This article was downloaded by: [2001:638:807:212:a8af:506c:25a1:b254] On: 09 November 2026, At: 06:25
Publisher: Institute for Operations Research and the Management Sciences (INFORMS)
INFORMS is located in Maryland, USA



Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Risk Analysis for Large Pools of Loans

Justin Sirignano, Kay Giesecke

To cite this article:
Justin Sirignano, Kay Giesecke (2019) Risk Analysis for Large Pools of Loans. Management Science 65(1):107-121.
https://doi.org/10.1287/mnsc.2017.2947

Full terms and conditions of use: https://pubsonline.informs.org/Publications/Librarians-Portal/PubsOnLine-Terms-and-Conditions

This article may be used only for the purposes of research, teaching, and/or private study. Commercial use or systematic downloading (by robots or other automatic processes) is prohibited without explicit Publisher approval, unless otherwise noted. For more information, contact permissions@informs.org.

The Publisher does not warrant or guarantee the article's accuracy, completeness, merchantability, fitness for a particular purpose, or non-infringement. Descriptions of, or references to, products or publications, or inclusion of an advertisement in this article, neither constitutes nor implies a guarantee, endorsement, or support of claims made of that product, publication, or service.

Copyright © 2018, INFORMS

Please scroll down for article—it is on subsequent pages

With 12,500 members from nearly 90 countries, INFORMS is the largest international association of operations research (O.R.) and analytics professionals and students. INFORMS provides unique networking and learning opportunities for individual professionals, and organizations of all types and sizes, to better understand and use O.R. and analytics tools and methods to transform strategic visions and achieve better outcomes.
For more information on INFORMS, its publications, membership, or meetings visit http://www.informs.org

http://pubsonline.informs.org/journal/mnsc/

MANAGEMENT SCIENCE
Vol. 65, No. 1, January 2019, pp. 107–121
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

Risk Analysis for Large Pools of Loans

Justin Sirignano,[a] **Kay Giesecke**[b]

[a] Department of Industrial and Enterprise Systems Engineering, University of Illinois at Urbana–Champaign, Urbana, Illinois 61801;
[b] Department of Management Science and Engineering, Stanford University, Stanford, California 94305

Contact: jasirign@illinois.edu (JS); giesecke@stanford.edu, http://orcid.org/0000-0002-5380-2918 (KG)

Received: February 3, 2016
Revised: October 2, 2016; July 26, 2017
Accepted: August 21, 2017
Published Online in Articles in Advance: April 25, 2018

https://doi.org/10.1287/mnsc.2017.2947

Copyright: © 2018 INFORMS

Abstract. Financial institutions, government-sponsored enterprises, and asset-backed security investors are often exposed to delinquency and prepayment risk from large numbers of loans. Examples include mortgages, credit cards, and auto, student, and business loans. Because of the size of the pools, the measurement of these exposures is computationally expensive. This paper develops and tests efficient numerical methods for the analysis of large pools of loans as well as asset-backed securities backed by such pools. For a broad class of statistical and machine learning models of loan-level risk in discrete time, we prove a law of large numbers and a central limit theorem for the pool-level risk. The asymptotics are then used to construct computationally efficient approximations for a large pool. The approximations aggregate the full loan-level dynamics, making it possible to take advantage of the detailed loan-level data often available in practice. We furthermore prove a convergence rate for the approximation and a uniform integrability result. The latter allows the approximation to be applied to a wide class of pool-level risk statistics. To demonstrate the effectiveness of our approach, we implement it on a data set of over 25 million mortgages. The results show the accuracy and speed of the approximation in comparison to brute-force simulation of a pool, for a variety of pools with different risk profiles.

History: Accepted by Noah Gans, stochastic models.
Funding: This work was supported by the National Science Foundation [Methodology, Measurement, and Statistics Grant 1325031].
Supplemental Material: The online appendix is available at https://doi.org/10.1287/mnsc.2017.2947.

Keywords: law of large numbers • central limit theorem • machine learning • loan default and delinquency • portfolio loss distribution • big data

1. Introduction

Financial institutions, government-sponsored enterprises such as Fannie Mae and Freddie Mac, and investors have credit exposures to large pools of loans, including mortgages, credit card receivables, auto loans, student loans, and business loans. A major U.S. bank might directly own a million mortgages. The government-sponsored enterprises are exposed to tens of millions of mortgages, either through directly owning the mortgages or providing credit guarantees against default for the mortgages. Major loan servicers can service up to 10 million mortgages, and even though they do not directly own the loans, mortgage servicing cash flows are still strongly affected by default and prepayment risk. Mortgage-backed securities (MBSs) typically have hundreds to tens of thousands of loans. Large pools of loans are also common for many other types of loans. For instance, a major U.S. bank might easily have on the order of 20,000 wholesale loans and 100,000 middle-market and commercial loans. A major credit card company can have tens of millions of credit card accounts. Over half of all consumer credit is eventually securitized, and each deal can consist of tens of thousands of credit cards. Because of their size, these loan pools and the securities they back are computationally challenging to analyze. A full simulation of a typical pool might take many hours.

This paper develops and tests efficient numerical methods for the analysis of large, heterogeneous pools of loans. We focus on a broad class of dynamic, discrete-time models of loan-level risk that are widely used in practice (see Lending Club 2015, CoreLogic 2014, Experian 2014, Dunsky et al. 2014, and others). A loan might be in one of several states, such as 30 days past due, prepaid, or foreclosure. The conditional transition probabilities could come from a range of statistical or machine learning formulations such as generalized linear models (e.g., logistic regression), neural networks, decision trees, and support vector machines. The conditional transition probability is allowed to depend on a vector of static loan-level features such as credit score and loan-to-value ratio, a vector of time-varying risk factors influencing multiple loans, such as the unemployment rate, as well as the past behavior of the loans in the pool (through a "mean-field" term). The dependence of the conditional transition probability on common systematic factors as well as

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

the mean-field term address the loan-to-loan correlation observed in practice (see Azizpour et al. 2018, Das et al. 2007).

For this class of models, we prove a law of large numbers and a central limit theorem that describe pool-level risk as the size of a pool grows large. A nonlinear stochastic equation governs the evolution of the law of large numbers limit. The fluctuation limit satisfies a stochastic equation that is linearized around the nonlinear law of large numbers limit. We provide a Monte Carlo algorithm for solving these equations and analyze its convergence as well as the optimal allocation of the computational budget. The limit theorems are used to construct computationally efficient approximations of the loss, prepayment, and cash flow distributions for a large loan pool. A convergence rate for the approximation is proven. We also prove a uniform integrability result, which guarantees the convergence of our approximation for a large class of continuous functions of pool-level quantities such as loss, prepayment, and cash flow. The approximations account for the full loan-level dynamics, making it possible to take advantage of the detailed loan-level data often available for loans in practice. Very importantly, the cost of our approximation remains constant no matter the dimension of the loan-level features. This is essential since loan-level data is often high-dimensional in practice (the dimension could be in the hundreds). Furthermore, since the law of large numbers and central limit theorem are dynamic, the approximation provides the loss, prepayment, and cash flow distributions across all time horizons at no extra computational cost. Given these distributions, risk analysis, pricing, and hedging of asset-backed securities such as MBSs backed by the pool is immediate.

We test the performance of our approximations on a detailed loan-level mortgage data set that includes over 25 million mortgages. We compare the approximate distribution with the "true" distribution (obtained by brute-force simulation) for various pools drawn from this data set. The comparison is performed using model parameters fitted to the data set. The approximation's computational cost is often several orders of magnitude less than the cost of brute-force simulation of the actual pool. It has a similar level of accuracy, in the center as well as the tail of the distribution. Although the approximation's accuracy increases with the size of the pool, it is highly accurate even for a pool having as little as 500 loans.

The remainder of this introduction discusses the related literature. Section 2 describes the class of discrete-time models of loan-level risk we consider. The limit theorems as well as uniform integrability for this class of models are presented in Section 3. These results are used to develop a computationally efficient approximation for large pools. Numerical methods to compute the approximation are described in Section 4. In Section 5, the approximation is tested on actual mortgage data. Section 6 concludes. An online appendix contains the proofs and supplementary results.

1.1. Related Literature

The computational expense associated with loan-by-loan models for large pools is widely recognized. Prior research has analyzed several approaches to tackle this issue.

One strand of the literature has developed top-down models of pool behavior, which are formulated without reference to the constituent loans and only model the behavior of a pool in aggregate. See Fermanian (2013), Wu et al. (2012), Richard and Roll (1989), Kang and Zenios (1992), Mattey and Wallace (2001), and others for top-down models of mortgage pools, and Arnsdorf and Halperin (2008), Giesecke et al. (2011), Ding et al. (2009), Errais et al. (2010), and others for models of corporate credit pools. The approximation that we develop for loan-level models is as tractable as a top-down model while taking full advantage of the loan-level feature data.

Another strand of the literature has developed large-pool approximations in various model frameworks. Several authors consider static one-period models in which loan-to-loan correlation is due to the dependence of loans on a common factor. Here, loans default independently conditional on the common factor, generating tractability. Vasicek (1987) applies a standard law of large numbers, conditionally on the common factor, to obtain an approximation for the distribution of the loss from default in a large pool. Frey et al. (2008) develops an approximation using a conditional central limit theorem. Gordy (2003), Gordy and Lutkebohmert (2013), Martin and Wilde (2002), Emmer and Tasche (2005), Gordy and Marrone (2012), and others develop "granularity adjustments" that address the residual idiosyncratic risk that remains undiversified in a (finite) portfolio; see Fermanian (2014) for a comprehensive review. Merino and Nyfeler (2002) uses a Fourier transform to analytically approximate the conditional loss distribution given the common factor. In contrast to these papers, we consider dynamic discrete-time, multistate models that allow one to measure pool-level risk over multiple future periods. Moreover, in our formulation, loans are not assumed to be conditionally independent given the common factor. We capture the influence of past events through a mean-field term, which addresses an important source of loan-to-loan correlation (see Azizpour et al. 2018 in this regard). We develop the limit theorems applying in this situation.

A number of authors prove limit theorems for continuous-time stochastic intensity and structural models of loan-by-loan default, and use them to construct

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

approximations for large pools. To address loan-to-loan correlation, these models may include a common factor and/or a mean-field or local interaction term. Dai Pra et al. (2009) and Dai Pra and Tolotti (2009) study a mean-field model in which an intensity is a function of the pool loss. In a model with local interaction, Giesecke and Weber (2006, 2004) take the intensity of a name as a function of the state of the names in a specified neighborhood of that name. Cvitanic et al. (2012) analyze a mean-field system in which an intensity is a function of an idiosyncratic risk factor, a systematic risk factor, and the portfolio loss rate. The risk factors follow diffusion processes whose coefficients may depend on the portfolio loss. Giesecke et al. (2013, 2015) and Spiliopoulos et al. (2014) analyze related mean-field systems in which the intensity is allowed to depend on the path of the portfolio loss. Therefore, the impact of a default may be transient and fade away with time. Davis and Rodriguez (2007) study an intensity model in a stochastic network setting, in which firms default independently of one another conditional on a systematic factor governed by a Markov chain. Diener et al. (2012) analyze an intensity model where defaults are conditionally independent given a systematic risk factor governed by a Brownian motion. Bush et al. (2011) examine a structural model in which defaults occur at first hitting times of correlated diffusion processes. Conditional on a systematic risk factor governed by a Brownian motion, defaults occur independently of one another.

