed as \( x_k \) or \( y_k \). These \( Z_{kj}, k = 1, \ldots, K, j = 1, \ldots, n \) are our inner samples.

The ANOVA-based estimator for \( Var_0(R(X,Y)) \) is then given by

\[
\hat{\sigma}_M^2 = \frac{1}{K-1} \sum_{k=1}^{K} (Z_k - \bar{Z})^2 - \frac{1}{n} \hat{\sigma}_\epsilon^2
\]

(53)

\[\text{We note that there are other proposals for solving similar problems, such as [19], which are left for future exploration.}\]
where
\[ \hat{\sigma}_c^2 = \frac{1}{K(n - 1)} \sum_{k=1}^{K} \sum_{j=1}^{n} (Z_{kj} - \bar{Z}_k)^2 \]
and
\[ \bar{Z}_k = \frac{1}{n} \sum_{j=1}^{n} Z_{kj}, \quad \bar{Z} = \frac{1}{K} \sum_{k=1}^{K} \bar{Z}_k. \]

The work [32] provides some guidelines for choosing K and n. Roughly speaking, for a growing simulation budget, n is better fixed at a small number while K increases with the budget. This will lead to a minimization of the standard error. The precise optimal values of n and K depend on the value of \( \text{Var}_0(R(X,Y)) \) itself and hence can only be guessed from pilot estimation. When K increases, the standard error goes to zero. In our simulation experiments we will choose n to be 200, and K to be 1,000.

To construct a confidence interval for \( sd_0(R(X,Y)) = \sqrt{\text{Var}_0(R(X,Y))} \), one can use sectioning [1]. Generate N (for example, \( N = 20 \)) replications of \( \hat{\sigma}_c^2 \) in (53). Say they are \( W_i, i = 1, \ldots, N \). An asymptotically valid 1 − \( \alpha \) confidence interval for \( sd_0(R(X,Y)) \) is then
\[ \left( \sqrt{\bar{W}} \pm \frac{\nu}{\sqrt{N}} \cdot \frac{t_{1-\alpha/2, N-1}}{\sqrt{N}} \right) \]
where \( \bar{W} = (1/N) \sum_{i=1}^{N} W_i \) is the sample mean, and \( \nu^2 = (1/(N - 1)) \sum_{i=1}^{N} (W_i - \bar{W})^2 \) is the sample variance of \( W_i \)'s. The quantity \( t_{1-\alpha/2, N-1} \) is the \( (1 - \alpha/2) \)-quantile of the t-distribution with degree of freedom \( N - 1 \).

Note that we can as well use the formula in (51) to generate our inner sample in (52), or in fact, any other formulas (such as a stratified sampling on the time horizon), as long as they are unbiased for the conditional expectation. The formula (50) should possess less variance than (51) in general, because of a derandomization of the averaging in time horizon. However, for long-run performance measures, some version of time randomization is necessary to ensure computational feasibility. We note that such randomization technique for approximating long summation bears a spirit close to Monte Carlo linear algebraic method used in solving large-scale dynamic programming (such as least squares temporal differences or policy evaluation; see, for example, [2]).

The nested simulation method as described above will be used in getting most of our numerical results in the sequel. Here we present another way to compute \( \text{Var}_0(R(X,Y)) \), which will also be useful for simulating the second order coefficients in the expansions in Theorem 1 (although it is less efficient than the method above). The idea is to simply observe that \( \text{Var}_0(R(X,Y)) = E_0[R_1(X,Y)R_2(X,Y)] \), where \( R_1(x,y) \) and \( R_2(x,y) \) can be any unbiased estimators for \( R(x,y) \) that are independent given \( (x,y) \). This leads to a scheme as follows: First sample an i.i.d. couple \( (X,Y) \) under \( P_0 \). Suppose the realization is \( (x,y) \). Then, given this \((x,y)\), sample 2n independent
copies of
\[ Z_j = \sum_{t=2}^{T} \left( h(\mathbf{X}_T^{(X_{t-1}=x,X_t=y)}) - h(\mathbf{X}_T^{(X_{t-1}=x)}) - h(\mathbf{X}_T^{(X_t=y)}) + h(\mathbf{X}_T) \right) \]
(or any version that is unbiased for \( R(x,y) \), such as (51)). Then \((1/n)\sum_{j=1}^{n} Z_j \cdot (1/n)\sum_{j=n+1}^{2n} Z_j \) will give one unbiased sample for \( \text{Var}_0(R(X,Y)) \). Note that one can merely put \( n = 1 \), but our numerical test (especially for the second order coefficient discussed below) suggests that it is better off to use a large \( n \), because then \( R_1(X,Y) \) and \( R_2(X,Y) \), being the sample averages of \( Z_j \), will be close to 0; this will mimic the magnitude of \( \text{Var}_0(R(X,Y)) \), which is typically small.

The second method described above can be extended to compute the second order coefficient. This is especially apparent by expressing \( R \) using formula (51), which proceeds as follows. First, sample a couple of time points \((s,t)\), with the restriction that \( s < t \), uniformly over \( \{2, \ldots, T\} \). Then:

1. If \( t - s > 1 \), then sample four copies of \( X \)'s under \( P_0 \), say \((x,y,z,w)\). Generate, given these four numbers, \( n \) samples of \( A_j = h(\mathbf{X}_T^{(X_{t-1}=x,X_t=y)}) - h(\mathbf{X}_T^{(X_{t-1}=x,z)}) - h(\mathbf{X}_T^{(X_{t-1}=x,y)}) + h(\mathbf{X}_T) \) and \( n \) samples of \( B_j = h(\mathbf{X}_T^{(X_{t-1}=z,X_t=w)}) - h(\mathbf{X}_T^{(X_{t-1}=z)}) - h(\mathbf{X}_T^{(X_t=w)}) + h(\mathbf{X}_T) \) independently. Also generate an independent sample of \( C = h(\mathbf{X}_T^{(X_{t-1}=x,y,x_{t-1}=z,x_t=w)}) \).

2. If \( t - s = 1 \), so that \( s = t - 1 \), then sample three copies of \( X \)'s under \( P_0 \), say \((x,y,z)\). Generate, given these three numbers, \( n \) samples of \( A_j = h(\mathbf{X}_T^{(X_{t-2}=x,X_{t-1}=y)}) - h(\mathbf{X}_T^{(X_{t-2}=x,z)}) - h(\mathbf{X}_T^{(X_{t-1}=y)}) + h(\mathbf{X}_T) \) and \( n \) samples of \( B_j = h(\mathbf{X}_T^{(X_{t-1}=y,X_t=z)}) - h(\mathbf{X}_T^{(X_{t-1}=y,z)}) - h(\mathbf{X}_T^{(X_{t-1}=z)}) + h(\mathbf{X}_T) \) independently. Also generate an independent sample of \( C = h(\mathbf{X}_T^{(X_{t-2}=x,X_{t-1}=y,x_{t-1}=z)}) \).

For each generation of \((s,t)\), the quantity \((\frac{1}{2})(1/n)\sum_{j=1}^{n} A_j \cdot (1/n)\sum_{j=1}^{n} B_j \cdot C \) will give one unbiased sample of \( \sum_{s,t=2,\ldots,T} E_0[R(X_{s-1},X_{s})R(X_{t-1},X_t)h(\mathbf{X}_T)] \). This can then be combined with the estimate of \( \text{Var}_0(R(X,Y)) \) as described before to get an estimate of the second order coefficient.

The coefficients for higher-order dependency assessment in Theorems 3 and 4 can be computed analogously.

### 6.2 Numerical Example 1: Queueing Model

We perform a “proof-of-concept” numerical experiment on the transient behavior of a first-come-first-serve \( M/M/s \) queue. In the benchmark model, customers arrive according to i.i.d. interarrival times and are served with i.i.d. service times. The performance measure that we shall look at is the tail probability of the waiting time of the \( T \)th customer in the queue, starting from, say, an empty system status. It is known that interarrival times in many real situations can be correlated. We compute our bounds to assess the discrepancy of our performance metric when the arrival process deviates in a neighborhood of i.i.d. to a 1-dependent process measured by the \( \phi^2 \)-coefficient.
To start with, we check the qualitative behavior of our “nonparametric derivatives”. We compute the first order coefficient in Theorem 1 using the nested simulation method described in Section 6.1. For each parameter set, we use 1,000 number of outer simulation runs and 200 inner simulations, and we use 20 replications to obtain the macro estimate and to construct confidence interval. As we have explained in previous sections, it is intuitive that our coefficient should increase with the number $T$, since $T$ controls the number of replications of the input model in the system. Figure 1 and Table 1 show the point estimates and the 95% confidence intervals of our coefficient $sd_0(R(X, Y))$ against $T$, for $T$ running from 10 to 50, in a single-server queue. As we can see, there is indeed a steady increase in the magnitude of $sd_0(R(X, Y))$. The point estimate of $sd_0(R(X, Y))$ increases from around 0.09 when $T = 10$ to 0.17 when $T = 50$. The small downward trend from $T = 40$ to $T = 50$ is possibly due to the kick-in of steady-state effect, so that the performance measure for $T = 40$ and 50 are very close to each other (or can be merely statistical error, as the confidence intervals overlap). Throughout the simulations we use an arrival rate of 0.8 and service rate 1, hence the system is in the moderate traffic regime. We use a threshold level of 2 for the waiting time tail probability.

Next, we also investigate the relation between the magnitude of the coefficient and the number of servers. We fix $T = 30$ and service rate 1. We then vary the number of servers from 1 to 5. The arrival rate is set to be 0.8 times the number of servers. This is to keep the traffic intensity to be 0.8 universally. Figure 2 and Table 2 show our estimates. When the number of servers increase, the coefficient appears to shrink, from around 0.16 for 1 server to 0.04 for 5 servers. This suggests that a system with more servers is more resistant to dependency misspecification of the arrival process.

