Availability estimation by simulations for systems including logistics

Ajit Rai

UNIVERSITE BRETAGNE LOIRE / MATHSTIC

UNIVERSITÉ DE RENNES 1

THESE DE DOCTORAT DE

L'UNIVERSITE DE RENNES 1
COMUE UNIVERSITE BRETAGNE LOIRE

ECOLE DOCTORALE N° 601
Mathématiques et Sciences et Technologies de l'Information et de la Communication
Spécialité : *Informatique*

Par

Ajit RAI

Estimation de la disponibilité par simulation, pour systèmes incluant des contraintes logistiques (Availability estimation by simulations for systems including logistics)

Thèse présentée et soutenue à Rennes, France, le 09/07/2018
Unité de recherche : IRISA (UMR CNRS 6074), Rennes, France.

Rapporteurs avant soutenance :

Hector CANCELA	Professeur à l'Universidad de la Republica, Montevideo, Uruguay
Sophie MERCIER	Professeure à l'Université de Pau et des Pays de l'Adour, France

Composition du Jury :

Gerardo RUBINO	Directeur de recherche, INRIA, France / Président
Hector CANCELA	Professeur à l'Universidad de la Republica, Montevideo, Uruguay
Sophie MERCIER	Professeure à l'Université de Pau et des Pays de l'Adour, France
Thiago ABREU	Maître de conférences, Université Paris-Est-Créteil, France / Examinateur
Bruno TUFFIN	Directeur de recherche, INRIA, France / Directeur de thèse
Pierre DERSIN	PHM et RAM directeur, Alstom, France / Co-directeur de thèse

Acknowledgements

First and foremost, I am thankful to my Ph.D. supervisor Dr. Bruno Tuffin at INRIA (Rennes, France) and my manager Dr. Pierre Dersin at ALSTOM (Saint-Ouen, France) for their invaluable guidance and support. The streamlined supervision and constructive reviews of Dr. Tuffin, and the resourcefulness of Dr. Dersin were immensely helpful, right from the commencement of this work to the present day. I would also like to thank: Dr. Gerardo Rubino at INRIA and Dr. Rene C. Valenzuela at ALSTOM (Madrid, Spain). All the valuable discussions with them always steered me in the right direction through the course of this work. Without the continuous support of these people, completion of this work would have been a rare event itself.

For my dissertation, I also thank the jury members for their invaluable time and effort to review my work. Especially, I am thankful to Prof. Sophie Mercier, Prof. Hector Cancela and Dr. Thiago Abreau for accepting to review my work and being part of the jury. Their comments, questions, and inputs were crucial for improving the quality and the essence of the dissertation.

I would also like to thank Prof. Enrico Zio and Prof. Francesco Di Maio for introducing me to the world of Monte Carlo simulations. Without their guidance, I would not have realized the fascinating side of uncertainties and probabilities.

My appreciation also goes to the former and current team members of the DIONYSOS team at IRISA/INRIA and the RAM team at ALSTOM. Specifically, I would also like to thank Mme. Fabienne Cuyollaa and Mme. Souvanna Marayphonh, who being the respective team assistants at INRIA and ALSTOM helped me always in all the cumbersome administrative tasks and were very supportive.

Last but not the least, I am thankful to my friends: Sid and Abhinav, whose friendship always brought happiness to me. Special thanks also go to my friends: Federico, Imad, and Louisa. I also thank my friend Justine for being immensely supportive. I am thankful to my family for their unconditional support, love and all the positive vibes they sent me all the way from India. Especially, I thank my parents for their undying love towards me and always inspiring me.

For all the above-mentioned people, words can not do enough justice to summarize their contribution and support. Without their guidance, encouragement, criticism, and concerns, realizing this work would have been impossible.