In contrast to the aforementioned papers, we prove limit theorems in a discrete-time setting. A discrete-time formulation is appropriate for the data structure common in practice, where loan status is reported on a monthly or quarterly basis. Unlike earlier results, our limit theorems cover a wide range of statistical and machine learning models that are in widespread industry use (see Lending Club 2015, CoreLogic 2014, Experian 2014, Dunsky et al. 2014, and others). The fitting of these models from historical loan performance data is standard. Our formulation is well adapted to settings with high-dimensional loan-level feature data, where the features can be continuously valued. Previous model frameworks for which limiting laws have been proven do not include loan-level data as a model input. Unlike the aforementioned papers, we allow for multiple types of events that may be mutually exclusive. In practice, many loans are subject to several types of events, such as delinquency, prepayment, and foreclosure. The limiting laws in the above papers do not cover important applications where the function of interest is unbounded. This includes the expectation, variance, higher-order moments, and utility functions of the pool-level loss, prepayment, or cash flow. To address these applications, we prove uniform integrability, which allows our convergence results to cover a large class of functions of practical interest. Furthermore, uniform integrability allows us to prove a convergence rate for the law of large numbers. The existing literature mentioned above does not provide convergence rates.

Dembo et al. (2012), Glasserman et al. (2008), Spiliopoulos and Sowers (2011, 2015), Zhang et al. (2015), and others develop large-deviation results for credit portfolios in various model frameworks. While we do not study large deviations in this paper, our numerical results show that our approximation (with only the law of large numbers and central limit theorem) is very accurate in the tails of the distribution even for moderate-size portfolios.

2. Class of Loan-Level Models

We consider a broad family of dynamic models of loan-level risk in a pool at times $t \in I = \{0, 1, \ldots, T\}$. We fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and an information filtration $(\mathcal{F}_t)_{t\in I}$. $\mathbb{P}$ is the actual probability measure. The total number of loans initially in the pool is N. The process $U^n = (U_t^n)_{t\in I}$ prescribes the state of the nth loan. The variable U_t^n takes values in a finite discrete space $\mathcal{U}$. Common states are current, 30, 60, and 90+ days delinquent, prepaid, charge-off, and foreclosure. Some states might be absorbing.

State transitions are influenced by several types of risk factors. Each loan has an $\mathcal{F}_0$-measurable loan-level feature vector $Y^n \in \mathcal{Y} \subseteq \mathbb{R}^{d_Y}$,[1] which contains static variables such as loan-to-value (LTV) ratio, credit score, geographic location, type of loan, etc. These loan-level factors are specific to each loan. We also consider systematic risk factors V that will have a common influence across many loans in the pool. The vector process $V = (V_t)_{t\in I}$, where $V_t \in \mathbb{R}^{d_V}$, might represent the behavior of local and national economic conditions such as the unemployment rate, housing prices, and interest rates.[2] The risk factor V is exogenous in the sense that its dynamics are not affected by the states U^n nor the Y^n. Finally, we allow borrowers' past behavior to influence state transitions. Define the "mean-field" process $H^N = (H_t^N)_{t\in I}$ as

$$H_t^N = \frac{1}{N}\sum_{n=1}^{N} f^H(U_t^n, Y^n), \tag{1}$$

where $f^H = (f_1^H, \ldots, f_K^H)$ and $f_k^H \colon \mathcal{U} \times \mathbb{R}^{d_Y} \mapsto \mathbb{R}$. We let $H_{t:s}^N = (H_t^N, \ldots, H_s^N)$ denote the path of H^N between times t and $s \geq t$. A specific application for H^N would be to model a contagion effect for mortgages where past defaults of mortgages in areas geographically close to the nth mortgage increase the likelihood of the nth mortgage defaulting. The process H^N would then keep track of the number of defaults at each geographic location.[3] In light of the mortgage meltdown, such a feedback mechanism has been supported

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

by several recent empirical papers; see Ambrose and Capone (1998), Goodstein et al. (2017), Harding et al. (2009), Lin et al. (2009), and Towe and Lawley (2013).

The dynamics of U^n are prescribed by the transition function:

$$\mathbb{P}[U_t^n = u \mid \mathcal{F}_{t-1}] = h_\theta(u, U_{t-1}^n, Y^n, V_{t-1}, H_{t-\tau:t-1}^N), \quad t = 1, \dots, T, \tag{2}$$

where θ is a parameter that takes values in the compact Euclidean space Θ, $\tau \geq 1$ is some fixed integer, and H_t^N is set to some predetermined constant (independent of N) for $t < 0$. Therefore, the dynamics of the states $U^1, \dots, U^N$ can potentially depend on the history of the mean-field term H^N. Extension of the model (and convergence results) when h_θ depends on the history of the common factor V is straightforward. Note that even conditional on the path of the common factor V, the dynamics of the loans $U^1, \dots, U^N$ are not independent because of the mean-field term H^N.

Equation (2) gives the marginal probability for the transitions of the loans from their state at time $t-1$ to time t. Furthermore, we stipulate that conditional on $\mathcal{F}_{t-1}$, the states $U_t^1, \dots, U_t^N$ are independent. This fully specifies the dynamics of the loans $n = 1, \dots, N$. In addition, we model the financial losses suffered by the loan originator, servicer, or investor on certain state transitions. Given a state transition $U_{t-1}^n \neq \mathrm{d} \to U_t^n = \mathrm{d}$, where $\mathrm{d} \in \mathcal{U}$ represents an absorbing "default" state such as 90+ days delinquent, charge-off, or foreclosure, the nth loan suffers a loss $\ell_t^n(Y^n, V_t) \in [0,1]$. Here, $\ell_t^n(Y^n, V_t)$ is itself a random variable conditional on Y^n and V_t, therefore allowing for idiosyncratic losses.[4] Conditional on V_t, the loan-level feature Y^n, and the nth loan defaulting at time t, the loss given default is $\ell_t^n(Y^n, V_t)$ where $\mathbb{P}[\ell_t^n(y,v) \in A] = \nu_{t,y,v}(A)$. Conditional on V_t, Y^n, $Y^{n'}$, and both loans n, n' defaulting, $\ell_t^n(Y^n, V_t)$ and $\ell_t^{n'}(Y^{n'}, V_t)$ are independent. Since the default state is absorbing, losses for the same loan cannot occur twice. More generally, one could consider "costs" or losses associated with each of the state transitions. For example, there might be servicer costs associated with loans that are behind payment but have not defaulted yet. The results in this paper could be extended to this more general case. Typical loss given default models include logit models or neural networks; an overview can be found in Loterman et al. (2012).

Many models commonly used to describe loan delinquency and prepayment fall under the model class (2): see Khandani et al. (2010), Baesens (2005), Banasik et al. (1999), Bastos (2010), Capozza et al. (1997), Stepanova and Thomas (2002), Westgaard and der Wijst (2001), and many others. The model framework (2) is popular in practice (see Lending Club 2015, CoreLogic 2014, Experian 2014, Dunsky et al. 2014) because it allows for detailed modeling of loan-level dynamics, inclusion of high-dimensional loan-level data, flexible choice of transition functions from statistical and machine learning, and the ability to incorporate important characteristics of loan pools such as correlation between delinquency and prepayment rates for different types of loans, geographic diversity, and burnout.

A typical choice for h_θ might be a generalized linear model (GLM), although many other choices are available. An example of a GLM is logistic regression. The results in this paper require only mild assumptions on the form of the function h_θ; these assumptions are satisfied by many standard models. Some examples are presented below.

Example 2.1 (Logistic Regression Model). Let the elements of $\mathcal{U}$ be $u_1, \dots, u_K$. If a loan is in state u_k, denote the states to which it can transition as $u_k^1, \dots, u_k^{N_k}$. Let $(y, v, H) \in \mathbb{R}^{d_Y + d_V + 1}$ be a vector containing the loan-level features, systematic factors, and mean-field factor. The vector $\theta_{u_k^n, u_k} \in \mathbb{R}^{d_Y + d_V + 1}$ contains the parameters corresponding to the state transition $u_k \to u_k^n$. Then,

$$h_\theta(u_k^n, u_k, y, v, H) = \frac{\exp(\theta_{u_k^n, u_k} \cdot (y, v, H))}{1 + \sum_{n'=1}^{N_k - 1} \exp(\theta_{u_k^{n'}, u_k} \cdot (y, v, H))}, \quad n < N_k,$$

$$h_\theta(u_k^{N_k}, u_k, y, v, H) = 1 - \sum_{n'=1}^{N_k - 1} h_\theta(u_k^{n'}, u_k, y, v, H),$$

where $\theta = \{\theta_{u_k^n, u_k}\}_{k=1:K, n=1:N_k-1}$. See Sirignano et al. (2016) for an implementation of this model for mortgages.

Example 2.2 (Neural Network Model). Let the elements of $\mathcal{U}$ be $u_1, \dots, u_K$. If a loan is in state u_k, denote the states to which it can transition as $u_k^1, \dots, u_k^{N_k}$. Then,

$$h_\theta(u_k^n, u_k, y, v, H) = \frac{\exp(\sigma_\theta(u_k^n, u_k, y, v, H))}{1 + \sum_{n'=1}^{N_k - 1} \exp(\sigma_\theta(u_k^{n'}, u_k, y, v, H))}, \quad n < N_k,$$

$$h_\theta(u_k^{N_k}, u_k, y, v, H) = 1 - \sum_{n'=1}^{N_k - 1} h_\theta(u_k^{n'}, u_k, y, v, H).$$

The activation function σ_θ is given by

$$\sigma_\theta(u_k^n, u_k, y, v, H) = W_\theta^2 q(W_\theta^1 Q^n),$$

where $Q^n = (u_k^n, u_k, y, v, H) \in \mathbb{R}^{d_Q \times 1}$, and $W_\theta^1 \in \mathbb{R}^{d_h \times d_Q}$ and $W_\theta^2 \in \mathbb{R}^{1 \times d_h}$ are weights on the input Q^n and the output of the hidden layer q, respectively. The hidden layer q: $\mathbb{R}^{d_h \times 1} \mapsto \mathbb{R}^{d_h \times 1}$ is the function $q(x) = (q_1(x_1), q_2(x_2), \dots, q_{d_h}(x_{d_h}))^\top$. Typical choices for $q_1, \dots, q_{d_h}$ are sigmoid functions and d_h is the number of "neurons" in the hidden layer. See Sirignano et al. (2016) for an implementation of this model for mortgages.