![Figure 1: Plot of first order coefficient in Theorem 1 against the customer position of interest](image)

<table>
<thead>
<tr>
<th>Customer Position of Interest</th>
<th>Point Estimate</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>0.0915</td>
<td>(0.0884,0.0946)</td>
</tr>
<tr>
<td>20</td>
<td>0.1296</td>
<td>(0.1232,0.1361)</td>
</tr>
<tr>
<td>30</td>
<td>0.1611</td>
<td>(0.1494,0.1728)</td>
</tr>
<tr>
<td>40</td>
<td>0.1669</td>
<td>(0.1529,0.1809)</td>
</tr>
<tr>
<td>50</td>
<td>0.1654</td>
<td>(0.1507,0.1800)</td>
</tr>
</tbody>
</table>

![Table 1: Point estimates and confidence intervals for the first order coefficients in Theorem 1 in relation to the customer position of interest](image)
Figure 2: Plot of first order coefficient in Theorem 1 against the number of servers

### Table 2: Point estimates and confidence intervals for the first order coefficients in Theorem 1 in relation to the number of servers

<table>
<thead>
<tr>
<th>Number of Servers</th>
<th>Point Estimate</th>
<th>95% C.I.</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.1611</td>
<td>(0.1494, 0.1728)</td>
</tr>
<tr>
<td>2</td>
<td>0.1703</td>
<td>(0.1618, 0.1788)</td>
</tr>
<tr>
<td>3</td>
<td>0.1296</td>
<td>(0.1259, 0.1333)</td>
</tr>
<tr>
<td>4</td>
<td>0.0814</td>
<td>(0.0761, 0.0867)</td>
</tr>
<tr>
<td>5</td>
<td>0.0423</td>
<td>(0.0392, 0.0454)</td>
</tr>
</tbody>
</table>

Our next step is to assess the performance of our local deviation bound against a parametric autoregressive arrival model. To this end, define a stationary AR(1) sequence $W_t = \beta_0 + \beta_1 W_{t-1} + \epsilon_t$, with $\epsilon_t \sim N(0, \sigma^2)$. We then construct the interarrival times $A_t$, with exponential rate $\lambda$, by doing a transformation $A_t = F^{-1}_\lambda(\Phi(W_t; \beta_0/(1 - \beta_1), \sigma^2/(1 - \beta_1^2)))$, where $F(\cdot)$ is the cdf of Exp($\lambda$) and $\Phi(\cdot; a, b)$ is the cdf of a Gaussian variable with mean $a$ and variance $b$. Note that $W_t$ has stationary distribution $N(\beta_0/(1 - \beta_1), \sigma^2/(1 - \beta_1^2))$. Hence $A_t$ as defined is distributed as correlated Exp($\lambda$).

The parameter $\beta_1$ encodes the correlation among successive $W_t$, and hence $A_t$. We do the following experiment. We set $\lambda = 0.8, \beta_0 = 1$ and $\sigma^2 = 0.5$, and service times as i.i.d. Exp(1) in an $M/M/1$ system. In Figure 3, the blue horizontal line, fixed at around 0.48, is the probability that the waiting time of the 30-th customer is greater than 2 (our performance measure) when we fix $\beta_1 = 0$, i.e. when the arrivals are i.i.d. This value is estimated using 1 million sample paths. We then vary the parameter $\beta_1$ from $-0.5$ to 0.5, and simulate using the stationary arrival sequence of $A_t$ with correlation parameter $\beta_1$ on the $M/M/1$ system. The sample means of the performance measures are shown as the black dots, each with confidence interval depicted by the two surrounding white dots. We use a simulation size of 50,000 for each $\beta_1$.

The red curves in Figure 3 are the upper and lower bounds computed from Theorem 1 using only the first order term (the dashed red curves are bounds calculated using the upper and lower limits of the confidence interval of the first order coefficient). The $\eta$ in the bound is computed by $\phi^2(P_f(A_{t-1}, A_t))$, or equivalently the $\phi^2$-coefficient of a bivariate Gaussian distribution with correlation $\beta_1$, i.e. $\beta_1^2/(1 - \beta_1^2)$ (note that $\chi^2$-distance is invariant to transformation of variables). In other words, the red curves are $\hat{E}_0 h \pm sd_0(R(X,Y)) \sqrt{\beta_1^2/(1 - \beta_1^2)}$. As we can see, the first order term in Theorem 1 seems to provide reasonably tight bounds for the performance output of the AR(1) model. The upper red curve matches closely with the performance measures when $\beta_1$ is positive, and the lower red curve provides a lower bound when $\beta_1$ is negative. Interestingly, this linear approximation (in $\eta$) works well even when the magnitude of $\beta_1$ is as large as 0.4.
Figure 3: Performance of the bounds in Theorem 1 against actual waiting time tail probabilities; red curves are the bounds computed using the first order terms in (16) and (17)

We comment that the estimation of the coefficients in Theorem 1 using the nested simulation approach, with the use of formula (50), is computationally intensive: to obtain one inner sample it requires $O(T^2)$ computational load. We believe that for large $T$, randomization of the time horizon using formula (51) will better balance the computational burden with sampling accuracy. The choice of sampling schemes, as well as the design of potential variance reduction techniques, will be left to future investigation.

Lastly, we also ran experiment to estimate the second order coefficient in Theorem 1 for the single-server setting with $T = 10$. We carried out the method described in Section 6.1, using 10,000 samples of $(s, t)$ pairs and for each $(s, t)$ pair, we generated $n = 2,500$ samples of $A_j$ and $B_j$ respectively. The magnitude of the second order coefficient appears to be tiny. The sample mean of

$$(1/n) \sum_{j=1}^n A_j \cdot \sum_{j=1}^n B_j \cdot C \text{ using the 10,000 trials is } -1.1644 \times 10^{-4}, \text{ with a 95\% confidence interval } (-6.8187 \times 10^{-4}, 4.4900 \times 10^{-4}) \text{ (which translates to } -0.4636 \text{ for the second order coefficient after multiplying by } \left(\frac{1}{2}\right) \text{ and dividing by the estimate of } \text{Var}_0(R(X, Y))).$$

We performed this estimation several times (the above numeric comes from a typical output), and each time the 95\% confidence interval covers zero, which means statistically speaking the coefficient is no different from zero to the degree of accuracy that the experiment can capture. In other words, the present method can only suggest a very tiny second order effect. The exact magnitude of the second order coefficient is yet to be determined, but it should be at most in the order of a tenth decimal place.
6.3 Numerical Example 2: Delta Hedging

Our next example is the delta hedging error for a European option. This setting was studied in [11], who also considered a wide range of other problems in finance. Here our comparison focuses on the case when the stock price of an underlying asset deviates from geometric Brownian motion assumption to autoregressive type process, a structure that is amenable to the method studied in this paper.

We adopt some notation from [11]. Consider a European call option, i.e. the payoff of the option is $(X_T - K)^+$, where $T$ is the maturity, $\{X_t\}_{t \geq 0}$ is the stock price, and $K$ is the strike price. We assume a Black-Scholes benchmark model for the stock, namely that the stock price follows a geometric Brownian motion

$$dX_t = \mu X_t dt + \sigma X_t dB_t$$

where $B_t$ is a standard Brownian motion, $\mu$ is the growth rate of the stock, and $\sigma$ is the volatility. Assuming trading can be executed continuously over time without transaction fee, then it is well known that a seller of the option at time 0 can perfectly hedge the payoff of the option by investing

$$\delta(t, X_t) = \Phi((\log(X_t/K) + (r - (1/2)\sigma^2)(T - t))/(\sigma \sqrt{T - t})),$$

i.e. the “delta”, units of stock at any point of time, with the rest of the portfolio held as cash. Here $r$ is the risk-free rate. This self-balancing portfolio will exactly offset the payoff of the option at maturity. In practice, however, trading cannot be done continuously, and implementing the Black-Scholes strategy will incure a discretization error. To be more specific, suppose that trading is allowed at times $k\Delta t$ for $k = 0, 1, 2, \ldots, T/(\Delta t)$, and that the trader holds the amount of cash $C(0)$ and stock $S(0)$:

$$C(0) = BS(0, T, X_0) - X_0 \delta(0, X_0)$$
$$S(0) = X_0 \delta(0, X_0)$$

at time 0, and cash $C(k\Delta t)$ and stock $S(k\Delta t)$:

$$C(k\Delta t) = e^{r\Delta t} C((k - 1)\Delta t) - X_{k\Delta t}[\delta(k\Delta t, X_{k\Delta t}) - \delta((k - 1)\Delta t, X_{(k-1)\Delta t})]$$
$$S(k\Delta t) = X_{k\Delta t}\delta(k\Delta t, X_{k\Delta t})$$

at time $k\Delta t$. Here $BS(0, T, X_0)$ is the Black-Scholes price for a European call option with maturity $T$ and initial stock price $X_0$, and $\delta(k\Delta t, X_{k\Delta t})$ is the “delta” of the option at time $k\Delta t$ with current stock price $X_{k\Delta t}$. The delta hedging error is given by

$$H_e = (X_T - K)^+ - C(T) - S(T)$$

and the performance measure is set to be $E|H_e|$.

We consider the case when $\{X_{k\Delta t}\}$ deviates from i.i.d. lognormal distribution to 1-dependent process. Although $|H_e|$ is unbounded in theory, and hence violating the assumption in Theorem 1,
for all practical purpose (and in all the simulation runs we have done) this variable only takes small numbers. Hence the boundedness technicality should not be a concern in applying our proposed method (for anxious users, one can merely consider a truncated hedging error as the cost function by putting a large enough cap on $|H_e|$, which would not alter any of our simulation results). Figure 4 depicts the performance of the first order term in Theorem 1 against an AR(1) model. We use $T = 1$ and $\Delta t = 0.01$, so that the number of periods in consideration is 100. The initial stock price and the strike price are both set to be 100, i.e. the call option is at-the-money. We set the benchmark geometric Brownian motion to have growth rate $\mu = 0.1$ and volatility $\sigma = \sqrt{0.2}$, and the risk-free rate as $r = 0.05$. The first order term of the bound in Theorem 1 is computed by using $1,000$ outer samples, $200$ inner samples and $20$ macro replications, which gives a point estimate of $1.5477$, with 95% confidence interval $(1.3820, 1.7134)$. The delta hedge error at the benchmark is estimated to be $0.5021$ from $100,000$ sample paths, which is depicted by the blue horizontal line in Figure 4. The brown curves in the figure are the upper and lower bounds calculated using only the first order term in Theorem 1. As a comparison to our bound, we compute $E|H_e|$ for $\{X_{k\Delta t}\}$'s that has logarithmic increment satisfying an AR(1) model, i.e. $\Delta \log X_t = \beta_0 + \beta_1 \Delta \log X_{t-1} + \epsilon_t$ with $\beta_1$ ranging from $-0.2$ to $0.2$, and the stationary distribution kept at lognormal($\mu - (1/2)\sigma^2 \Delta t, \sigma^2 \Delta t$).