Contents

List of Figures

List of Tables

Résumé en Français

La compétitivité croissante sur le marché des systèmes ferroviaires de voyageurs, tout en considérant des contrats de performance (c'est-à-dire performance-based contract en anglais) a conduit les fournisseurs de systèmes ferroviaires à se concentrer simultanément sur deux objectifs : réduire le coût des solutions proposées, et respecter les exigences de haute performance et de sûreté de fonctionnement [1]. Les ententes de niveaux de service (*Service-Level-Agreements : SLA* en anglais) font également partie intégrante de ces contrats de performance, dans lesquels les objectifs de disponibilité du système sont stricts, et où le non-respect des niveaux de performance requis mène souvent à des pénalités [1]. La mesure du coût du cycle de vie (CCV) est très utile dans de tels cas pour estimer le coût total des frais encourus durant le cycle de vie d'un système [2] ; et de plus permet de prendre des décisions d'achat bien renseignées [1].

L'estimation du CCV peut inclure, de manière non exhaustive, le coût de l'arrêt du matériel roulant ferrovaire dû à des défaillances du système de signalisation, le coût d'une opération de maintenance corrective sur une voie ferrée bloquant la circulation, le coût d'accidents causant de graves blessures voire mortels, etc. [3]. Les paramètres de Fiabilité, de Disponibilité, de Maintenabilité et de Sécurité (FDMS) sont importants pour déterminer le CCV [3], et la gestion des FDMS doit être considérée dans le projet d'ingénierie de système afin d'atteindre des objectifs de haute performance [4]. L'analyse des FDM traite la mesure des performances (liées à la sûreté de fonctionnement) de systèmes ferroviaires (par exemple matériel roulant ferrovaire, réseaux de communication, compteurs d'essieux, etc.) ainsi que les facteurs les influençant [4]. Plus important encore, les facteurs de FDM constituent une approche stratégique pour l'intégration de la fiabilité, de la disponibilité et de la maintenabilité en utilisant des méthodes, outils et techniques d'ingénierie afin d'identifier, quantifier et analyser les défaillances d'un équipement ou d'un système qui empêchent la réalisation de leurs objectifs [5]. Par conséquent, l'analyse des FDM est un élément intégral pour évaluer et réaliser efficacement des obligations de contrat [3] et est également un des domaines les plus importants pour l'amélioration de la rentabilité [6]. Afin d'étudier de manière précise ces systèmes hautement fiables et complexes, trois aspects doivent être considérés : le choix des paramètres de fiabilité [7], une modélisation mathématique simple mais efficace (incluant la conception d'un système) ainsi qu'une méthodologie d'analyse efficiente [8].

Choix des mesures de fiabilité pour l'analyse de FDM

Il y a plusieurs mesures de performance/fiabilité associées à l'analyse des FDM, comme par exemple le temps moyen d'atteinte de la défaillance (*Mean Time To Failure : MTTF* en anglais), la moyenne des temps de bon fonctionnement (*Mean Time Between Failures : MTBF* en anglais), la durée moyenne de panne (*Mean Time To Repair : MTTR* en anglais), le temps moyen d'indisponibilité (*Mean Down Time : MDT* en anglais), la fiabilité (ou la non-fiabilité), la disponibilité (ou indisponibilité), etc [3, 9, 10]. Le choix des paramètres de fiabilité peut être utile dans différentes situations et dans certains cas il est possible de déduire certains paramètres à partir d'autres [3]. Ce choix requiert également de prendre en considération le cas où la pénalité ou le coût de la défaillance du système dépend de la durée totale des défaillances ou de la fréquence des défaillances [7]. Dans cette thèse, nous nous concentrons sur l'estimation de la fiabilité (ou non-fiabilité) dans le cas de réseaux statiques (où le temps ne joue aucun rôle) et la disponibilité (ou indisponibilité) asymptotique pour des systèmes dynamiques (sous des hypothèses markoviennes). La fiabilité d'un système est définie comme la probabilité de fonctionner tel que requis pendant un intervalle de temps donné (par exemple t_1, t_2), dans des conditions données [2]. Ce paramètre est utile pour les réseaux statiques dans lesquels le temps ne jouè aucun rôle. D'autre part, la disponibilité (ou au contraire l'indisponibilité) asymptotique d'un système est la fraction de temps durant laquelle le système est dans un état de fonctionnement (ou au contraire défaillance) lorsque le temps tend vers l'infini (c'est-à-dire en régime asymptotique) [2, 10]. Ainsi nous basons notre choix sur l'estimation de la disponibilité (ou indisponibilité) asymptotique pour les systèmes dynamique, où l'état du système change au cours du temps.