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

Example 2.3 (Decision Tree Model). Divide the space $\mathcal{U}\times\mathcal{Y}\times\mathbb{R}^{d_V}\times\mathbb{R}^K$ into the regions $\{R_\theta^m\}_{m=1}^M$, with the condition that these regions are disjoint and their union covers the entire space. The transition probabilities in region m are $p_\theta^m(u,u',y,v,H)$. The transition probability h_θ then is

$$h_\theta(u,u',y,v,H)=\sum_{m=1}^M p_\theta^m(u,u',y,v,H)\mathbf{1}_{(u',y,v,H)\in R_\theta^m}.$$

The regions $\{R_\theta^m\}_{m=1}^M$ are chosen by sequentially splitting the space via recursive binary splitting. For each region m, the transition probabilities $p_\theta^m(u,u',y,v,H)$ could be a logistic regression or, more simply, a constant transition matrix between the different states. See Hastie et al. (2009) for more details on decision trees.

For risk management and other applications, one is interested in the behavior of a pool of N loans under the model class (2). The behavior of the pool at time t can be described by the empirical measure of the variables $(U_t^1,Y^1),\dots,(U_t^N,Y^N)$, which is denoted by $\mu_t^N\in B=\mathcal{P}(\mathcal{U}\times\mathcal{Y})$, where $\mathcal{P}$ is the space of probability measures. Formally,

$$\mu_t^N=\frac{1}{N}\sum_{n=1}^N\delta_{(U_t^n,Y^n)}, \tag{3}$$

where δ is the Dirac measure. The empirical measure $\mu_t^N(u,A)=\int_A\mu_t^N(u,dy)$ gives the fraction of the loans in pool at time t that are in state u and that have loan-level features in the set $A\subset\mathbb{R}^{d_Y}$. It completely encodes, up to permutations, the pool dynamics and features. For instance, the pool-level default rate can be expressed as $\int_{\mathcal{Y}}\mu_t^N(\mathrm{d},dy)$. Representing prepayment by an absorbing state $\mathrm{p}\in\mathcal{U}$, the pool-level prepayment rate takes the form $\int_{\mathcal{Y}}\mu_t^N(\mathrm{p},dy)$. The process $H^N\in\mathbb{R}^K$ specifying the mean-field term in (2) can also be expressed in terms of the empirical measure: $H_t^N=\langle f^H,\mu_t^N\rangle_{\mathcal{U}\times\mathcal{Y}}$. Here and below, we use the notation $\langle f,\nu\rangle_E=\int_E f(x)\nu(dx)$.[5] For notational convenience, let $\langle f(u,y),\nu(u,dy)\rangle=\langle f(u,\cdot),\nu(u,\cdot)\rangle_{\mathcal{Y}}$. The default loss rate L_t^N at time t is

$$L_t^N=\frac{1}{N}\sum_{n=1}^N\ell_t^n(Y^n,V_t)(\mathbf{1}_{U_t^n=\mathrm{d}}-\mathbf{1}_{U_{t-1}^n=\mathrm{d}}). \tag{4}$$

The cumulative loss from default up until time t is simply $\sum_{s=1}^t L_s^N$. Let the loss process over all times be $L^N=(L_t^N)_{t\in I}$. Given the empirical measure and the loss from default of a pool, analysis of securities backed by the pool is immediate. For instance, the distribution of the cash flow from a pool of loans can be analyzed, an example of which is given below.

Example 2.4 (Cash Flow from a Loan Pool). Let $\mathcal{U}=\{c,p,d\}$, where c represents a loan that is current on its payments, p represents prepayment, and d represents default. Let the cash flow for a loan with feature y in state u at time t, when the common factor $V_t=v$, be $r_{t,v}(u,y)$. It is comprised of both interest and principal payments. The dependence on V_t allows one to capture a variable loan interest rate tied to a benchmark rate. Let the outstanding balance for a loan with feature y at time t be $b_t(y)$. On prepayment, the borrower will pay off the entire outstanding balance. The functions $r_{t,v}(u,y)$ and $b_t(y)$ are deterministic. The recovery from default at time t is $1-L_t^N$. The total cash flow rate from a pool of loans at time t is

$$C_t=\langle r_{t,V_t},\mu_t^N\rangle_{\mathcal{U}\times\mathcal{Y}}+\left\langle b_t,\mu_t^N(\mathrm{p},\cdot)-\mu_{t-1}^N(\mathrm{p},\cdot)\right\rangle+(1-L_t^N).$$

The total cash flow rate is just the summation of C_t over all times $t\in I$.

3. Limiting Laws and Asymptotic Approximation

For large pools of loans (i.e., large N), the computation of the distribution of the default rate, prepayment rate, and loss from default is computationally burdensome. We develop an efficient asymptotic approximation of these distributions for large pools. The approximation is based on a law of large numbers (LLN) and a central limit theorem (CLT) for the empirical measure μ^N and the loss process L^N. The proofs for the theorems below are given in the online appendix.

Assumption 3.1. *Suppose that μ_0^N converges in distribution to $\bar\mu_0$ in $B=\mathcal{P}(\mathcal{U}\times\mathcal{Y})$, where $\bar\mu_0$ is deterministic, and that the feature space $\mathcal{Y}$ is compact. Also, h_θ is continuous in its third and fifth arguments and the function f^H is continuous and bounded in its second argument.*

Theorem 3.2. *Provided Assumption 3.1, the empirical measure μ^N converges in distribution to $\bar\mu$ in B^{T+1} as $N\to\infty$, where $\bar\mu$ satisfies*

$$\bar\mu_t(u,dy)=\sum_{u'\in\mathcal{U}}h_\theta(u,u',y,V_{t-1},\bar H_{t-\tau:\,t-1})\bar\mu_{t-1}(u',dy), \tag{5}$$

where $\bar H_t=\langle f^H,\bar\mu_t\rangle_{\mathcal{U}}\times\mathcal{Y}$ and $\bar H_{t:s}=(\bar H_t,\dots,\bar H_s)$.

The assumptions required for Theorem 3.2 are relatively mild and are satisfied by standard models. For instance, logistic regression and neural networks both satisfy Assumption 3.1. The compactness assumption for the feature space $\mathcal{Y}$ is also satisfied by commonly used data features (such as credit score, LTV ratio, interest rate, and original balance). The assumption that the function f^H is bounded and continuous still allows the mean-field term H^N to keep track of the fraction of the pool in any combination of states in $\mathcal{U}$ and categories in $\mathcal{Y}$. For example, H^N could track the fraction of the pool that has defaulted at each geographic location as well as the fraction of the pool that has defaulted for each loan product type (such term loan, credit line, etc.).

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

It is important to note that the LLN in Theorem 3.2 is dynamic and is also a random equation. Randomness enters through the path of the common factor V: the quantities $\bar{H}_t$ and $\bar{\mu}_t$ are deterministic functions of $V_{0:t-1} = (V_0, \ldots, V_{t-1})$. If the transition function h_θ depends on the past pool dynamics through $H^N_{t-\tau:t-1}$, the LLN is nonlinear.

We supplement the LLN with a CLT. Define the empirical fluctuation process $\Xi^N_t = \sqrt{N}(\mu^N_t - \bar{\mu}_t)$, which takes values in $W = \prod_{u=1}^{|\mathcal{U}|} S'(\mathbb{R}^{d_\mathcal{Y}})$, where S' is the space of tempered distributions. An element in the space of tempered distributions is the finite-order (distributional) derivative of some continuous function with polynomial growth. The space of measures is a subset of the space of tempered distributions. Convergence of Ξ^N_t must be considered in a larger space than the space of measures since Ξ^N_t can take negative values and $\int_A \Xi^N_t(u, dy)$ is not bounded. Although Ξ^N_t is not an element of $\mathcal{P}(\mathcal{U} \times \mathbb{R}^{d_\mathcal{Y}})$, it does exist in the larger space W. Moreover, its limit $\bar{\Xi}_t$ also exists in W. The CLT satisfies an equation linearized around the nonlinear dynamics of the LLN. Randomness enters both through V and a martingale term. Like the LLN, the CLT is also dynamic.

Assumption 3.3. *$\sqrt{N}(\mu^N_0 - \bar{\mu}_0)$ converges in distribution to $\bar{\Xi}_0$ in W. In addition, the function f^H is smooth in its second argument and h_θ is smooth in its third and fifth arguments.*[6]

Theorem 3.4. *Provided Assumptions 3.1 and 3.3, Ξ^N converges in distribution to $\bar{\Xi}$ in W^{T+1} as $N \to \infty$, where $\bar{\Xi}$ satisfies the equation*

$$\begin{aligned}\bar{\Xi}_t(u, dy) = &\sum_{u' \in \mathcal{U}} h_\theta(u, u', y, V_{t-1}, \bar{H}_{t-\tau:t-1}) \bar{\Xi}_{t-1}(u', dy) \\ &+ \sum_{u' \in \mathcal{U}} \left(\frac{\partial}{\partial H} h_\theta(u, u', y, V_{t-1}, \bar{H}_{t-\tau:t-1}) \cdot \bar{E}_{t-\tau:t-1} \right) \\ &\cdot \bar{\mu}_{t-1}(u', dy) + \bar{\mathcal{M}}_t(u, dy), \end{aligned} \tag{6}$$

where $E^N_t = \langle f^H, \Xi^N_t \rangle_{\mathcal{U} \times \mathcal{Y}}$, $\bar{E}_t = \langle f^H, \bar{\Xi}_t \rangle_\mathcal{U} \times \mathcal{Y}$, and $\bar{E}_{t:s} = (\bar{E}_t, \ldots, \bar{E}_s)$. Given V, $\bar{\mathcal{M}}(u, dy)$ is a conditionally Gaussian process with zero mean and covariance:

$$\begin{aligned}&\mathrm{Cov}[\bar{\mathcal{M}}_t(u_1, dy), \bar{\mathcal{M}}_t(u_2, dy) \mid V_{0:t-1}] \\ &\quad = -\sum_{u' \in \mathcal{U}} h_\theta(u_1, u', y, V_{t-1}, \bar{H}_{t-\tau:t-1}) \\ &\qquad \cdot h_\theta(u_2, u', y, V_{t-1}, \bar{H}_{t-\tau:t-1}) \bar{\mu}_{t-1}(u', dy), \\ &\mathrm{Var}[\bar{\mathcal{M}}_t(u, dy) \mid V_{0:t-1}] \\ &\quad = \sum_{u' \in \mathcal{U}} h_\theta(u, u', y, V_{t-1}, \bar{H}_{t-\tau:t-1}) \\ &\qquad \cdot (1 - h_\theta(u, u', y, V_{t-1}, \bar{H}_{t-\tau:t-1})) \bar{\mu}_{t-1}(u', dy), \end{aligned}$$

where $u_1 \neq u_2$ and $V_{0:t} = (V_0, \ldots, V_t)$. Given V, $\bar{\mathcal{M}}_{t_2}$ is independent of $\bar{\mathcal{M}}_{t_1}$ for $t_2 \neq t_1$ and $\bar{\mathcal{M}}_t(u, dy_1)$ is independent of $\bar{\mathcal{M}}_t(u', dy_2)$ for $y_1 \neq y_2$.