The black dots in Figure 4 show the point estimates of $E|H_e|$ at different $\beta_1$, with the two bars above and below each point depicting the 95% confidence interval. Each point estimate is obtained using $20,000$ sample paths. Similar to the example in the previous subsection, the $\eta$ in our bound is set to be $\phi^2(P_f(\Delta \log X_{t-1}, \Delta \log X_t))$ for each $\beta_1$. One can see that as $\beta_1$ moves from negative to positive, the hedging error generally decreases. The brown curves provide valid covering interval for the hedging error throughout the whole range of $\beta_1$ from $-0.2$ to $0.2$. The upper curve is especially tight in the negative region, with the curve almost touching the trend of black dots all the way to $\beta_1 = -0.2$.

Next, we compare our bounds to the method suggested in [11], which considers general model uncertainty quantification. This comparison aims to demonstrate the gain in accuracy when assessment scheme is derived in a way that targets at specific statistical feature, namely serial dependency in this paper. The method in [11] is based on relative entropy. Part of their work suggested viewing the sample path $\{X_{k\Delta t}\}$ as a single random object and solving the optimization problem (adapted to our notation) $\max E_f[h(\{X_{k\Delta t}\})]$, where the optimization is over a Kullback-Leibler constraint $D(P_f(\{X_{k\Delta t}\})\parallel P_0(\{X_{k\Delta t}\})) \leq \eta$. The optimal solution to this program is characterized by an exponential change of measure $\propto \exp\{h(\{X_{k\Delta t}\})/\alpha\}$, where $\alpha > 0$ is the dual optimal solution. For a deviation from i.i.d. lognormal to AR(1) model over $T$ time steps (while keeping the stationary distribution unchanged), the involved Kullback-Leibler (KL) divergence on the sample path $\{X_{k\Delta}\}$ can be calculated to be $(T - 1)\log(1/(1 - \beta_1^2))$ (we note that KL divergence is more convenient
than $\chi^2$-distance when computing the model deviation of the whole sample path, since the former is decomposable as the sum of individual divergences).

The green dashed curves in Figure 5 depicts the bound obtained from solving the above optimization problem. To compute this bound for each value of $\beta_1$, or the corresponding $\eta$, we first approximately solve for the inverse of the optimal dual solution, i.e. $1/\alpha > 0$, via a sample average approximation using 100,000 benchmark sample paths. This can be done either by setting the complementarity condition or by minimizing the dual program, both of which lead to identical result. Then we average the 100,000 sample paths using the exponential tilting weights with the tilting parameter being the estimated $1/\alpha$, which gives the upper green curve. The lower green curve is obtained analogously by solving for a negative $1/\alpha$. We can see that the green curves bound the black dots that represent the actual hedging errors as $\beta_1$ varies.

The brown curves in Figure 5 are the first order bounds in Theorem 1, i.e. the same brown curves as in Figure 4. The increase in the tightness of the bounds is quite substantial; for example, the gap goes down from a single digit to the order of one tenth in the case $\beta_1 = \pm 0.2$. This suggests that the bound in Theorem 1 is much more accurate in assessing the effect of first order dependency compared to a general aggregate KL constraint. Of course, this is because we have made a stronger structural assumption on model discrepancy and have devised a specialized method, whereas the general bound in [11] treats the whole trajectory as a single object, and as such they are robust against deviations other than serial dependency. This phenomenon highlights the tradeoff between robustness and accuracy: for a user who is well-informed with the marginal distribution and who is concerned about serial dependency, it is better off to adopt the approach in this paper to get tighter error bound; on the other hand, a user who has only minimal knowledge about the sample
Figure 5: Comparison of the bounds in Theorem 1 and the method based on aggregate constraint not targeted at serial dependency; brown curves are computed from Theorem 1, green curves are based on aggregate constraint path, including marginal distribution, could turn to the bound in [11] that is coarser but is capable to safeguard against a wider sort of model misspecification.

We now test the performance of our sensitivity estimates for higher order dependency discussed in Section 5. In particular, we focus on the bounds in Theorem 3, and compare it against a parametric AR(2) model. Consider a stock price with logarithmic increment $\Delta \log X_t = \beta_0 + \beta_1 \Delta \log X_{t-1} + \beta_2 \Delta \log X_{t-2} + \epsilon_t$. We allow $\beta_1$ and $\beta_2$ to vary, with $\beta_0$ and the variance of $\epsilon_t$ chosen so that the stationary distribution is kept at lognormal($((\mu - (1/2)\sigma^2)\Delta t, \sigma^2 \Delta t$).

Figure 6 compares our bounds with the AR(2) model when $\beta_1$ is fixed at 0.1, and $\beta_2$ varies from $-0.2$ to 0.2. The black dots and the bars above and below them depict the point estimates and the 95% confidence intervals of the actual hedging errors using 20,000 sample paths. There is a downward trend in the hedging error as $\beta_2$ increases. Note that when $\beta_2 = 0$, it reduces to an AR(1) model. The green dashed lines are the upper and lower bounds computed using only the first order coefficient in Theorem 1, or in other words, the bounds that are suitable for only first order dependency assessment. As we can see, they provide a valid covering interval for the case $\beta_2 = 0$: the lower green line almost touches the black dot. However, when $\beta_2$ moves away from 0, the interval between the lines no longer contain the black dots, suggesting that the model under scrutiny goes beyond first order dependency.

Encouragingly, our bounds are effective once we extend them to second order dependency effect. The two brown curves in Figure 6 are computed from the upper and lower bounds in Theorem 3. The second order coefficient in (40) and (41) is estimated to be 0.8879, with a 95% confidence interval of (0.8687, 0.9071), using 1,000 outer samples, 200 inner samples and 20 macro replications. The
η₁ in the formula is set to be $\beta_1^2/(1 - \beta_1^2)$, while the η₂ parameter, which controls the deviation from first to second order dependency, is numerically calculated using (37) and the observation in Example 1. After adding the term controlling the second order dependency, we see that the brown curves succeed in bounding the black dots. In fact, they seem to be reasonably tight, which also suggests the approximate tightness of the inequalities in Theorem 3.

As a further test, we also vary the value of $\beta_1$ from $-0.2$ to $0.2$, this time with $\beta_2$ fixed at $0.1$. Figure 7 shows our result. Again, the black dots and the bars locate the point estimates and the 95% confidence intervals for the actual hedging errors, which show a downward trend as $\beta_1$ increases. The green dashed curves are the upper and lower bounds computed using Theorem 1 with only the first order term, i.e. they are bounds that are targeted only for first order dependency. When $\beta_1 = 0$, the bounds degenerate to a single point at the benchmark performance measure. When $\beta_1$ deviates from 0, the bounds naturally widen, but still are not enough to contain the black dots, especially for small positive $\beta_1$. The brown curves depict the bounds computed using Theorem 3, which covers the case of second order dependency. We see that there is an improvement for the region of small positive $\beta_1$. In particular, the lower bound is approaching the locations of the black dots in this region.

7 Concluding Remarks

In this paper we propose a nonparametric, adversarial approach in quantifying the sensitivity of input serial dependency to a stochastic system. The approach outputs local estimates for the
Figure 7: Performance of bounds in Theorem 3 against an AR(2) model as $\beta_1$ varies; green dashed lines are first order bounds computed using Theorem 1, and brown curves are second order bounds computed using Theorem 3.

worst- and best-case changes in system output subject to input deviation in terms of statistical distance. These estimates are tight in the case of 1-dependency, and are in the form of conservative bounds when higher order dependency is present. As we have mentioned in the introduction, the approach that we are proposing is a general methodology: by placing appropriate constraints on the associated optimization programs, other statistical features from the input model can be assessed. One example of these other types of constraint is moment condition, which is widely used in distributional robust optimization. In many practical situations, moments of the input model are stably estimable from past data. Encapsulating such information can reduce input uncertainty and help yielding less conservative error estimates. Analyzing this type of constraints thus constitutes one of our future research directions. Moreover, one other important future direction is to extend our framework beyond finite horizon problems, studied in the current paper, to other types of system estimation, such as random horizon problems, steady-state and quantile estimation. We believe that appropriate adaptation of the analysis proposed in this paper can bring in new insight as well as practical tools to assess input model uncertainty for these other problem variations.

Lastly, we point out a potential use of our methodology that we have kept silent so far: the construction of iterative method for computing bounds for more “global” assessment of model uncertainty. The technique that we developed in this paper is clearly a local analysis. The assumption that the input model only deviates in a neighborhood from the reality means that our method, at least in theory, does not provide any guarantee for models that are far apart. One can argue that this is a limitation of “derivative” type sensitivity analysis in general, and that in this case the benchmark model should not be used in the first place. Nevertheless, we believe our local
analysis can potentially help assess model uncertainty even when the truth is far off, by suggesting a well-defined local ascendency move in optimization problems that possess more “global” type of distributional constraints. Notice that our results, such as Theorem 1, define a precise “steepest descent” direction in the space of probability measures. Naturally, this points to the construction of iterative scheme that converges to local optimum for optimization programs with similar kind of dependency constraints. The construction of this sort of iterative schemes over the probability space appears to be new.