Technique de modélisation

Un autre aspect important pour l'analyse des FDM est une technique de modélisation efficiente. Cela est nécessaire pour comprendre comment un système réel particulier fonctionne et quelles hypothèses peuvent être faites pour mathématiquement modéliser un tel système [8]. La technique de modélisation doit être simple et suffisamment représentative du système réel [8], et de manière plus importante, soluble. Pour l'estimation de la fiablité des réseaux statiques, où le temps ne joue aucun rôle, les techniques de modélisation graphique fournissent des modèles plus simples qui sont faciles à valider [11]. Pour les systèmes dynamiques, où le facteur temps intervient, la modélisation des réseaux de Petri rend possible la visualisation et la modélisation de comportements complexes qui comprennent la concurrence, la synchronisation et le partage des ressources tout en représentant une description concise [12].

Méthodologie d'analyse

L'analyse de modèles mathématiques est la partie la plus importante de l'étude de systèmes complexes et est le sujet principal de cette thèse. Les techniques analytiques et numériques deviennent rapidement inutiles à cause des exigences rigoureuses qu'elles nécessitent en terme de complexité et (ou) d'hypothèses sur le modèle [13, 14]. Elles deviennent également inefficaces dès lors que les dimensions mathématiques deviennent importantes [13, 14]. La simulation de Monte Carlo dans sa forme standard peut être utilisée pour simuler des modèles mathématiques à larges dimensions. Elle est basée sur une méthode d'approximation statitistique [13, 14] et fournit une alternative plus pratique. Plus précisément, les intégrales mutli-dimensionnelles ne peuvent être évaluées que numériquement, et de ce fait la simulation de Monte Carlo est plus pratique que la méthode déterministe [13, 14, 15, 16]. Cependant, bien que la simulation de Monte Carlo dans sa forme standard soit facile pour estimer les paramètres de fiabilité auxquels nous nous intéressons (par exemple l'indisponibilité asymtpotique) pour des modèles larges, elle présente également des limites. Particulièrement lorsque l'événement d'intérêt est rare (par exemple la défaillance d'un système hautement fiable), la simulation de Monte Carlo dans sa forme standard souffre d'inefficacité [13, 14]. La taille d'échantillon et par conséquent le temps de calcul doivent être augmentés quand la rareté d'un événement d'intérêt augmente (c'est-à-dire lorsque la probabilité de l'occurrence d'un événement d'intérêt diminue). La simulation d'événements rares est un terme générique qui couvre le domaine de recherche d'estimation de paramètres spécifiques quand la probabilité d'un événement d'intérêt est très faible.

Simulations d'Evénements Rares

Les événements rares, comme leur nom l'indique, sont des événements dont les occurrences sont très peu probables. Le terme *très peu* dépend du contexte et du domaine d'application [14]. De très nombreux domaines présentent des événements d'intérêt rares mais critiquement importants (par exemple, les paramètres de sûreté de fonctionnement [17], les probabilités de dépassement de tampon dans les réseaux de télécommunication [17, 18], la fréquence de dommage au cœur dans l'analyse de risques et de sécurité de centrales nucléaires [19], etc). Ces probabilités sont associées à la défaillance de systèmes ou infrastructures critiques spécifiques pouvant entraîner la perte de services essentiels, la perte catastrophique de vies humaines, instabilité financière, etc.