The additional assumptions necessary for Theorem 3.4 are mild. The main additional assumption is that h_θ and f^H are smooth in the loan feature y. Typical models for h_θ such as logistic regression, neural networks, probit, generalized linear models with a smooth link function, kernel methods with smooth kernels (e.g., Gaussian kernel), and Gaussian process regression satisfy Assumptions 3.1 and 3.3. Because of their discontinuities, decision trees do not satisfy Assumptions 3.1 and 3.3. However, if the Y^n are i.i.d., Theorems 3.2 and 3.4 will hold if h_θ is piecewise smooth in its third argument and smooth in its fifth argument, which would cover decision trees. Although we prove convergence in S', whose dual is the space of Schwartz functions, smoothness of h_θ and f^H on $\mathcal{Y}$ suffices for Theorem 3.4 to hold since $\mathcal{Y}$ is compact. Finally, the restriction on f^H is not stringent; the mean-field term H^N can still keep track of the fraction of the pool in any combination of states in $\mathcal{U}$ and categories in $\mathcal{Y}$.

Theorems 3.2 and 3.4 can be extended to the loss from default (4). To this end, define the empirical loss fluctuations $\Lambda^N_t = \sqrt{N}(L^N_t - \bar{L}_t) \in \mathbb{R}$.

Assumption 3.5. *The first and second moments of the loss given default, $g_1(t, v, y) = \int_0^1 z \nu_{t,y,v}(dz)$ and $g_2(t, v, y) = \int_0^1 z^2 \nu_{t,y,v}(dz)$, are smooth in y for each t, v.*

Proposition 3.6. *Let Assumptions 3.1, 3.3, and 3.5 hold. The loss process L^N converges in distribution to $\bar{L}$ in $[0,1]^{T+1}$ as $N \to \infty$, where $\bar{L}$ satisfies*

$$\bar{L}_t = \left\langle \int_0^1 z \nu_{t,y,V_t}(dz), \bar{\mu}_t(\mathrm{d}, dy) - \bar{\mu}_{t-1}(\mathrm{d}, dy) \right\rangle. \tag{7}$$

The fluctuations for the loss Λ^N converge in distribution to $\bar{\Lambda}$ in $\mathbb{R}^{T+1}$ as $N \to \infty$ where $\bar{\Lambda}$ satisfies

$$\bar{\Lambda}_t = \left\langle \int_0^1 z \nu_{t,y,V_t}(dz), \bar{\Xi}_t(\mathrm{d}, dy) - \bar{\Xi}_{t-1}(\mathrm{d}, dy) \right\rangle + \bar{\mathcal{Z}}_t, \tag{8}$$

where, conditional on V, $\bar{\mathcal{Z}}_t$ is a mean-zero Gaussian with variance

$$\begin{aligned}\mathrm{Var}[\bar{\mathcal{Z}}_t \mid V_{0:t}] = \Bigg\langle &\left[\int_0^1 z^2 \nu_{t,y,V_t}(dz) - \left(\int_0^1 z \nu_{t,y,V_t}(dz) \right)^2 \right], \\ &\bar{\mu}_t(d, dy) - \bar{\mu}_{t-1}(d, dy) \Bigg\rangle. \end{aligned}$$

Given V, $\bar{\mathcal{Z}}_{t_1}$ is independent of $\bar{\mathcal{Z}}_{t_2}$ for $t_1 \neq t_2$ as well as $\bar{\Xi}$.

Assumption 3.5 requires smoothness of the first and second moments of the loss given default. Standard models of the loss given default such as neural networks (see Loterman et al. 2012 for an overview) satisfy this assumption.

Once the law of large numbers $\bar{\mu}$ and central limit theorem $\bar{\Xi}$ have been computed, the law of large numbers $\bar{L}$ and the central limit theorem $\bar{\Lambda}$ for the loss can

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

be easily computed. The law of large numbers for the loss $\bar{L}$ is a function of $\bar{\mu}$. Similarly, the central limit theorem for the loss $\bar{\Lambda}$ is a function of $\bar{\Xi}$ and an independent Gaussian random variable $\bar{Z}$.

Corollary 3.7. *Under Assumptions 3.1, 3.3, and 3.5, $(V, \mu^N, L^N, \Xi^N, \Lambda^N)$ converges in distribution to $(V, \bar{\mu}, \bar{L}, \bar{\Xi}, \bar{\Lambda})$ in $(\mathbb{R}^{d_V})^{T+1} \times B^{T+1} \times \mathbb{R} \times W^{T+1} \times \mathbb{R}$ as $N \longrightarrow \infty$.*

Using Corollary 3.7, the law of large numbers and central limit theorem can be combined to form an approximation for a finite pool of N loans:

$$\mu^N = \bar{\mu} + \frac{1}{\sqrt{N}}\Xi^N \overset{d}{\approx} \bar{\mu}^N = \bar{\mu} + \frac{1}{\sqrt{N}}\bar{\Xi},$$
$$L^N = \bar{L} + \frac{1}{\sqrt{N}}\Lambda^N \overset{d}{\approx} \bar{L}^N = \bar{L} + \frac{1}{\sqrt{N}}\bar{\Lambda}. \tag{9}$$

The LLN $\bar{\mu}$ is a first-order approximation of μ^N while the CLT $\bar{\Xi}$ provides a second-order correction. The large-pool approximations (9) are conditionally Gaussian given V and can be utilized to efficiently estimate pool-level quantities such as prepayment, default, loss, and cash flow rates for large pools of loans. Section 4 below explains the computation of the approximation.

The Theorems 3.2 and 3.4 prove that the distribution of the finite pool and its fluctuations converge to the distribution of the LLN and CLT. However, this does not guarantee that the statistics, such as the moments, of the distribution of the finite pool and its fluctuations converge to the statistics of the LLN and CLT.[7] In particular, $X^N \overset{d}{\to} \bar{X}$ only implies $\mathbb{E}[f(X^N)] \to \mathbb{E}[f(\bar{X})]$ for f continuous and *bounded*. To prove $\mathbb{E}[f(X^N)] \to \mathbb{E}[f(\bar{X})]$ for continuous f, one needs to show that $f(X^N)$ is uniformly integrable. Since $\mathcal{Y}$ is assumed to be compact, $\mathcal{P}(\mathcal{Y} \times \mathcal{U})$ is also compact. Consequently, the empirical measure μ^N lies in a compact space (the space of measures on a compact space is compact). The loss L^N also lies in the compact set $[0, 1]$. $f(\mu^N, L^N)$ is therefore uniformly integrable for any continuous function f and $\mathcal{E}[f(\mu^N, L^N)] \to \mathcal{E}[f(\bar{\mu}, \bar{L})]$ for any continuous function f. The main challenge resides in showing that $f(\Xi^N, \Lambda^N)$ is also uniformly integrable. Unlike μ^N, Ξ^N and Λ^N are not restricted to compact sets.

For instance, convergence in distribution is not sufficient for the approximation (9) to hold for the second moment of the loss at time t:

$$\mathbb{E}[(L_t^N)^2] = \mathbb{E}[\bar{L}_t^2] + \frac{2}{\sqrt{N}}\mathbb{E}[\bar{L}_t \Lambda_t^N] + \frac{1}{N}\mathbb{E}[(\Lambda_t^N)^2]. \tag{10}$$

To justify the approximation $\mathbb{E}[(L_t^N)^2] \approx \mathbb{E}[(\bar{L}_t^N)^2]$, one needs the convergence results $\mathbb{E}[\bar{L}_t \Lambda_t^N] \to \mathbb{E}[\bar{L}_t \bar{\Lambda}_t]$ and $\mathbb{E}[(\Lambda_t^N)^2] \to \mathbb{E}[(\bar{\Lambda}_t)^2]$ as $N \to \infty$. Because of $|\bar{L}_t|$ being bounded, it is sufficient to show that Λ_t^N and $(\Lambda_t^N)^2$ are uniformly integrable.

Assumption 3.8. *For functions of the form $f(\Xi^N, \Lambda^N) = g(\sum_{t=1}^T \sum_{u \in \mathcal{U}} \langle \phi_{t,u}(y), \Xi_t^N(u, dy) \rangle + \sum_{t=1}^T c_t \Lambda_t^N)$, where $c_t \in \mathbb{R}$, g is continuous and polynomially bounded,[8], and the functions $\phi_{t,u}$ are continuous. In addition, $V_{0:T} = (V_0, V_1, \ldots, V_T)$ takes values in a compact set $\mathcal{V}$, and for any continuous ϕ, there are positive constants $K_{1,0,u}$ and $K_{2,0,u}$ such that*

$$\mathbb{P}\left[\left(\sup_{y \in \mathcal{Y}} |\phi(y)|\right)^{-1} |\langle \phi(y), \Xi_0^N(u, dy) \rangle| > \alpha \,\middle|\, V_{0:T}\right] \leq K_{1,0,u} \exp(-K_{2,0,u}\alpha^2). \tag{11}$$

Theorem 3.9. *Under Assumptions 3.1, 3.3, 3.5, and 3.8, $f(\Xi^N, \Lambda^N)$ is uniformly integrable.*

The class of functions covered by Assumption 3.8 can represent any function g of the default, loss, prepayment, or cash flow rates where g does not have faster than polynomial growth in its tails. We are not aware of any applications with exponential growth in the tails, so this class of functions is quite general. A sufficient condition for the bound (11) is that the loan features Y^n are independent; then, by the Azuma–Hoeffding inequality, the bound is satisfied.

The ideas and results in this paper, which are developed for discrete time, can be extended to a continuous-time framework where events are governed by counting processes with intensities. The law of large numbers and central limit theorem for the continuous-time framework satisfy a random ODE and an SDE, respectively. Their forms are very similar to the forms given above for the discrete-time model considered in this paper. In fact, the discrete-time limiting laws presented here are the Euler discretization for the continuous time limiting laws. Some examples of continuous-time loan models that would be covered under such an extension are the mortgage models studied by Deng et al. (2000) and Mattey and Wallace (2001).

The model (2) can also be extended to allow for transitions to depend on the entire path of V up until time t. The loan dynamics would be $\mathbb{P}[U_t^n = u \mid \mathcal{F}_{t-1}, V_t] = h_\theta(u, U_{t-1}^n, Y^n, V_{t-\tau_v:t}, H_{t-\tau:t-1}^N)$. $V_{t-\tau_v:t} = (V_{t-\tau_v}, \ldots, V_t)$ denotes the path of V between times $t - \tau_v$ and t. In this extended model, loan transitions from time $t-1$ to t are not independent conditional on $\mathcal{F}_{t-1}$. The proofs of all theorems and results hold trivially for this extended model.

4. Computing the Approximation

The approximation (9) can be computed by Monte Carlo simulation. First, many paths are simulated from the systematic process V. Conditional on each path of V, $\bar{\mu}$ can be calculated deterministically and the conditional covariance of $\bar{\Xi}$ can be evaluated in closed form. Since $\bar{\Xi}$ is conditionally Gaussian, this directly yields the conditional distribution of $\bar{\Xi}$. Then, the unconditional distribution for the approximation $\bar{\mu}^N$

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

can be found by averaging the conditional distributions across the paths of V. We highlight that the approximate loss $\bar{L}^N$ is simply a linear function of $(\bar{\mu}, \bar{\Xi})$ plus some independent Gaussian noise, and therefore is also very straightforward to simulate. In effect, the approximate loss $\bar{L}^N$ can be simulated (at no extra cost) on top of the simulated paths of $(\bar{\mu}, \bar{\Xi})$.