A Selected Proofs

Proof of Proposition 2. The first step is to relax the constraints \( E_0[L|x] = E_0[L|y] = 1 \) a.s., and consider

\[
E_0[h(X, Y)L] - \alpha (E_0(L-1)^2 - \eta) + \int (E_0[L|x] - 1)d\beta(x) + \int (E_0[L|y] - 1)d\gamma(y)
\]  

(54)

where \( \beta \) and \( \gamma \) are signed measures with bounded variation, i.e. \( \int |d\beta|(x), \int |d\gamma|(y) < \infty \). Then, differentiating (54) as a heuristic Euler-Lagrange equation gives

\[
h(x, y) - 2\alpha L(x, y) + \beta(x) + \gamma(y) = 0
\]

which leads to

\[
L = \frac{1}{2\alpha}(h(x, y) + \beta(x) + \gamma(y)).
\]

(55)

Since \( E_0[L] = 1 \), we must have \( E_0[L] = \frac{1}{2\alpha}(E_0[h(X, Y)] + E_0\beta(X) + E_0\gamma(Y)) = 1 \) or that

\[
E_0\beta(X) + E_0\gamma(Y) = 2\alpha - E_0[h(X, Y)].
\]

(56)

Moreover, \( E_0[L(X, Y)|x] = 1 \) and \( E_0[L(X, Y)|y] = 1 \) give

\[
\frac{1}{2\alpha}(E_0[h(X, Y)|x] + \beta(x) + E_0\gamma(Y)) = \frac{1}{2\alpha}(E_0[h(X, Y)|y] + E_0\beta(X) + \gamma(y)) = 1
\]

so that

\[
\beta(x) = 2\alpha - E_0[h(X, Y)|x] - E_0\gamma(Y)
\]

(57)

\[
\gamma(y) = 2\alpha - E_0[h(X, Y)|y] - E_0\beta(Y).
\]

(58)

Therefore, substituting (56), (57) and (58) into (55), we get

\[
L(x, y) = \frac{1}{2\alpha}(h(x, y) + 2\alpha - E_0[h(X, Y)|x] - E_0\gamma(Y) + 2\alpha - E_0[h(X, Y)|y] - E_0\beta(Y))
\]

\[
= \frac{1}{2\alpha}(h(x, y) + 2\alpha - E_0[h(X, Y)|x] + 2\alpha - E_0[h(X, Y)|y] - 2\alpha + E_0[h(X, Y)])
\]

\[
= 1 + \frac{1}{2\alpha}(h(x, y) - E_0[h(X, Y)|x] - E_0[h(X, Y)|y] + E_0h(X, Y))
\]

(59)
which coincides with (10). The expression (59) is our candidate optimal solution, and our next task is to verify that it is indeed an optimal solution. This means we have to prove that

\[ E_0[hL] - \alpha E_0(L - 1)^2 \leq E_0[hL^*] - \alpha E_0(L^* - 1)^2 \]  \tag{60}

for any \( L \in \mathcal{L} \), where \( L^* \) denotes the candidate optimal solution (59). To this end, recall the notation in Section 3 that \( \tilde{h} = h - r \), and that \( h \) possesses the orthogonal decomposition \( h = \tilde{h} + r \) on \( \mathcal{M}^\perp \oplus \mathcal{M} = \mathcal{L}_2 \). Then

\[
E_0[hL^*] - \alpha E_0(L^* - 1)^2 = E_0h + \frac{1}{2\alpha} E_0[r] - \frac{1}{4\alpha} E_0r^2 \\
= E_0h + \frac{1}{2\alpha} Var_0(r) - \frac{1}{4\alpha} Var_0(r) \quad \text{since } \tilde{h} \perp r \text{ and } E_0r = 0 \\
= E_0h + \frac{1}{4\alpha} Var_0(r)
\]

Going back to (60), consider any \( L \in \mathcal{L} \). For convenience, we write \( U = L - 1 \) so that \( E_0[U|X] = E_0[U|Y] = E_0U = 0 \) a.s.. Then we need to prove, for any such \( U \),

\[
E_0[h(U + 1)] - \alpha E_0U^2 \leq E_0h + \frac{1}{4\alpha} Var_0(r) \\
\iff E_0[hU] - \alpha E_0U^2 \leq \frac{1}{4\alpha} Var_0(r) \\
\iff E_0r^2 - 4\alpha E_0[hU] + 4\alpha^2 E_0U^2 \geq 0 \\
\iff E_0r^2 - 4\alpha E_0[rU] - 4\alpha E_0[hU] + 4\alpha^2 E_0U^2 \geq 0 \\
\iff E_0(r - 2\alpha U)^2 - 4\alpha E_0[hU] \geq 0. \tag{61}
\]

But \( E_0[hU] = 0 \) since \( U \in \mathcal{M} \). So (61) clearly holds.

\begin{proof}[Proof of Proposition 1] Since the formulation in (8) is convex, strong duality holds, so the optimal value of (9) equals (8). By complementary slackness, we have \( \alpha^*(E_0(L^* - 1) - \eta) = 0 \). Let us assume for now that \( \alpha^* > 0 \), which implies \( E_0(L^* - 1)^2 = \eta \). From (10), this in turn implies 

\[
\frac{1}{4\alpha^*} E_0r^2 = \eta \quad \text{or} \\
\frac{1}{4\alpha^*} Var_0(r) = \eta \tag{62}
\]

since \( E_0r = 0 \). Now the optimal objective value is

\[
E_0[hL^*] = E_0 \left[ h \left( 1 + \frac{r}{2\alpha^*} \right) \right] = E_0h + \frac{Cov_0(h, r)}{2\alpha^*} = E_0h + \frac{Var_0(r)}{2\alpha^*} \tag{63}
\]

since \( h = \tilde{h} + r \) with \( \tilde{h} \) orthogonal to \( r \). Combining with (62), we get \( E_0[hL^*] = E_0h + sd_0(r)\sqrt{\eta} \).

Lastly, suppose instead that \( \alpha^* = 0 \). Then the dual objective becomes \( \max_{L \in \mathcal{L}} E_0[hL] \). Since \( h \) is non-constant, there exists a measurable set \( A \in \text{supp}(P_0) \subset \mathcal{X}^2 \) such that \( h(x, y) > E_0[h] \) for any \( (x, y) \in A \). Now consider a feasible solution \( L \in \mathcal{L} \) that has support on \( A \). The resulting \( E_0[hL] \)
must be strictly larger than $E_0[h]$. Since we know, from (63), that as $\eta \to 0$, the minimum dual objective value $E_0[hL^*]$ over $\alpha > 0$ can be arbitrarily close to $E_0[hL]$, we conclude that $\alpha^* = 0$ is not possible.

**Proof of Theorem 1.** We follow the proof idea described at the end of Section 4. We first use the complementarity condition to find the relation between $\eta$ and $\alpha$, which allows us to use Theorem 2. This will then also facilitate our expansion of the objective value in terms of $\eta$. Now recall from (25) that $L^*$ satisfies $L^*(x, y) = 1 + \frac{R_{L^*}(x, y)}{2\alpha}$. We start with

$$H_{L^*}(x, y) = \sum_{t=2}^T E_0 \left[ h \prod_{s=2, \ldots, T \atop s \neq t} \left( 1 + \frac{R_{L^*}(X_{s-1}, X_s)}{2\alpha} \right) \bigg| X_{t-1} = x, X_t = y \right]$$

$$= \sum_{t=2}^T E_0[h|X_{t-1} = x, X_t = y] + \frac{1}{2\alpha} \sum_{s,t=1, \ldots, T \atop s \neq t} E_0[hR_{L^*}(X_{s-1}, X_s)|X_{t-1} = x, X_t = y] + \bar{O}\left(\frac{1}{\alpha^2}\right)$$

where $\bar{O}(1/\alpha^q)$ for $q > 0$ satisfies $|\bar{O}(1/\alpha^q)\alpha^q| \leq C$ for some $C$, for any $\alpha > 0$, uniformly over $x$ and $y$; the equality follows from the boundedness of $R_{L^*}(x, y)$ inherited from $h$

$$= H_2(x, y) + \frac{1}{2\alpha} W_{L^*}(x, y) + \bar{O}\left(\frac{1}{\alpha^2}\right)$$

(64)

where $H_2(x, y) = \sum_{t=1}^T E_0[h|X_{t-1} = x, X_t = y]$ is defined in (12), and

$$W_{L^*}(x, y) = \sum_{s,t=1, \ldots, T \atop s \neq t} E_0[hR_{L^*}(X_{s-1}, X_s)|X_{t-1} = x, X_t = y].$$

Now, through a “filtering” of the linear effect of $X$ and $Y$, (64) implies

$$R_{L^*}(x, y) = R(x, y) + \frac{S_{L^*}(x, y)}{2\alpha} + \bar{O}\left(\frac{1}{\alpha^2}\right)$$

(65)

where $R_L$ and $R$ are defined in (26) and (15), and

$$S_{L^*}(x, y) = W_{L^*}(x, y) - E_0[W_{L^*}|x] - E_0[W_{L^*}|y] + E_0 W_{L^*}.$$ 

Therefore we can write

$$E_0(L^* - 1)^2 = E_0 \left( \frac{R_{L^*}(X, Y)}{2\alpha} \right)^2$$

$$= \frac{1}{4\alpha^2} E_0 \left( R(X, Y) + \frac{S_{L^*}(X, Y)}{2\alpha} + \bar{O}\left(\frac{1}{\alpha^2}\right) \right)^2$$

$$= \frac{1}{4\alpha^2} \left( E_0 R(X, Y)^2 + \frac{1}{\alpha} E_0[R(X, Y)S_{L^*}(X, Y)] + O\left(\frac{1}{\alpha^2}\right) \right).$$
Hence setting $E_0(L^* - 1)^2 = \eta$ gives

$$\frac{1}{2\alpha} = \frac{\sqrt{\eta}}{s_0(R(X,Y))} \left( 1 + \frac{1}{\alpha} \frac{E_0[R(X,Y)S^{L^*}(X,Y)]}{Var_0(R(X,Y))} + O \left( \frac{1}{\alpha^2} \right) \right)^{-1/2}$$

since $E_0R(X,Y) = 0$

$$= \frac{\sqrt{\eta}}{s_0(R(X,Y))} \left( 1 - \frac{1}{2\alpha} \frac{E_0[R(X,Y)S^{L^*}(X,Y)]}{Var_0(R(X,Y))} + O \left( \frac{1}{\alpha^2} \right) \right)$$

$$= \frac{\sqrt{\eta}}{s_0(R(X,Y))} \left( 1 - \frac{1}{2\alpha} \frac{E_0[R(X,Y)S^{L^*}(X,Y)]}{Var_0(R(X,Y))} + O \left( \frac{1}{\alpha^2} \right) \right)$$

by recursing the expression. Note that the asymptotic is valid since we assume $s_0(R(X,Y)) > 0$.