4.1. Algorithm

The numerical evaluation of $\bar{\mu}$ and $\bar{\Xi}$ conditional on each path of V requires discretizing $\mathbb{R}^{d_Y}$ into computational cells and then calculating $\bar{\mu}_t(u, dy)$ and $\bar{\Xi}_t(u, dy)$ at the center of each cell. Let the set of cells be $\mathscr{C} = \{c_1, \ldots, c_K\}$ where $\bigcup_{k=1}^K c_k = \mathscr{Y}$ with associated centers (i.e., grid points) $\mathscr{R} = \{y_1, \ldots, y_K\}$ where $y_k \in c_k$. The initial measures $\bar{\mu}_0$ would be calibrated to the data of the pool, and the second-order correction $\bar{\Xi}_0$ would be set to zero. For each V, $\bar{\Xi}_t$'s distribution on the set of grid points can be computed in closed form (since it is conditionally Gaussian). The approach is outlined below:

- Define $\bar{\mu}^\Psi$ and $\bar{\Xi}^\Psi$ to satisfy the LLN and CLT evolution equations, but with initial condition $\bar{\mu}^\Psi_{t=0} = \Psi(u, dy)$ where $\Psi(u, dy) = \sum_{k=1}^K \bar{\mu}_{t=0}(u, c_k)\delta_{y_k}$.[9]
- Simulate paths $V^1, \ldots, V^L$ from the random variable V. Conditional on a path V^l, let $\bar{\mu}^\Psi$ and $\bar{\Xi}^\Psi$ be $\bar{\mu}^{\Psi,l}$ and $\bar{\Xi}^{\Psi,l}$, respectively. For each V^l:

  —Calculate $\bar{\mu}_t^{\Psi,l}(u, y_k)$ for all $y_k \in \mathscr{R}$ at times $t = 1, \ldots, T$.

  —Calculate the (closed-form) covariance of $\bar{\Xi}_t^{\Psi,l}$ across the grid points $y_1, \ldots, y_K$ at $t = 1, \ldots, T$.[10]

  —For each V^l and any function f, a numerical approximation can be made:

$$\begin{aligned}
&\langle f, \mu_t^{N,l}\rangle_{\mathscr{U}\times\mathscr{Y}} \\
&\quad \overset{d}{\approx} \langle f, \bar{\mu}_t^l\rangle_{\mathscr{U}\times\mathscr{Y}} + \frac{1}{\sqrt{N}}\langle f, \bar{\Xi}_t^l\rangle_{\mathscr{U}\times\mathscr{Y}} \\
&\quad \approx \sum_{k=1}^K \sum_{u\in\mathscr{U}} f(u, y_k)\left[\bar{\mu}_t^{\Psi,l}(u, y_k) + \frac{1}{\sqrt{N}}\bar{\Xi}_t^{\Psi,l}(u, y_k)\right] \\
&\quad \sim \mathscr{N}(m^l, \Sigma^l),
\end{aligned}$$

where $\mu_t^{N,l}$ is the empirical measure μ_t^N for the path V^l, $m^l = \sum_{k=1}^K \sum_{u\in\mathscr{U}} f(u, y_k)\bar{\mu}_t^{\Psi,l}(u, y_k)$, and $\Sigma^l = (1/N) \cdot \mathrm{Var}[\sum_{k=1}^K \sum_{u\in\mathscr{U}} f(u, y_k)\bar{\Xi}_t^{\Psi,l}(u, y_k)]$ (which can be calculated in closed form).

- Collecting the conditional distributions for each V^l, the density of $\langle f, \mu_t^N\rangle_{\mathscr{U}\times\mathscr{Y}}$ can be directly computed as $(1/L)\sum_{l=1}^L \phi(z, m^l, \Sigma^l)$ where $\phi(\cdot, m, \Sigma)$ is the density of Gaussian random variable with mean m and covariance Σ.

Alternatively, $\bar{\Xi}_t$ could be directly simulated instead of finding its closed-form covariance conditional on each V^l. In this latter approach, one would simply simulate $\bar{\Xi}_t^\Psi$ for each V^l. Such simulation is straightforward since $\bar{\Xi}_t$ is conditionally Gaussian given V. The simulation scheme above has the advantage of lower variance per sample with the same bias.

The convergence for the numerical error of the simulation scheme is analyzed in Online Appendix B. We show that as the sizes of the computational cells tend to zero (i.e., the number of grid points tends to infinity), the error also converges to zero. Moreover, a specific convergence rate is provided that can guide the choice of the location of grid points, number of grid points, and number of Monte Carlo trials. Given a fixed computational budget, there is a trade-off between more trials (reduction of variance) and more grid points (reduction of bias). In Online Appendix B, we derive a closed-form formula for the optimal allocation between the number of grid points and the number of trials provided a fixed budget.

We also prove a convergence rate for the approximation itself (including both approximation error of the law of large numbers as well as numerical error from the numerical scheme that computes the law of large numbers). Let the maximum "radius" of the computational cells be

$$r_{K,i} = \tfrac{1}{2}\max_{k=1,\ldots,K}\max_{y_i, y_i' \in c_k} |y_i - y_i'|.$$

The mean-squared error of the numerical scheme for the finite system expected value is

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathbb{E}g(\langle f, \mu_t^N\rangle_{\mathscr{U}\times\mathscr{Y}}) - \frac{1}{L}\sum_{l=1}^L g(m^l)\right)^2\right] \\
&\quad \le \underbrace{\frac{C_6}{L}\mathrm{Var}[g(\langle f, \bar{\mu}_t\rangle_{\mathscr{U}\times\mathscr{Y}})]}_{\text{variance}} + \underbrace{C_7 d_Y \sum_{i=1}^{d_Y} \mathbb{E}[C_{3,t,i}(V)^2] r_{K,i}^2}_{\text{bias from numerical quadrature}} \\
&\qquad + \underbrace{\frac{C_8}{N}}_{\substack{\text{bias from approximation} \\ \text{of finite system}}}
\end{aligned} \tag{12}$$

The first term is the error from the Monte Carlo variance and can be reduced by increasing the number of Monte Carlo samples L. The second term is the numerical quadrature error and can be reduced by decreasing the cell size $r_{K,i}$. The third term is the approximation error for the finite system and decreases as the size of the system gets large. Numerical experiments in Section 5 suggest that the approximation bias for the law of large numbers is very small even for N in the low thousands. The constant C_8 does not depend on the number of Monte Carlo samples L nor the cell size $r_{K,i}$.

Traditional Monte Carlo simulation of the system is unbiased and has mean-squared error that decreases as C_{MC}/L. However, each simulation of the system requires simulating N individual loans. That is, NL paths must be simulated to get L samples from

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

$g(\langle f, \mu_t^N \rangle_{\mathcal{U}\times\mathcal{Y}})$. The cost of the approximation depends on the number of grid points required; as can be seen from (12), this typically increases as d_Y increases. Therefore, a rough comparison of the computational costs of the approximation versus traditional Monte Carlo is the number of grid points versus the size of the system N; for instance, see Equation (15) and the corresponding discussion later in this paper.

The computational performance of the LLN and CLT can be further improved by the use of a nonuniform grid for discretizing $\mathbb{R}^{d_Y}$. Using a nonuniform grid, more points would be placed where $\bar{\mu}_0(u, dy)$ is large and less points would be placed where $\bar{\mu}_0(u, dy)$ is small. Online Appendix D provides an example of a particular nonuniform grid well adapted to our problem. The proposed grid is highly accurate even with only a small number of grid points. A sparse grid can also be used to further decrease computational time. Online Appendix E describes this approach. Using a sparse grid, simulation is performed at only a few points and then the solution is evaluated on a finer grid via interpolation. A final advantage of the approximation is its reusability; one set of simulations on a prechosen grid can be used to calculate the distribution for many different pools by reweighting the simulated CLT and LLN. Another choice, implemented in Section 5.6, is a nonuniform sparse grid using k-means clustering.

4.2. Efficient Low-Dimensional Approximation

If the loan-level feature space $\mathcal{Y}$ is very high-dimensional, the law of large numbers and central limit theorem will be high-dimensional and computations using traditional *uniform grids* can become expensive because of the curse of dimensionality. The number of mesh points in a uniform grid will grow exponentially with the dimension d_Y of the loan-level feature space.

Fortunately, for a reasonably large subclass of models, one can perform a transformation of the LLN in Equation (5) that converts the LLN into a low-dimensional problem. Suppose that there is a function g_θ such that $h_\theta(u, u', y, v, H) = g_\theta(u, u', f_\theta(y), v, H)$ where $f_\theta: \mathcal{Y} \mapsto \mathbb{R}^{d_W}$. Then, h_θ is invariant under the coordinate transformation $w = f_\theta(y)$ and one can reduce the high-dimensional equation for $\bar{\mu}_t(u, dy)$ to the low-dimensional equation:

$$\bar{\mu}_t(u, dw) = \sum_{u' \in \mathcal{U}} g_\theta(u, u', w, V_{t-1}, \bar{H}_{t-\tau:t-1}) \bar{\mu}_{t-1}(u', dw), \quad w \in \mathbb{R}^{d_W}. \tag{13}$$

The same coordinate transformation can be used to arrive at a low-dimensional CLT in terms of w. No matter how large the original dimension d_Y, the transformation reduces the dimension d_Y of the LLN and CLT to a constant dimension d_W, paving the way for tractable computations.

An example of a function h_θ that satisfies the necessary requirement is any generalized linear model (GLM), such as logistic regression (see Example 2.1). GLMs are widely used in practice for default and prepayment modeling. Below is an example of the low-dimensional transformation for GLMs.

Example 4.1 (One-Dimensional Efficient Approximation). Let $\mathcal{U} = \{c, p\}$ where c represents a loan that is current on its payments and p represents prepayment. Let Y^n be a vector containing loan-level features for the nth loan and V_t be a vector of systematic factors at time t. For this example, h_θ does not depend on the mean-field term H^N. A logistic function is chosen for h_θ. Let the parameter vector $\theta = (\theta_{\mathrm{p}}^Y, \theta_{\mathrm{p}}^X)$ where $X_t = (1, V_t)$. Then,

$$h_\theta(\mathrm{p}, \mathrm{c}, Y^n, V_{t-1}, H^N_{t-\tau:t-1}) = \frac{e^{\theta_{\mathrm{p}}^X \cdot X_{t-1} + \theta_{\mathrm{p}}^Y \cdot Y^n}}{1 + e^{\theta_{\mathrm{p}}^X \cdot X_{t-1} + \theta_{\mathrm{p}}^Y \cdot Y^n}},$$
$$h_\theta(\mathrm{c}, \mathrm{c}, Y^n, V_{t-1}, H^N_{t-\tau:t-1}) = 1 - h_\theta(\mathrm{p}, \mathrm{c}, Y^n, V_{t-1}, H^N_{t-\tau:t-1}). \tag{14}$$

Let $w = \theta_{\mathrm{p}}^Y \cdot y$. Then, both the low-dimensional LLN (13) and the low-dimensional CLT only have one spatial dimension, even though y may be high-dimensional.