This also shows that given any small enough $\eta > 0$, we can find $\alpha$ and $L$ that satisfies $E_0(L^* - 1)^2 = \eta$, which allows us to invoke Theorem 2.

Now consider the objective value of (18):

$$E_0[hL^*] = E_0 \left[ h \prod_{t=2}^{T} \left( 1 + \frac{R^{L^*}(X_{t-1}, X_t)}{2\alpha} \right) \right]$$

$$= E_0h + \frac{1}{2\alpha} \sum_{t=2}^{T} E_0[hR^{L^*}(X_{t-1}, X_t)] + \frac{1}{4\alpha^2} \sum_{s,t=2,\ldots,T} E_0[hR^{L^*}(X_{t-1}, X_t)R^{L^*}(X_{s-1}, X_s)]$$

$$+ O \left( \frac{1}{\alpha^3} \right).$$

Using (65), we can write (67) as

$$E_0h + \frac{1}{2\alpha} \sum_{t=2}^{T} E_0[hR(X_{t-1}, X_t)] + \frac{1}{4\alpha^2} \sum_{t=2}^{T} E_0[hS^{L^*}(X_{t-1}, X_t)]$$

$$+ \frac{1}{4\alpha^2} \sum_{s,t=2,\ldots,T} E_0[hR^{L^*}(X_{t-1}, X_t)R^{L^*}(X_{s-1}, X_s)] + O \left( \frac{1}{\alpha^3} \right)$$

$$= E_0h + \frac{1}{2\alpha} E_0[H_2(X,Y)R(X,Y)] + \frac{1}{4\alpha} \left( E_0[H_2(X,Y)S^{L^*}] \right.$$}

$$+ \sum_{s,t=2,\ldots,T} E_0[hR(X_{s-1}, X_s)R(X_{t-1}, X_t)] \right) + O \left( \frac{1}{\alpha^3} \right)$$

by conditioning on $(X_{t-1}, X_t)$ inside some of the expectations to get the equality. Note that $E_0[H_2(X,Y)R(X,Y)] = Var_0(R(X,Y))$, by the property that $R$ is an orthogonal projection of $H$. 

41
Then, by substituting (66), we have (68) equal to

\[ E_0 h + s \sigma_0 (R(X,Y)) \sqrt{\eta} - \frac{E_0[R(X,Y)S^{L*}(X,Y)]\eta}{Var_0(R(X,Y))} + \frac{\eta}{Var_0(R(X,Y))} \left( E_0[H(X,Y)S^{L*}(X,Y)] + \sum_{s,t=2}^{T} E_0[hR(X_{s-1},X_s)R(X_{t-1},X_t)] \right) + O(\eta^{3/2}) \]

\[ = E_0 h + s \sigma_0 (R(X,Y)) \sqrt{\eta} + \frac{\eta}{Var_0(R(X,Y))} \left( E_0[H_2(X,Y)S^{L*}(X,Y)] - E_0[R(X,Y)S^{L*}(X,Y)] \right) \]

\[ + \sum_{s,t=2}^{T} E_0[hR(X_{s-1},X_s)R(X_{t-1},X_t)] \right) + O(\eta^{3/2}) \]  

(69)

Now note that \( H - R \) is orthogonal to \( S^{L*} \) by construction. This concludes that (69) is equal to

\[ E_0 h + s \sigma_0 (R(X,Y)) \sqrt{\eta} + \frac{\eta}{Var_0(R(X,Y))} \sum_{s,t=2}^{T} E_0[hR(X_{s-1},X_s)R(X_{t-1},X_t)] + O(\eta^{3/2}) \]

which gives the theorem.

References


B Proofs of Technical Results

B.1 Equivalence of Formulations

Detailed justification of the equivalence between (5) and (8). We show that, for any feasible solution in the max program in (5), one can find a feasible solution in (8) that attains the same objective value, and vice versa. First, given any $P_f$ that has the same marginal distributions as $P_0$, we can let $L(X,Y) = dP_f(X,Y)/dP_0(X,Y) = dP_f(X,Y)/dP_f(X)dP_f(Y)$. Then $\phi^2(P_f) \leq \eta$ is equivalent to $E_0[(L-1)^2] \leq \eta$. Also, as discussed in Section 3, $P_f(x) = P_0(x)$ implies $E_0[L|X] = 1$ a.s. since $P_f(X \in A) = E_0[E_0[L|X]I(X \in A)]$, and similarly $P_f(y) = P_0(y)$ implies $E_0[L|Y] = 1$ a.s.

On the other hand, given any $L$ in the feasible set in (8), we define $P_f(x,y) = L(x,y)dP_0(x)dP_0(y)$. The constraint $E_0[(L-1)^2] \leq \eta$ implies immediately that $\phi^2(P_f) \leq \eta$. Moreover,

$$P_f(X \in A) = E_0[L(X,Y)I(X \in A)] = E_0[E_0[L|X]I(X \in A)] = P_0(X \in A)$$
since \(E_0[L|X] = 1\) a.s.. Similarly, \(P_f(Y \in B) = P_0(Y \in B)\) for any measurable \(B\).

Hence the two formulations are equivalent. \(\square\)

**Proof of Proposition 3.** We prove in both directions. First, we show that each feasible solution in (11) can be re-expressed as a feasible solution in (18) that has the same objective value. Given \(P_f\) that satisfies \(P_f(x_t) = P_0(x_t)\), we can let \(L_t(x_{t-1}, x_t) = dP_f(x_t|x_{t-1})/dP_0(x_t) = dP_f(x_{t-1}, x_t)/dP_0(x_{t-1})P_f(x_t)\) be the likelihood ratio of a transition step under \(P_f\) w.r.t. the benchmark. Since \(P_f(x_t) = P_0(x_t)\), we have \(L_t(x_{t-1}, x_t) = dP_f(x_{t-1}, x_t)/dP_0(x_{t-1})P_0(x_t) = dP_f(x_{t-1}, x_t)/dP_f(x_{t-1})P_f(x_t)\). Also, since \(P_f(x_{t-1}, x_t) = P_f(x_{s-1}, x_s)\) for all \(s, t\), this means that \(L_t(\cdot, \cdot)\) are all the same, so we can let \(L\) be the identical likelihood ratio. Then

\[
E_f[h(X_T)] = \int h(x_T)dP_f(x_1)dP_f(x_2|x_1)\cdots dP_f(x_T|x_{T-1})
\]

\[
= \int h(x_T)dP_f(x_1)\frac{dP_f(x_2|x_1)}{dP_0(x_1)}\cdots \frac{dP_f(x_T|x_{T-1})}{dP_0(x_T)}dP_0(x_1)\cdots dP_0(x_T)
\]

\[
= \int h(x_T)L(x_1, x_2)\cdots L(x_{T-1}, x_T)dP_0(x_1)\cdots dP_0(x_T)
\]

\[
= E_0\left[h(X_T) \prod_{t=2}^{T} L(X_{t-1}, X_t) \right].
\]

Hence the objectives are the same. To show that \(L\) is feasible, first note that \(\phi^2(P_f(X, Y)) = E(L(X, Y) - 1)^2\) by the definition of \(L\), so \(\phi^2(P_f(X, Y)) \leq \eta\) implies \(E(L(X, Y) - 1)^2 \leq \eta\). Now, much like the justification of (8), since

\[
P_f(X_{t-1} \in A) = E_0[E_0[L(X_{t-1}, X_t)|X_{t-1}]I(X_{t-1} \in A)] = P_0(X_t \in A)
\]

for any measurable \(A\), by the constraint that \(P_f(X_{t-1}) = P_0(X_{t-1})\), we must have \(E_0[L(X_{t-1}, X_t)|X_{t-1}] = 1\). Similarly, \(E_0[L(X_{t-1}, X_t)|X_t] = 1\). Hence \(L\) satisfies all the constraints in (18).

To prove the other direction, given any \(L\) in the feasible set of (18), define the Markov transition kernel \(dP_f(x_t|x_{t-1}) = L(x_{t-1}, x_t)dP_0(x_t)\) for any \(t = 2, \ldots, T\), and \(dP_f(x_1) = dP_0(x_1)\). Note that \(dP_f(x_t|x_{t-1})\) is well-defined since

\[
\int dP_f(x_t|x_{t-1}) = \int L(x_{t-1}, x_t)dP_0(x_t) = E_0[L(x_{t-1}, x_t)|x_{t-1}] = 1
\]
a.e., by the first marginal constraint in (18). Now

\[
E_0\left[h(X_T) \prod_{t=2}^{T} L(X_{t-1}, X_t) \right] = \int h(x_T)dP_0(x_1)\prod_{t=2}^{T} L(x_{t-1}, x_t)dP_0(x_t)
\]

\[
= \int h(x_T)dP_f(x_1)\prod_{t=2}^{T} dP_f(x_t|x_{t-1})
\]

\[
= E_f[h(X_T)].
\]
Hence the objective is matched. Next we prove $P_f(x_t) = P_0(x_t)$ by induction. Note that $P_f(X_1 \in A) = P_0(X_1 \in A)$ for any measurable $A$ by definition. Assume $P_f(x_{t-1}) = P_0(x_{t-1})$. Then

$$P_f(X_t \in A) = \int_{x_t \in A} dP_f(x_t|x_{t-1})dP_f(x_{t-1})$$

$$= \int_{x_t \in A} L(x_{t-1}, x_t)dP_0(x_t)dP_0(x_{t-1})$$

$$= \int_{x_t \in A} E_0[L(X_{t-1}, X_t)|X_t = x_t]dP_0(x_t)$$

$$= P_0(X_t \in A)$$

for any measurable $A$, by the second marginal constraint in (18). Hence $P_f(x_t) = P_0(x_t)$, and we conclude our induction.

Finally, note that $dP_f(X_t|X_{t-1}) = L(X_{t-1}, X_t)dP_0(X_t)$ implies $L = dP_f(X_t|X_{t-1})/dP_0(X_t) = dP_f(X_{t-1}, X_t)/dP_0(X_{t-1})P_0(X_t)$. Hence $E_0(L(X,Y) - 1)^2 \leq \eta$ implies $\phi^2(P_f(X,Y)) \leq \eta$.

\[ \square \]

### B.2 Proofs Related to the Contraction and Fixed Point Machinery

We first point out some convenient properties of $L$, which defines a change of measure sequentially from the i.i.d. samples in $P_0$ to the Markov chain under $P_f$, over the time steps $t = 1, 2, \ldots, T$. Note that this definition is two-way, i.e. a Markov chain can be constructed both forward and backward in time under stationarity. Moreover, the chain can be started from any time point conditioned on stationarity. Let us now define $L_{s:t} = \prod_{k=s}^{t} L(X_{k-1}, X_k)$ for any $2 \leq s \leq t \leq T$. Also, for any $1 \leq s \leq t \leq T$, define $F_{s:t} = F(X_s, \ldots, X_t)$ as the $\sigma$-algebra generated by $\{X_s, \ldots, X_t\}$. The following two properties of $L_{s:t}$ are immediate:

**Lemma 4.**

1. Under $P_0$, and fix the starting point of time $2 \leq s \leq T$ under stationarity. The process $\{L_{s:t} : t = s-1, s, s+1, \ldots\}$ (where we define $L_{s:s-1} = 1$) is a martingale adapted to $\{F_{(s-1):t} : t = s-1, s, s+1, \ldots\}$, with initial value identical to 1. Similar property holds for the time-reverse setting. Namely, under $P_0$, and fix the starting time $2 \leq t \leq T$. The process $\{L_{s:t} : s = t+1, t, t-1, \ldots\}$ (where we now define $L_{(t+1):t} = 1$) is a martingale adapted to $\{F_{(s-1):t} : s = t+1, t, t-1, \ldots\}$ with initial value 1.

2. For any measurable $g(X_{s:t}) := g(X_s, \ldots, X_t)$, for $s \leq t$, we have

$$E_f[g(X_{s:t})] = E_0[g(X_{s:t})L_{2:t}] = E_0[g(X_{s:t})L_{(s+1):T}]$$

Both properties can be generalized to the scenario when the factors $L$ in the likelihood ratio product $L_{s:t}$ are non-identical.
Proof of Lemma 4. To prove the first part, consider, for any \( t \geq s \),

\[
E_0[L_{s(t+1)}|\mathcal{F}_{(s-1):t}] = E_0[L(X_{s-1}, X_s) \cdots L(X_t, X_{t+1})|\mathcal{F}_{(s-1):t}]
\]

\[
= L(X_{s-1}, X_s) \cdots L(X_{t-1}, X_t) E_0[L(X_t, X_{t+1})|\mathcal{F}_{(s-1):t}]
\]

\[
= L(X_{s-1}, X_s) \cdots L(X_{t-1}, X_t)
\]

since \( E_0[L(X_t, X_{t+1})|\mathcal{F}_{(s-1):t}] = E_0[L(X_t, X_{t+1})|X_t] = 1 \)

by the first marginal constraint in (18)

\[
= L_{s:t}
\]

The backward direction follows analogously, by using the second marginal constraint in (18). The second part of the lemma follows immediately from the forward and backward martingale properties of \( L_{s:t} \). Also, the generalization to non-identical \( L \)'s follows trivially. \( \square \)

Proof of Lemma 1. We divide the proofs into three components: first, we show that \( \mathcal{K} \) is well-defined and closed; second, we prove that \( \mathcal{K} \) is a contraction map; lastly, we show that the fixed point of \( \mathcal{K} \) has all identical components.

We start with showing well-definedness. Note that each summand in the definition of \( H L^{(1):(T-2)}(x, y) \) satisfies

\[
|E_0[hL^{(T-t+1)}(X_1, X_2)L^{(T-t-2)}(X_2, X_3) \cdots L^{(T)}(X_{t-2}, X_{t-1})L^{(1)}(X_t, X_{t+1})L^{(2)}(X_{t+1}, X_{t+2})
\]

\[
\cdots L^{(T-t)}(X_{T-1}, X_T)|X_{t-1} = x, X_t = y|
\]

\[
\leq C E_0[L^{(T-t+1)}(X_1, X_2)L^{(T-t-2)}(X_2, X_3) \cdots L^{(T)}(X_{t-2}, X_{t-1})L^{(1)}(X_t, X_{t+1})L^{(2)}(X_{t+1}, X_{t+2})
\]

\[
\cdots L^{(T-t)}(X_{T-1}, X_T)|X_{t-1} = x, X_t = y \] for some \( C > 0 \)

\[
= C E_0[L^{(1)}(X_1, X_{t+1})L^{(2)}(X_{t+1}, X_{t+2}) \cdots L^{(T-t)}(X_{T-1}, X_T)|X_t = y]
\]

\[
= C \] by the forward and backward martingale properties of Lemma 4.

Hence \( |H L^{(1):(T-2)}(x, y)| \leq (T - 1)C \), and the stepwise operator \( \mathcal{K} \) satisfies

\[
|\mathcal{K}(L^{(1):(T-2)}) - 1| \leq \frac{4(T - 1)C}{2\alpha} < \infty.
\]

The same bound holds for each \( \mathcal{K}(\tilde{L}^{(1)}, \tilde{L}^{(2)}, \ldots, \tilde{L}^{(t)}, L^{(t+1)}, \ldots, L^{(T-2)}) \) for \( t = 1, \ldots, T - 3 \). Hence each component of \( \mathcal{K}(L^{(1):(T-2)}) \) is finite a.s.. Moreover, for large enough \( \alpha \), we have

\[
|H L^{(1):(T-2)}(x, y) - E_0[H L^{(1):(T-2)}|x] - E_0[H L^{(1):(T-2)}|y] + E_0 H L^{(1):(T-2)}| \leq 4(T - 1)C \leq 2\alpha
\]

for any \( L^{(1):(T-2)} \in \mathcal{L}^c(M)^{T-2} \) and so each component of \( \mathcal{K}(L^{(1):(T-2)}) \) is non-negative a.s.. Note that by definition \( \mathcal{K}(L^{(1):(T-2)}) \) satisfies \( E_0[\mathcal{K}(L^{(1):(T-2)})|x] = E_0[\mathcal{K}(L^{(1):(T-2)})|y] = 1 \) a.s., and similarly along the iteration of \( \mathcal{K} \) in (29).
To conclude that $\mathcal{K}$ is well-defined, we are left to show that $\mathcal{K}$ preserves the boundedness in the $L_2$-norm. Since

$$E_0 K(L^{(1):(T-2)})^2 - 1 = E_0 (K(L^{(1):(T-2)}) - 1)^2$$

$$\leq \frac{1}{4\alpha^2} E_0 \left( H L^{(1):(T-2)}(X, Y) - E_0 [H L^{(1):(T-2)} | X] - E_0 [H L^{(1):(T-2)} | Y] + E_0 H L^{(1):(T-2)} \right)^2$$

$$\leq \frac{(4(T-1)C)^2}{4\alpha^2} \leq M - 1$$

for any $L^{(1):(T-2)} \in \mathcal{L}^c(M)^{T-2}$, when $\alpha$ is large enough, we have $K(L^{(1):(T-2)}) \in \mathcal{L}^c(M)$. Iterating through $K(\tilde{L}^{(1)}, \tilde{L}^{(2)}, \ldots, \tilde{L}^{(t)}, L^{(t+1)}, \ldots, L^{(T-2)})$ for $t = 1, \ldots, T - 3$, we conclude that $\mathcal{K}$ is closed in $\mathcal{L}^c(M)^{T-2}$.

Now we show that $\mathcal{K}$ is a contraction. Consider, for any $L^{(1):(T-2)} = (L^{(1)}, \ldots, L^{(T-2)}) \in \mathcal{L}^c(M)^{T-2}$ and $L^{(1):(T-2)'} = (L^{(1)'}, \ldots, L^{(T-2)'}) \in \mathcal{L}^c(M)^{T-2}$,

$$E_0 |K(L^{(1):(T-2)}) - K(L^{(1):(T-2)'})|^2$$

$$\leq E_0 |H L^{(1):(T-2)} - H L^{(1):(T-2)'}|^2 \text{ since projection is a contraction}$$

$$\leq \frac{C}{4\alpha^2} E_0 \left[ \left( \sum_{t=2}^{T} E_0 [L^t - L'^t | X_{t-1}, X_t] \right)^2 \right]$$

$$\leq \frac{(T-1)^2 C}{4\alpha^2} \sum_{t=2}^{T} E_0 \left( E_0 [L^t - L'^t | X_{t-1}, X_t] \right)^2 \tag{70}$$

for some constant $C > 0$, where we denote

$$L^t = L^{(T-t+1)}(X_1, X_2) L^{(T-t+2)}(X_2, X_3) \cdots L^{(T)}(X_{t-2}, X_{t-1}) L^{(1)}(X_t, X_{t+1}) L^{(2)}(X_{t+1}, X_{t+2}) \cdots L^{(T-t)}(X_{T-1}, X_T)$$

and

$$L'^t = L^{(T-t+1)'}(X_1, X_2) L^{(T-t+2)'}(X_2, X_3) \cdots L^{(T)'}(X_{t-2}, X_{t-1}) L^{(1)'}(X_t, X_{t+1}) L^{(2)'}(X_{t+1}, X_{t+2}) \cdots L^{(T-t)'}(X_{T-1}, X_T)$$

for convenience.
Let us focus on \( E_0 \left( E_0[|L^t - L^t'|] |X_{t-1}, X_t| \right)^2 \). Note that, by telescoping, we have

\[
L^t - L^t' = \left( L^{(T-t+1)}(X_1, X_2) - L^{(T-t+1)'}(X_1, X_2) \right) L^{(T-t+2)}(X_2, X_3) \cdots L^{(T)}(X_{t-2}, X_{t-1}) L^{(1)}(X_t, X_{t+1}) \\
L^{(2)}(X_{t+1}, X_{t+2}) \cdots L^{(T-t)}(X_{T-1}, X_T) \\
+ L^{(T-t+1)'}(X_1, X_2) \left( L^{(T-t+2)}(X_2, X_3) \cdots L^{(T)}(X_{t-2}, X_{t-1}) L^{(1)}(X_t, X_{t+1}) L^{(2)}(X_{t+1}, X_{t+2}) \right) \\
\cdots \right) \\
- L^{(T-t+1)'}(X_1, X_2) \cdots L^{(T)}(X_{t-2}, X_{t-1}) L^{(1)'}(X_t, X_{t+1}) L^{(2)'}(X_{t+1}, X_{t+2}) \cdots L^{(T-t)'}(X_{T-1}, X_T) \\
\vdots \\
= \sum_{s=2}^{T} \prod_{k=2, k \neq t}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left( L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right) \prod_{k=s+1}^{T} L^{(k-t)}(X_{k-1}, X_k)
\]

where we denote \( L^{(k)} \) for a negative \( k \) to be \( L^{(k+T-1)} \). Hence

\[
E_0[|L^t - L^t'|] |X_{t-1}, X_t| \leq \sum_{s=2, \ldots, T} \left[ \prod_{k=2, k \neq t}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| \prod_{k=s+1}^{T} L^{(k-t)}(X_{k-1}, X_k) \right] \\
\left| X_{t-1}, X_t \right|
\]

and so

\[
E_0 \left( E_0[|L^t - L^t'|] |X_{t-1}, X_t| \right)^2 \leq (T-2)^2 \sum_{s=2, \ldots, T} E_0 \left[ \prod_{k=2, k \neq t}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| \prod_{k=s+1}^{T} L^{(k-t)}(X_{k-1}, X_k) \right]^2.
\]