Furthermore, beyond GLMs, the transformation allows arbitrary complexity with respect to the input factors y and v; the sole restriction is that v can only interact with y through the function $f_\theta(y)$. For instance, h_θ could be of the form where it is a softmax function of two neural networks: the first neural network's input is v, and the second neural network's input is y. GLMs can also be made arbitrarily nonlinear in the input space y by adding features that are nonlinear functions of y (e.g., using polynomials or other basis functions). It is important to emphasize that the dimension d_W does not depend on how large the original dimension d_Y is; thus, even as d_Y grows, the computational cost of the low-dimensional LLN and CLT remains the same.

The initial distribution $\bar{\mu}_0(u, dw)$ in (13) can be calibrated directly from the data available for the pool of loans. Simply transform the features $Y = (Y^1, \ldots, Y^N)$ to $W = (W^1, \ldots, W^N)$ via the map f_θ and then calculate the empirical distribution of W. Because of their broad applicability and computational tractability, the low-dimensional LLN and CLT are very useful in practice.

The combination of the low-dimensional transformation and the nonuniform grid schemes mentioned in the previous section make simulating the pool loss and prepayment levels via the LLN and CLT very computationally tractable. For models of h_θ that do not fall under the subclass of models for which the low-dimensional transformation is applicable, or where $d_W > 3$ and therefore the low-dimensional transformation is too computationally costly, one can still compute the approximation (9) for high dimensions (without any transformation) via sparse nonuniform grids.

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

We describe and numerically implement this latter approach in Section 5.6, which also strongly outperforms brute-force simulation of the actual pool.

5. Computational Performance of the Approximation

We compare the accuracy and computational cost of our approximation with brute-force Monte Carlo simulation using actual mortgage data. We find that the approximation has very high accuracy even for small pools in the hundreds of mortgages. Its computational cost is typically orders of magnitude lower than brute-force Monte Carlo simulation of the actual pool.

5.1. Models and Data

Unless noted otherwise, we test the performance of the approximation using logistic regression models of the transition function h_θ in (2); see Example 2.1. The model parameters θ are estimated from loan performance data by the method of maximum likelihood; see Online Appendix F for details. For simplicity, the loss given default is normalized to one.

We consider two loan performance data sets, allowing us to run tests using different parameter sets. The first consists of over 10 million subprime mortgages during the time period 1995–2013. The mortgages are spread across the entire United States, covering over 36,000 different zip codes. The feature data (the Y^n) includes zip code, FICO score, loan-to-value (LTV) ratio, initial interest rate, initial balance, type of mortgage, deal ID, and time of origination. We consider three states: current, prepayment, and default. Prepayment and default events are recorded in the data set on a monthly basis, suggesting a monthly time step for the discrete-time model (2). The second data set, obtained from Freddie Mac, contains agency mortgages over the time period 1999–2014 and consists of 16 million mortgages. Freddie Mac insures against any default losses, so prepayment and default can be treated as the same event. Thus, we only consider two states, current and prepayment. Prepayment events are recorded on a monthly basis in the data set. The feature data (Y^n) includes FICO score, first-time homebuyer indicator, type of mortgage, maturity date, number of units, occupancy status, combined loan-to-value, original debt-to-income ratio, LTV ratio, initial interest rate, prepayment penalty mortgage indicator, loan purpose (purchase, cash-out refinance, or no cash-out refinance), original loan term, and number of borrowers. The zip codes for each mortgage have been partially anonymized in this data set, but the metropolitan statistical area (MSA) is reported for each mortgage. There are 430 different MSAs reported in the data set.

The loan performance and feature data sets are complemented with data on macroeconomic variables, including the national unemployment rate and the national mortgage rate. These variables represent the systematic factors V. They are modeled as (independent) discrete-time random walks with standard deviations fitted to their historical values.

Online Appendix F reports the fitted parameter θ for each of the two logistic regression models (one for subprime loans and the other for agency loans). The systematic factors are especially influential for both default and prepayment. The LTV ratio and initial interest rate are notably important factors for subprime default, while FICO score and initial interest rate are important factors for agency prepayment. Subprime prepayment appears to be largely driven by the systematic factors.

5.2. Performance of LLN and CLT

First, we demonstrate the accuracy of the approximate distribution by comparing it with the "actual distribution" for the pool. The actual distribution is the distribution under the fitted model and is found via brute-force Monte Carlo simulation using a large number of Monte Carlo samples. The pool is drawn at random from the subprime data set. (We note that the "actual distribution" itself has Monte Carlo error, but this Monte Carlo error is small as shown in Online Appendix G.)

We perform 25,000 Monte Carlo trials for both the brute-force simulation of the actual pool as well as the simulation of the approximation. We simulate the pool for a one-year time horizon, with monthly discretization ($T = 12$). Since we model both default and prepayment for subprime loans, $d_W = 2$ for the approximation. Our implementation of the approximation uses nonuniform grids; see Online Appendix D. Table 1 reports the accuracy of the approximation for 100 randomly drawn pools. Each pool is of size $N = 2{,}500$. Figures 1 and 2 of Online Appendix G further illustrate the accuracy of the approximation.

Computational costs are reported in Table 2 for different sized pools N. In general, a rough approximation of the ratio of the computational costs is

$$\frac{\text{Cost of LLN}}{\text{Cost of Simulation of Actual Pool}} = \frac{N_g}{N}, \tag{15}$$

where N_g is the number of grid points needed for the numerical solution of the LLN and CLT equations across the space $\mathbb{R}^{d_W}$, and N is the number of mortgages in the actual pool. For instance, if one only needs 25 grid points, the ratio of computational costs is $25/N$. For a million mortgages, that leads to a reduction in computational time of well over four orders in magnitude. Computational times listed in Table 2 for brute-force simulation of the pool, simulation of the LLN, and simulation of the full approximation (LLN combined with the CLT) are for $d_W = 2$ (model includes both prepayment and default). These computational times

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

Table 1. For 100 Pools, We Calculate Various Measures of the Accuracy of the Approximation for the 99% VaR for the Loss from Default

Approx.	MSE	Bias	Error in percent	Variance	10% quantile	90% quantile
LLN	1.54×10^{-5}	3.91×10^{-3}	0.39	5.06×10^{-7}	4.3×10^{-3}	3.51×10^{-3}
LLN and CLT	1.31×10^{-7}	1.35×10^{-4}	0.03	5.17×10^{-7}	6.06×10^{-4}	-3.81×10^{-4}

Notes. Each pool is of size $N = 2{,}500$ and is randomly drawn from the data set. We calculate the mean-squared error, the bias, and the average percent error (|Actual VaR − Approx. VaR| × 100%). We also calculate the variance of the approximation. Finally, as a measure of the rarity of "large approximation errors," we report the 10% and 90% quantiles of the error of the approximation across the 100 pools.

are for a 12-month time horizon and 25,000 Monte Carlo trials.

We also study the accuracy of the approximation for the moments of default losses for randomly selected subprime pools. Table 3 reports the percent error of the approximation for the moments of the pool loss for different sizes N at a $T = 12$ month time horizon. The actual values for the moments are calculated via brute-force Monte Carlo simulation with 50,000 trials. 50,000 Monte Carlo trials are also used for the approximation of the moments. There is lower comparative accuracy for higher-order moments that represent the non-Gaussian characteristics of the finite pool. These non-Gaussian characteristics fade as N becomes large.

5.3. Numerical Performance Across Actual Deals

The subprime data set covers over 6,000 deals.[11] As mentioned earlier, for each mortgage, a deal ID is available. One can therefore reconstruct the actual pools for deals. We will perform some numerical tests to study the approximation's accuracy across the wide diversity of deals in the data universe.

Figure 1 compares the approximate distribution with the actual distribution at a time horizon of 12 months for 10 actual deals in the data set. The deals are chosen at random and each contains between 5,000 and 10,000 mortgages. It is interesting that the default rate can vary considerably between deals as a consequence of the quality of the underlying mortgages, demonstrating how important it is for a model to consider the loan-level characteristics of the mortgages in the pool.

To further assess the accuracy of the approximation, a set of deals is selected at random from the data set, and the approximation's 99% value at risk (VaR) is compared with the true 99% value at risk. Tables 4 and 5 report the numerical performance of the approximation.

5.4. One-Dimensional Approximation

So far, we have focused on the case where $d_W = 2$ (both default and prepayment events are considered). To demonstrate the one-dimensional approach, we use the agency model, which only includes prepayment.

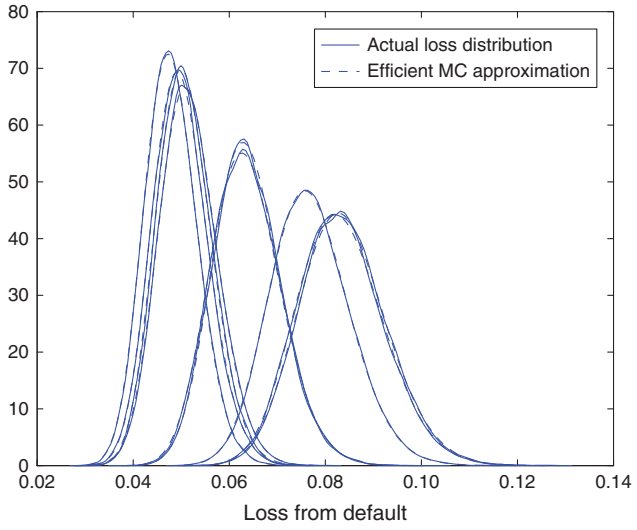
Table 2. Comparison of Computational Times (Seconds) for Approximation and Brute-Force Monte Carlo Simulation of the Pool

N	Time for brute-force simulation	Time for LLN	Time for LLN and CLT
1,000	44.34	1.03	2.67
5,000	153.78	1.03	2.67
10,000	273.39	1.03	2.67
25,000	608.94	1.03	2.67
100,000	2,847.43	1.03	2.67
1,000,000	28,563.68	1.03	2.67

Table 3. Percent Error for Monte Carlo Approximation for Moments of the Loss

N	First moment	Second moment	Third moment	Fourth moment	Fifth moment
500	0.69	1.3	1.9	2.5	3.2
1,000	0.44	0.87	1.3	1.7	2.2
5,000	0.09	0.19	0.28	0.36	0.45
10,000	0.07	0.14	0.22	0.30	0.37

Figure 1. (Color online) Comparison of Actual Distribution with Approximate Distribution for 10 Actual Deals with $5{,}000 < N < 10{,}000$



Notes. Loss reported as fraction of pool that defaulted. The horizon is 12 months.