Let us now consider each summand above. Without loss of generality, consider the case \( s < t \). We
have
\[
E_0 \left[ \prod_{k=2}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| \prod_{k=s+1}^{t} L^{(k-t)}(X_k) \right] = E_0 \left[ \prod_{k=2}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| \prod_{k=s+1}^{t-1} L^{(k-t)}(X_k) \right] = E_0 \left[ \prod_{k=2}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| \prod_{k=s+1}^{t-1} L^{(k-t)}(X_k) \right]
\]
where we denote \( \prod_{k=s+1}^{t-1} L^{(k-t)}(X_k) = 1 \) if \( s = t - 1 \)
\[
= E_0 \left[ E_0 \left[ \prod_{k=2}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| \prod_{k=s+1}^{t-1} L^{(k-t)}(X_k) \right] \right]
\]
\[
= E_0 \left[ E_0 \left[ \prod_{k=2}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| \prod_{k=s+1}^{t-1} L^{(k-t)}(X_k) \right] \right]
\]
\[
= E_0 \left[ \prod_{k=2}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| \prod_{k=s+1}^{t-1} L^{(k-t)}(X_k) \right]
\]
(72)

where \( E_f \) is under the change of measure generated by \( \prod_{k=s+1}^{t-1} L^{(k-t)}(X_k) \). On the other hand, we have
\[
E_0 \left[ \prod_{k=2}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| \right] = E_0 \left[ \prod_{k=2}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| \right]
\]
\[
= E_0 \left[ \prod_{k=2}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| \right]
\]
(73)
by the backward martingale property of \( \prod_{k=2}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \). So, by (72), (73) and Jensen’s inequality, we have

\[
E_0 \left( E_0 \left[ \prod_{k=2}^{s-1} L^{(k-t)'}(X_{k-1}, X_k) \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| \right] \right)^2 \\
= E_0 \left( E_\mathcal{P} \left[ E_0 \left[ \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right| X_s \right] \right] \right)^2 \\
\leq E_0 \left[ \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right|^2 \right] 
\]

(74)

where \( E_\mathcal{P} \) denotes the expectation under the change of measure induced by \( \prod_{k=s+1}^{t-1} L^{(k-t)}(X_{k-1}, X_k) \) from step \( t-1 \) backward to \( s \) and then following the benchmark \( P_0 \) before \( s \). The last inequality follows from Jensen’s inequality, and by the construction that \( X_s \) is marginally distributed as \( P_0 \) and that the transition from \( X_s \) and \( X_{s-1} \) is under the benchmark \( P_0 \).

Therefore, (71) is bounded by

\[
(T - 2)^2 \sum_{s=2, \ldots, T \atop s \neq t} E_0 \left[ \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right|^2 \right],
\]

and hence (70) is bounded by

\[
\frac{(T - 1)^2(T - 2)^2}{4\alpha^2} \sum_{t=2}^{T} \sum_{s=2, \ldots, t \atop s \neq t} E_0 \left[ \left| L^{(s-t)}(X_{s-1}, X_s) - L^{(s-t)'}(X_{s-1}, X_s) \right|^2 \right] \\
= \frac{(T - 1)^3(T - 2)^2}{4\alpha^2} \sum_{t=1}^{T-2} E_0 \left[ \left| L^{(t)} - L^{(t)'} \right|^2 \right] \\
\leq \frac{(T - 1)^3(T - 2)^3}{4\alpha^2} d(\mathcal{L}^{(1)}; (T-2), \mathcal{L}^{(1)}; (T-2)').^2.
\]

When \( \alpha \) is large enough, this gives \( E_0 \left[ K(\mathcal{L}^{(1)}; (T-2)) - K(\mathcal{L}^{(1)}; (T-2)') \right]^2 \leq d(\mathcal{L}^{(1)}; (T-2), \mathcal{L}^{(1)}; (T-2)').^2 \), and the computation above can be iterated over \( (\mathcal{L}^{(1)}, \mathcal{L}^{(2)}, \ldots, \mathcal{L}^{(t)}), \mathcal{L}^{(t+1)}, \ldots, \mathcal{L}^{(T-2)} \) for \( t = 1, \ldots, T - 3 \). Then we shall get that

\[
\frac{(T - 1)^3(T - 2)^3}{4\alpha^2} d(\mathcal{K}^{(1)}; (T-2), \mathcal{L}^{(1)}; (T-2)').^2 \leq \frac{(T - 1)^3(T - 2)^3}{4\alpha^2} d(\mathcal{L}^{(1)}; (T-2), \mathcal{L}^{(1)}; (T-2)').^2
\]
which gives
\[
\begin{align*}
d(K(L^{(1)};T-2), K(L^{(1)};T-2')) & \leq \frac{(T-1)^{3/2}(T-2)^{3/2}C^{1/2}}{2\alpha} d(L^{(1)};T-2, L^{(1)};T-2') \\
& \leq cd(L^{(1)};T-2, L^{(1)};T-2')
\end{align*}
\]
for some constant \(0 < c < 1\), when \(\alpha\) is large enough. Hence we conclude that \(K\) is a contraction on \(L^c(M)^{T-2}\).

By Banach fixed point theorem, \(K\) has a fixed point \(L^*\). We are left to show that all components of \(L^*\) are identical. To this end, let \(L^* = (L^{(1)*}, L^{(2)*}, \ldots, L^{(T-2)*})\) and note that by the definition of fixed point and the iteration in (29),
\[
\begin{align*}
\tilde{L}^{(1)*} & := K(L^{(1)*}, L^{(2)*}, \ldots, L^{(T-2)*}) = L^{(1)*} \\
\tilde{L}^{(2)*} & := K(\tilde{L}^{(1)*}, L^{(2)*}, \ldots, L^{(T-2)}) = K(L^{(1)*}, L^{(2)*}, \ldots, L^{(T-2)*}) = L^{(2)*} \\
\tilde{L}^{(3)*} & := K(\tilde{L}^{(1)*}, \tilde{L}^{(2)*}, L^{(3)*}, \ldots, L^{(T-2)}) = K(L^{(1)*}, L^{(2)*}, L^{(3)*}, \ldots, L^{(T-2)*}) = L^{(3)*} \\
& \vdots \\
\tilde{L}^{(T-2)*} & := K(\tilde{L}^{(1)*}, \tilde{L}^{(2)*}, \ldots, \tilde{L}^{(T-3)*}, L^{(T-2)}) = K(L^{(1)*}, L^{(2)*}, \ldots, L^{(T-3)*}, L^{(T-2)*}) = L^{(T-2)*}
\end{align*}
\]
and hence
\[
L^{(1)*} = L^{(2)*} = \ldots = L^{(T-2)*} = K(L^{(1)*}, L^{(2)*}, \ldots, L^{(T-2)*}).
\]
We thus have concluded the lemma. \(\square\)

**Proof of Lemma 3.** We first note that, as discussed in Section 4.2, convergence to the fixed point associated with the operator \(K\) under the \(d\)-metric implies componentwise convergence, i.e. the sequence \(L^{(k)} = K(L^{(k-T+2)}, L^{(k-T+3)}, \ldots, L^{(k-1)})\) for \(k = T-1, T-2, \ldots\) defined in Lemma 2 converges to \(L^*\), the identical component of the fixed point \(L^*\) of \(K\).

Now consider each term in (32). For the first term,
\[
\tilde{H}L^{(k)(k+T-2)} - \tilde{H}L^* \leq C \sum_{t=2}^{T} E_0 |L^{k,t} - L^*| \tag{75}
\]
for some constant \(C > 0\), where we denote
\[
L^{k,t} = L^{(k+T-t+2)}(X_1, X_2) L^{(k+T-t+3)}(X_2, X_3) \cdots L^{(k+T-1)}(X_{t-2}, X_{t-1}) L^{(k+1)}(X_{t-1}, X_t) L^{(k+2)}(X_t, X_{t+1}) \cdots L^{(k+T-t+1)}(X_{T-1}, X_T)
\]
and
\[
L^* = \prod_{t=2}^{T} L^*(X_{t-1}, X_t).
\]

52
Then, by the same technique in the proof of contraction in Lemma 1, we have (75) bounded by

\[ C \sum_{t=2}^{T} \sum_{s=2}^{T} E_0 |L^{(k+s-t+1)}(X_{s-1}, X_s) - L^*(X_{s-1}, X_s)| \]

where we denote \( L^{(k+j)} = L^{(k+j+T-1)} \) for any negative \( j \), which is then equal to

\[ C(T-1) \sum_{t=2}^{T} E_0 |L^{(k+t-2)} - L^*| \to 0 \]

as \( k \to \infty \). For the second term in (32), since \( \|L^{(k)} - L^*\|_2 \to 0 \), we have immediately that \( \|L^{(k)}\|_2 \to \|L^*\|_2 \) and so

\[ \sum_{t=2}^{T} E_0 (L^{(k+t-2)} - 1)^2 - (T-1)E_0(L^* - 1)^2 \to 0 \]

as \( k \to \infty \). Hence we have proved the lemma. \( \square \)

**B.3 Proofs Related to Higher Order Dependency Assessment**

*Missing component of proof for Theorem 3.* As discussed in Section 5, in order to show Theorem 3, we have to argue that \( S(X,Y,Z) \) is the projection of \( \tilde{H}_{0,2}(X,Y,Z) \) onto \( \tilde{M}_0 \) and that \( sd_{j_1}(S_{j_1}(X,Y,Z)) \to sd_0(S(X,Y,Z)) \), where \( X, Y \) and \( Z \) are three consecutive states under the respective measure. We lay out our argument for each of these below.