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

Table 4. Accuracy of the Approximation for the 99% VaR for the Loss from Default and Prepayment, for 100 Pools of Size $4,000 \leq N \leq 10,000$

Approx.	MSE	Bias	Error in percent	Variance	10% quantile	90% quantile
LLN	3.19×10^{-6}	1.75×10^{-3}	0.18	2.66×10^{-4}	1.32×10^{-3}	2.2×10^{-3}
LLN and CLT	5.72×10^{-8}	2.80×10^{-5}	0.01	2.66×10^{-4}	-1.69×10^{-4}	2.59×10^{-4}

Notes. Each pool is an actual subprime mortgage deal with sizes $4{,}000 \leq N \leq 10{,}000$. Each mortgage deal is drawn at random from the data set. We calculate the mean-squared error, the bias, and the average percent error (|Actual VaR – Approx. VaR| × 100%). We also calculate the variance of the approximation. Finally, as a measure of the rarity of "large approximation errors," we report the 10% and 90% quantiles of the error of the approximation across the 100 pools. The time horizon is 12 months, and 50,000 Monte Carlo simulations are used.

Table 5. Accuracy of the Approximation for the 99% VaR for the Loss from Default and Prepayment, for 100 Pools of Size $300 \leq N \leq 2,500$

Approx.	MSE	Bias	Error in percent	Variance	10% quantile	90% quantile
LLN	5.34×10^{-5}	6.93×10^{-3}	0.69	4.24×10^{-4}	4.96×10^{-3}	8.86×10^{-3}
LLN and CLT	2.94×10^{-7}	4.01×10^{-4}	0.04	4.48×10^{-4}	-7.72×10^{-6}	-8.21×10^{-4}

Notes. Each pool is an actual subprime mortgage deal with sizes $300 \leq N \leq 2{,}500$. Each mortgage deal is drawn at random from the data set. We calculate the mean-squared error, bias, and average percent error (|Actual VaR – Approx. VaR| × 100%). We also calculate the variance of the approximation. Finally, as a measure of the rarity of "large approximation errors," we report the 10% and 90% quantiles of the error of the approximation across the 100 pools. The time horizon is 12 months, and 50,000 Monte Carlo simulations are used.

Table 6. Accuracy of the One-Dimensional Approximation for the 99% VaR for the Loss from Prepayment, for 100 Pools of Size $N = 2,500$

Approx.	MSE	Bias	Error in percent	Variance	10% quantile	90% quantile
LLN	4.27×10^{-6}	1.77×10^{-3}	0.17	3.99×10^{-5}	5.35×10^{-4}	2.91×10^{-3}
LLN and CLT	1.36×10^{-6}	4.59×10^{-4}	0.08	3.97×10^{-5}	-7.59×10^{-4}	1.59×10^{-3}

Notes. Each pool is drawn at random from the data set. We calculate the mean-squared error, bias, and average percent error (|Actual VaR – Approx. VaR| × 100%). We also calculate the variance of the approximation. Finally, as a measure of the rarity of "large approximation errors," we report the 10% and 90% quantiles of the error of the approximation across the 100 pools. The time horizon is 12 months, and 25,000 Monte Carlo simulations are used.

As indicated in Table 2 of Online Appendix F, this model has 22 feature variables (Y^n), which amount to 22 dimensions in the feature space. To obtain an even higher-dimensional model, we also include the metropolitan statistical area (MSA) as a feature variable. Since there are 430 MSAs in the data set and each of them is represented by a dummy variable, we now have $d_Y = 452$. We emphasize that, even though $d_Y = 452$, d_W only equals 1.

Table 6 reports results for the one-dimensional approximation. Figure 3 of Online Appendix G compares the approximation with the actual prepayment distribution for a randomly drawn pool of loans.

5.5. Precomputation for Financial Institutions

In Online Appendix E, we propose to presimulate $\bar{\mu}^N$ on a prechosen grid and then use this one set of simulations on the single grid to find the distribution for many different mortgage-backed securities (MBSs). Using this approach, the approximation can provide great computational cost savings even for very small MBSs, as long as a financial institution is dealing with many of these small MBSs in aggregate.

We now illustrate this approach using the agency model from Section 5.4. One hundred pools, each of size $N = 2{,}500$, are drawn at random from the agency mortgage data set. Each of these pools is simulated 25,000 times using brute-force Monte Carlo simulation. Using the approximation, we also presimulate on a prechosen grid. Twenty-five thousand Monte Carlo paths are also used for the simulation on this grid. The prechosen grid only has 20 grid points, placed at uniform intervals. The distribution of $\bar{\mu}^N$ is smooth in w, so one can numerically interpolate from the sparse grid points to get a finer solution. We use piecewise cubic spline interpolation.[12] This saves computa-

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

Table 7. Accuracy of the One-Dimensional Approximation Using Presimulation for the 99% VaR for the Loss from Prepayment, for 100 Pools of Size 2,500

Approx.	MSE	Bias	Error in percent	Variance	10% quantile	90% quantile
LLN	9.24×10^{-6}	1.51×10^{-3}	0.17	1.79×10^{-5}	-8.97×10^{-3}	4.11×10^{-3}
LLN and CLT	5.87×10^{-6}	2.60×10^{-4}	0.14	1.79×10^{-5}	-1.89×10^{-3}	2.96×10^{-3}

Notes. Each pool is of size $N = 2{,}500$ and is randomly drawn from the data set. We calculate the mean-squared error, the bias, and the average percent error (|Actual VaR − Approx. VaR| × 100%). We also calculate the variance of the approximation. Finally, as a measure of the rarity of "large approximation errors," we report the 10% and 90% quantiles of the error of the approximation across the 100 pools. The time horizon is 12 months, and 25,000 Monte Carlo simulations are used.

Table 8. Accuracy of the Approximation Without Low-Dimensional Transformation for the 99% VaR for the Loss from Default and Prepayment, for 100 Pools of Size $5{,}000 \le N \le 10{,}000$

Approx.	MSE	Bias	Error in percent	Variance	10% quantile	90% quantile
LLN	2.54×10^{-5}	5.00×10^{-3}	0.50	2.94×10^{-8}	4.18×10^{-3}	5.7×10^{-3}
LLN and CLT	2.59×10^{-8}	1.11×10^{-4}	0.14	3.57×10^{-7}	-4.72×10^{-5}	2.56×10^{-4}

Notes. Each pool is an actual subprime mortgage deal with sizes $5{,}000 \le N \le 10{,}000$. Each mortgage deal is drawn at random from the data set. We calculate the mean-squared error, bias, and average percent error (|Actual VaR − Approx. VaR| × 100%). We also calculate the variance of the approximation. Finally, as a measure of the rarity of "large approximation errors," we report the 10% and 90% quantiles of the error of the approximation across the 100 pools. The time horizon is 12 months, and 25,000 Monte Carlo simulations are used.

tional time since it allows one to simulate on a very sparse grid but still ultimately achieve a very accurate solution on a fine grid.

Table 7 reports statistics for the accuracy of the approximation using presimulation. The time horizon is 12 months. The approximation is very accurate.

5.6. Approximation Without Low-Dimensional Transformation

The previous numerical solutions of the law of large numbers and central limit theorem rely on an *exact* low-dimensional transformation, which requires a restriction on the interaction between the factors Y^n and V. Although that low-dimensional transformation still covers a wide range of models, for completeness, we now present some methods for the computation of the LLN and CLT for the full class of models where a high-dimensional feature space $\mathcal{Y}$ may make a traditional Cartesian grid infeasible.

We let h_θ be a function of two neural networks, one for defaults and one for prepayments (see Example 2.2). Both neural networks have a single hidden layer of five neurons. The loan-level features are FICO score, LTV ratio, original balance, and initial interest rate. The neural network for defaults also takes as an input the national unemployment rate, while the neural network for prepayments takes both the national unemployment rate and the national mortgage rate as inputs. The model is trained on the subprime mortgage data set. The model parameter estimates are available on request.

When the low-dimensional transformation is not applicable, different methods must be used to efficiently evaluate the LLN and CLT. One approach is to cluster the pool into K clusters using k-means clustering on $Y^1, \ldots, Y^N$. The K centroids are chosen as the grid points. This is a sparse nonuniform grid in a high-dimensional space. Table 8 reports the performance of this method.

Another approach is to perform a low-dimensional transformation with a neural network. This alternative approach is described in Online Appendix G.3. Additional figures showing the performance of these sparse grid methods are included in Online Appendix G.3.

6. Conclusion

This paper develops and tests efficient numerical methods for the analysis of large pools of loans as well as asset-backed securities backed by such pools. For a broad class of statistical and machine learning models of loan-level risk in discrete time, we prove a law of large numbers and a central limit theorem for the pool-level risk. The asymptotics are then used to construct computationally efficient approximations for a large pool. We extensively test our approach on actual mortgage data. The approximation is highly accurate even for relatively small pools (as little as 500 loans). In practice, pools commonly range from a few thousand to

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

hundreds of thousands of loans. Brute-force simulation for large pools can be computationally expensive; our approximation can save several orders of magnitude in computational time. The asymptotic approximation accounts for the full loan-level dynamics, taking advantage of the detailed loan-level information typically available (such as credit score, loan-to-value ratio, initial interest rate, and type of loan). A key feature of our approximation is that its computational cost is constant no matter the dimension of the loan-level data. This feature is desirable since loan-level data can be high-dimensional.

Acknowledgments

The authors thank several anonymous referees for very helpful comments. The authors also thank seminar participants at the University of Illinois at Urbana–Champaign, Imperial College London, London–Paris Bachelier Workshop on Mathematical Finance, and the International Monetary Fund for comments. Special thanks are due to Kostas Spiliopoulos and Stefan Weber for insightful comments. In addition, the authors thank participants at the 2014 Annual INFORMS Meeting and the 2014 SIAM Conference on Financial Mathematics and Engineering, where this paper won the 2014 SIAM Financial Mathematics and Engineering Conference Paper Prize.

Endnotes

[1] Y^n includes both continuous and categorical variables. Categorical variables are encoded as a vector whose elements are each in $\{0,1\}$. Of course, $\{0,1\}$ is a subset of the real line, so $\mathcal{Y} \subseteq \mathbb{R}^{d_Y}$.

[2] The time t can also be included in the systematic factor vector V.

[3] To be concrete, consider a single geographic area and let $\mathcal{U} = \{c, d\}$, where c represents a loan that is current on its payments and d represents default. Specify $U_t^n = \mathrm{c}$ for $t \le 0$. Let $f^H(U_t^n, Y^n) = \mathbf{1}_{U_t^n = \mathrm{d}}$ in (1), where $\mathbf{1}_A$ is the indicator function for an event A. The fraction of loans that defaulted in the $(t-1)$th period in the area takes the form $H_{t-1}^N - H_{t-2}^N$, and can therefore be written as a function of $H_{t-2:t-1}^N$. The loan dynamics can then depend on the defaults in the pool (or geographic area) via H^N.

[4] Here we have implicitly assumed a unit notional for each loan. This paper's results can be extended to the case of loans with different notional sizes as long as the loans' notional sizes are bounded (which they are in practice). The notional size of the loan would be included in the Y^n variable.

[5] For instance, $\langle f, \nu\rangle_{\mathcal{U}\times\mathcal{Y}} = \sum_{u\in\mathcal{U}} \int_{\mathcal{Y}} f(u,y)\nu(u,dy)$ and $\langle f(u,\cdot), \nu(u,\cdot)\rangle_{\mathcal{Y}} = \int_{\mathcal{Y}} f(u,y)\nu(u,dy)$.

[6] A function f is smooth if $f \in C^\infty$.

[7] For instance, consider the random variable X^N where $X^N = 2^N$ with probability $1/N$ and is otherwise 0. $X^N \xrightarrow{p} 0$ but $\mathbb{E}[X^N] \to \infty$ as $N \to \infty$.

[8] In other words, there is some positive constant C_0 such that $|g(x)| \le |x|^k$ for $|x| \ge C_0$ and $k \in \mathbb{N}$.

[9] This is the equivalent of solving the LLN and CLT if $\mathcal{Y}$ was finite-dimensional—i.e., at a finite set of grid points $y_1, \ldots, y_K$. A generalization allowing different sets of grid points for different states u is straightforward.

[10] Provided $\mathrm{Cov}[\bar{\Xi}_{t-1}(u,y_i), \bar{\Xi}_{t-1}(u',y_j)\,|\,V^I]$ and $\mathrm{Cov}[\bar{\Xi}_{t-1}(u,y_i), \bar{E}_{0:t-1}\,|\,V^I]$ for all $y_i, y_j \in \mathcal{R}$ and $u, u' \in \mathcal{U}$, one can calculate the quantities $\mathrm{Cov}[\bar{\Xi}_t(u,y_i), \bar{\Xi}_t(u',y_j)\,|\,V^I]$ and $\mathrm{Cov}[\bar{\Xi}_t(u,y_i), \bar{E}_{0:t-1}\,|\,V^I]$ *in closed form*. Thus, one can march forward through time, calculating the (closed-form) covariance of $\bar{\Xi}_t^{\Psi,I}$ across the grid points $y_1, \ldots, y_K$ at each time $t = 1, \ldots, T$.

[11] "Deal" refers to the securitization of a pool of mortgages into an MBS. An MBS's structure may vary widely: it could be anything from a pass-through to a collateralized mortgage obligation (CMO). A CMO has a tranched structure, with different payment rules for the different tranches.

[12] An interesting question is whether one is guaranteed that the interpolated solution at any point y converges to the correct value as the number of Monte Carlo samples increases and the distance between grid points decreases. If h is continuously differentiable and X takes values in a compact set, such convergence holds (by the application of a Taylor expansion, boundedness of continuous function on a compact metric space, and the dominated convergence theorem). More general cases may require additional technical conditions.

References

Ambrose B, Capone C (1998) Modeling the conditional probability of foreclosure in the context of single-family mortgage default resolutions. *Real Estate Econom.* 26(3):391–429.

Arnsdorf M, Halperin I (2008) BSLP: Markovian bivariate spread-loss model for portfolio credit derivatives. *J. Comput. Finance* 12(2):77–100.

Azizpour S, Giesecke K, Schwenkler G (2018) Exploring the sources of default clustering. *J. Financial Econom.* 129(1):154–183.

Baesens B (2005) Neural network survival analysis for personal loan data. *J. Oper. Res. Soc.* 56(9):1089–1098.

Banasik J, Crook J, Thomas L (1999) Not if but when will borrowers default. *J. Oper. Res. Soc.* 50(12):1185–1190.

Bastos J (2010) Forecasting bank loans loss-given-default. *J. Banking Finance* 34(10):2510–2517.

Bush N, Hambly BM, Haworth H, Jin L, Reisinger C (2011) Stochastic evolution equations in portfolio credit modelling. *SIAM J. Financial Math.* 2(1):627–664.

Capozza D, Kazarian D, Thomson T (1997) Mortgage default in local markets. *Real Estate Econom.* 25(4):631–655.

CoreLogic (2014) Riskmodel. Technical report, CoreLogic, Irvine, CA.

Cvitanic J, Ma J, Zhang J (2012) The law of large numbers for self-exciting correlated defaults. *Stochastic Processes Their Appl.* 122(8):2781–2810.

Dai Pra P, Tolotti M (2009) Heterogeneous credit portfolios and the dynamics of the aggregate losses. *Stochastic Processes Their Appl.* 119(9):2913–2944.

Dai Pra P, Runggaldier WJ, Sartori E, Tolotti M (2009) Large portfolio losses: A dynamic contagion model. *Ann. Appl. Probab.* 19(1):347–394.

Das S, Duffie D, Kapadia N, Saita L (2007) Common failings: How corporate defaults are correlated. *J. Finance* 62:93–117.

Davis M, Rodriguez J (2007) Large portfolio credit risk modelling. *Internat. J. Theoret. Appl. Finance* 10(4):653–678.

Dembo A, Deuschel J, Duffie D (2012) Large portfolio losses. *Finance Stochastics* 1(8):3–16.

Deng Y, Quigley J, Van Order R (2000) Mortgage terminations, heterogeneity, and the exercise of mortgage options. *Econometrica* 68(2):275–307.

Diener N, Jarrow R, Protter P (2012) Relating top-down with bottom-up approaches in the evaluation of abs with large collateral pools. *Internat. J. Theoret. Appl. Finance* 15(2):1250011.

Ding X, Giesecke K, Tomecek P (2009) Time-changed birth processes and multi-name credit derivatives. *Oper. Res.* 57(4):990–1005.

Dunsky R, Zhou X, Kane M, Chow M, Hu C, Varrieur A (2014) FHFA mortgage analytics platform. Technical report, Federal Housing Finance Agency, Washington, DC.

Emmer S, Tasche D (2005) Calculating credit risk capital charges with the one factor model. *J. Risk* 7(2):85–101.

Errais E, Giesecke K, Goldberg L (2010) Affine point processes and portfolio credit risk. *SIAM J. Financial Math.* 1:642–665.

Downloaded from informs.org by [2001:638:807:212:a8af:506c:25a1:b254] on 09 November 2026, at 06:25 . For personal use only, all rights reserved.

Experian (2014) Home-equity indicators with new credit data methods for improved mortgage risk analytics. Technical report, Experian, Costa Mesa, CA.

Fermanian J (2013) A top-down approach for MBS, ABS and CDO of ABS: A consistent way to manage prepayment, default and interest rate risks. *J. Real Estate Finance Econom.* 46(3):480–515.

Fermanian J (2014) The limits of granularity adjustments. *J. Banking Finance* 45(August):9–25.

Frey R, Popp M, Weber S (2008) An approximation for credit portfolio losses. *J. Credit Risk* 4(1):3–20.

Giesecke K, Weber S (2004) Cyclical correlations, credit contagion, and portfolio losses. *J. Banking Finance* 28(12):3009–3036.

Giesecke K, Weber S (2006) Credit contagion and aggregate losses. *J. Econom. Dynamics Control* 30(5):741–767.

Giesecke K, Goldberg L, Ding X (2011) A top-down approach to multiname credit. *Oper. Res.* 59(2):283–300.

Giesecke K, Spiliopoulos K, Sowers RB (2013) Default clustering in large portfolios: Typical events. *Ann. Appl. Probab.* 23(1): 348–385.

Giesecke K, Spiliopoulos K, Sowers RB, Sirignano JA (2015) Large portfolio asymptotics for loss from default. *Math. Finance* 25(1): 77–114.

Glasserman P, Kang W, Shahabuddin P (2008) Fast simulation of multifactor portfolio credit risk. *Oper. Res.* 56(5):1200–1217.

Goodstein R, Hanouna P, Ramirez C, Stahel C (2017) Contagion effects in strategic mortgage defaults. *J. Financial Intermediation* 30(April):50–60.

Gordy M (2003) A risk-factor model foundation for ratings-based bank capital rules. *J. Financial Intermediation* 12(3):199–232.

Gordy M, Lutkebohmert E (2013) Granularity adjustment for regulatory capital assessment. *Internat. J. Central Banking* 9(3):33–70.

Gordy M, Marrone J (2012) Granularity adjustment for mark-to-market credit risk models. *J. Banking Finance* 36(7):1896–1910.

Harding J, Rosenblatt E, Yao V (2009) The contagion effect of foreclosed properties. *J. Urban Econom.* 66(3):164–178.

Hastie T, Tibshirani R, Friedman J (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Springer, New York).

Kang P, Zenios S (1992) Complete prepayment models for mortgage-backed securities. *Management Sci.* 38(11):1665–1685.

Khandani A, Kim A, Lo A (2010) Consumer credit-risk models via machine-learning algorithms. *J. Banking Finance* 34(11): 2767–2787.

Lending Club (2015) Private communication.

Lin Z, Rosenblatt E, Yao V (2009) Spillover effects of foreclosures on neighborhood property values. *J. Real Estate Finance Econom.* 38(4):387–407.

Loterman G, Brown I, Martens D, Mues C, Baesens B (2012) Benchmarking regression algorithms for loss given default modeling. *Internat. J. Forecasting* 28(1):161–170.

Martin R, Wilde T (2002) Unsystematic credit risk. *Risk* 15:123–128.

Mattey J, Wallace N (2001) Housing-price cycles and prepayment rates of us mortgage pools. *J. Real Estate Finance Econom.* 23(2): 161–184.

Merino S, Nyfeler M (2002) Calculating portfolio loss. *Risk* 15(8): 82–86.

Richard S, Roll R (1989) Prepayments on fixed-rate mortgage-backed securities. *J. Portfolio Management* 15(3):73–82.

Sirignano J, Sadhwani A, Giesecke K (2016) Deep learning for mortgage risk. Working paper, Stanford University, Stanford, CA.

Spiliopoulos K, Sowers R (2011) Recovery rates in investment-grade pools of credit assets: A large deviations analysis. *Stochastic Processes Their Appl.* 121(12):2861–2898.

Spiliopoulos K, Sowers R (2015) Default clustering in large pools: Large deviations. *SIAM J. Financial Math.* 6(1):86–116.

Spiliopoulos K, Sirignano JA, Giesecke K (2014) Fluctuation analysis for the loss from default. *Stochastic Processes Their Appl.* 124(7): 2322–2362.

Stepanova M, Thomas L (2002) Survival analysis methods for personal loan data. *Oper. Res.* 50(2):277–289.

Towe C, Lawley C (2013) The contagion effect on neighboring foreclosures. *Amer. Econom. Assoc.* 5(2):313–335.

Vasicek O (1987) Probability of loss distribution. Technical report, Moody's KMV, San Francisco.

Westgaard S, Van der Wijst N (2001) Default probabilities in a corporate bank portfolio: A logistic model approach. *Eur. J. Oper. Res.* 135(2):338–349.

Wu S, Jiang L, Liang J (2012) Intensity-based models for pricing mortgage-backed securities with repayment risk under a CIR process. *Internat. J. Theoret. Appl. Finance* 15(3):1250021.

Zhang X, Blanchet J, Giesecke K, Glynn PW (2015) Affine point processes: Approximation and efficient simulation. *Math. Oper. Res.* 40(4):797–819.