\( S(X,Y,Z) \) is a projection. We first check that \( S(X,Y,Z) \in \tilde{M}_0 \). Note that from (42) we can write

\[ S(x,y,z) = \tilde{H}_{0,2}(x,y,z) - E_0[\tilde{H}_{0,2}|x,y] - E_0[\tilde{H}_{0,2}|y,z] + E_0[\tilde{H}_{0,2}|y]. \]

So

\[ E_0[S(X,Y,Z)|X,Y] = E_0[\tilde{H}_{0,2}|X,Y] - E_0[\tilde{H}_{0,2}|X,Y] - E_0[\tilde{H}_{0,2}|Y] + E_0[\tilde{H}_{0,2}|Y] = 0 \]

since \( X \) and \( Z \) are independent. Similarly, \( E_0[S(X,Y,Z)|Y,Z] = 0 \), and so \( S(X,Y,Z) \in \tilde{M}_0 \).
Next we show that \( S \) and \( \tilde{H}_{0,2} - S \) are orthogonal. Consider

\[
E_0[S(\tilde{H}_{0,2} - S)] \\
= E_0(\tilde{H}_{0,2}(X,Y,Z) - E_0[\tilde{H}_{0,2}|X,Y] - E_0[\tilde{H}_{0,2}|Y,Z] + E_0[\tilde{H}_{0,2}|Y]) \\
(E_0[\tilde{H}_{0,2}|X,Y] + E_0[\tilde{H}_{0,2}|Y,Z] - E_0[\tilde{H}_{0,2}|Y]) \\
= E_0(E_0[\tilde{H}_{0,2}|X,Y])^2 + E_0(E_0[\tilde{H}_{0,2}|Y,Z])^2 - E_0(E_0[\tilde{H}_{0,2}|Y])^2 \\
- E_0(E_0[\tilde{H}_{0,2}|X,Y] + E_0[\tilde{H}_{0,2}|Y,Z] - E_0[\tilde{H}_{0,2}|Y])^2 \\
= -2E_0[\tilde{H}_{0,2}|X,Y]E_0[\tilde{H}_{0,2}|Y,Z] + 2E_0[\tilde{H}_{0,2}|X,Y]E_0[\tilde{H}_{0,2}|Y] \\
+ 2E_0[\tilde{H}_{0,2}|Y,Z]E_0[\tilde{H}_{0,2}|Y] - 2E_0(E_0[\tilde{H}_{0,2}|Y])^2 \\
= -2E_0(\tilde{H}_{0,2}|Y)^2 + 2E_0(E_0[\tilde{H}_{0,2}|Y])^2 - E_0(E_0[\tilde{H}_{0,2}|Y])^2 \\
= 0
\]

where the second-to-last step is via conditioning on \( Y \), and by using the fact that given \( Y \), \( E_0[\tilde{H}_{0,2}|X,Y] \) is independent of \( E_0[\tilde{H}_{0,2}|Y,Z] \). Therefore \( S(X,Y,Z) \) is the projection of \( \tilde{H}_{0,2} \) onto \( \tilde{\mathcal{M}}_0 \).

The convergence \( sd_f \( S f \( X,Y,Z \) \) \rightarrow sd_0(S(X,Y,Z)) \). To facilitate the proof, denote \( P_{f_1} \) as the projection operator onto \( \mathcal{M}_{f_1} \) in the space \( L_2(P_{f_1}) \), and similarly denote \( P_0 \) as the projection onto \( \tilde{\mathcal{M}}_0 \) in the space \( L_2(P_0) \). Note that since \( E_{f_1}S_{f_1} = 0 \), \( sd_{f_1}(S_{f_1}) \) is exactly \( \|S_{f_1}\|_{f_1} \), where we denote \( \| \cdot \|_{f_1} \) as the \( L_2(P_{f_1}) \)-norm. In a similar fashion, \( sd_0(S) = \|S\|_0 \), where we denote \( \| \cdot \|_0 \) as the \( L_2(P_0) \)-norm. With these notations, we are set to prove that \( \|P_{f_1} \tilde{H}_{f_1,2}\|_{f_1} \rightarrow \|P_0 \tilde{H}_{0,2}\|_0 \) as \( \eta_1 \rightarrow 0 \).

Next, we observe that by the boundedness of \( h \) and that convergence in \( \chi^2 \)-distance implies convergence in distribution, we have \( \tilde{H}_{f_1,2}(x,y,z) \rightarrow \tilde{H}_{0,2}(x,y,z) \) pointwise a.e. as \( \eta_1 \rightarrow 0 \). Hence by dominated convergence \( \|\tilde{H}_{f_1,2} - \tilde{H}_{0,2}\|_0 \rightarrow 0 \), where \( \tilde{H}_{f_1,2} \) and \( \tilde{H}_{0,2} \) are coupled in the natural way under the measure \( P_0 \). Moreover, we have \( \|G\|_{f_1} \rightarrow \|G\|_0 \) for any bounded \( G = G(X,Y) \).

Now consider

\[
\|P_{f_1} \tilde{H}_{f_1,2}\|_{f_1} \rightarrow \|P_0 \tilde{H}_{0,2}\|_0 = (\|P_{f_1} \tilde{H}_{f_1,2}\|_{f_1} - \|P_{f_1} \tilde{H}_{f_1,2}\|_{0}) + (\|P_{f_1} \tilde{H}_{f_1,2}\|_{0} - \|P_0 \tilde{H}_{0,2}\|_{0}). \tag{76}
\]

The first term in (76) goes to 0 as \( \eta_1 \rightarrow 0 \) by our observations above. The second term is bounded from above by \( \|P_{f_1} \tilde{H}_{f_1,2} - P_0 \tilde{H}_{0,2}\|_0 \), which in turn is bounded by

\[
\|P_{f_1} \tilde{H}_{f_1,2} - P_0 \tilde{H}_{0,2}\|_0 + \|P_0 \tilde{H}_{f_1,2} - P_0 \tilde{H}_{0,2}\|_0. \tag{77}
\]

The second term in (77) is dominated by \( \|\tilde{H}_{f_1,2} - \tilde{H}_{0,2}\|_0 \) by the contraction property of the projection \( P_0 \), which converges to 0 as \( \eta_1 \rightarrow 0 \) by the observations before. We are left to show that the first term in (77) also goes to 0.
To this end, write
\[
\|\mathcal{P}^0 \tilde{H}_{f_1} - \mathcal{P}^0 \tilde{H}_{f_1}\|_2 = \|\mathcal{P}^0 \tilde{H}_{f_1} - \mathcal{P}^0 \mathcal{P}^0 \tilde{H}_{f_1}\|_2 + \|\mathcal{P}^0 \mathcal{P}^0 \tilde{H}_{f_1}\|_2 = 0. \tag{78}
\]
We tackle the first and the second term in (78) one-by-one. For the first term, using the explicit expression for $\mathcal{P}^0$, we can write
\[
\begin{align*}
\mathcal{P}^0 \tilde{H}_{f_1} - \mathcal{P}^0 \mathcal{P}^0 \tilde{H}_{f_1} &= E_0[\mathcal{P}^0 \tilde{H}_{f_1}|X, Y] + E_0[\mathcal{P}^0 \tilde{H}_{f_1}|Y, Z] - E_0[\mathcal{P}^0 \tilde{H}_{f_1}|Y] \\
&= E_0[\mathcal{P}^0 \tilde{H}_{f_1}|X, Y] + E_0[\mathcal{P}^0 \tilde{H}_{f_1}|Y, Z] - E_0[\mathcal{P}^0 \tilde{H}_{f_1}|Y] + o(1)
\end{align*}
\]
a.s., by the weak convergence from $P_{f_1}$ to $P_0$. But since $\mathcal{P}^0 \tilde{H}_{f_1}|X, Y]$, $E_0[\mathcal{P}^0 \tilde{H}_{f_1}|Y, Z]$, and hence $E_0[\mathcal{P}^0 \tilde{H}_{f_1}|Y]$ are all 0. By dominated convergence we have $\|\mathcal{P}^0 \tilde{H}_{f_1} - \mathcal{P}^0 \mathcal{P}^0 \tilde{H}_{f_1}\|_2 \to 0$.

For the second term, we use the following observation. The conditions $E_{\tilde{f}_1}[V|X, Y] = 0$ and $E_{\tilde{f}_1}[V|Y, Z] = 0$ in the definition of the closed subspace $\hat{M}_{f_1}$ is equivalent to the conditions $\langle V, \nu(X, Y) \rangle_{\tilde{f}_1} = 0$ and $\langle V, \rho(Y, Z) \rangle_{\tilde{f}_1} = 0$ for any measurable functions $\nu$ and $\rho$. Consequently, any elements in $\hat{M}_{f_1}$ can be expressed in the form $\nu(X, Y) + \rho(Y, Z)$. In particular, $\tilde{H}_{f_1} - \mathcal{P}^0 \tilde{H}_{f_1}$ is in this form. But then under the natural coupling $\mathcal{P}^0(\tilde{H}_{f_1} - \mathcal{P}^0 \tilde{H}_{f_1}) = 0$. So the second term in (78) is 0. We thus conclude that $\|\mathcal{P}^0 \tilde{H}_{f_1} - \mathcal{P}^0 \tilde{H}_{f_1}\|_2 \to 0$, and from (76), (77) and (78), we obtain that $\|\mathcal{P}^0 \tilde{H}_{f_1}\|_2 \to \|\mathcal{P}^0 \tilde{H}_{0}\|_2 \to 0$. This proves the theorem. □