Les systèmes ferroviaires de voyageurs sont généralement constitués de composants hétérogènes hautement fiables, et afin de satisfaire les critères de fiabilité, on utilise la redondance à des niveaux hiérarchiques différents (composant, produit, sous-système, équipement, etc.) [1]. Cela rend le système hautement fiable et en fait une structure complexe. Comme expliqué précédemment, la simulation de Monte Carlo dans sa forme standard devient inefficace dans l'estimation des paramètres de fiabilité dans le contexte d'événements rares. Il existe de nombreuses techniques d'accélération (ou techniques de réduction de la variance) qui ont été proposées pour augmenter la fréquence d'événements d'intérêt rares. Il faudrait, par ailleurs, prendre une taille d'échantillon inacceptablement grande pour obtenir suffisamment d'échantillons positifs (utiles) pour l'estimation de n'importe quel paramètre lié à un événement d'intérêt rare [17] et cela rendrait le temps de calcul peu pratique. Les deux principales techniques d'accélération qui ont reçu une attention considérable dans ce contexte sont : l'échantillonnage préférentiel (*Importance Sampling* en anglais) [14] et la technique du Splitting [20]. L'idée générale derrière la méthode du Splitting est l'utilisation d'un mécanisme de sélection pour favoriser l'échantillon de chemins considérés comme menant à des événements rares [21]. L'autre méthode pour l'accélération d'événements rares qui a mérité une attention particulière et fait l'objet de notre travail actuel est l'échantillonnage préférentiel.

L'idée générale derrière l'échantillonnage préférentiel est de changer les lois de probabilité (procédé appelé changement de mesure de l'échantillonnage) du modèle utilisé pour la simulation afin d'augmenter l'occurrence d'événements d'intérêt rares [13, 22]. Les résultats sont ensuite utilisés pour calculer les paramètres d'intérêt dans les conditions des lois de probabilité d'origine en compensant le biais. Cette compensation est faite en multipliant l'estimateur par un facteur de correction nommé *fonction de vraisemblance* afin d'obtenir des estimateurs sans biais des paramètres d'intérêt [13]. Ce concept a pour origine le travail fait sur l'échantillonnage aléatoire (nommé la méthode Monte Carlo) de John von Neumann et Stanislaw Ulam dans le projet Manhattan mené durant les années 1940 et est utilisé afin de résoudre des problèmes de physique nucléaire [23, 24].

L'échantillonnage préférentiel est maintenant une technique avancée de réduction de variance qui a été appliquée avec succès en conjonction avec la simulation de Monte Carlo afin d'obtenir de précises estimations dans une grande variété de domaines, dont les problèmes de files d'attente et de fiabilité [10, 17, 25, 26]. Cependant, le changement de mesure dans l'échantillonnage préférentiel n'est pas connu a priori et la difficulté principale est de déterminer un bon changement de mesure. Le terme ambigu *bon* est en fait le changement de mesure qui réduit la variance de l'estimateur final, le terme *optimal* fournissant évidemment un estimateur avec une variance zéro [13, 14]. Le changement de mesure optimal est également inconnu a priori, et même s'il était connu, il serait difficile d'effectuer l'échantillonnage à partir de celui-ci [17]. Une sélection inappropriée du changement de mesure de l'échantillonnage préférentiel peut donner une estimation complètement incorrecte. Cependant, un changement de mesure optimal (ou même bon) de l'échantillonnage préférentiel permet une considérable amélioration en terme de temps de calcul, de réduction de la variance, et ainsi de précision de l'estimateur. De ce fait, cette thèse apporte plusieurs contributions en matière d'estimation des paramètres d'intérêt dans le contexte d'événements rares, comme nous l'expliquons dans la partie suivante.

Objectifs et Contributions de la Thèse

En raison des différents problèmes mentionnés plus haut, l'objectif principal de ce travail est de proposer des méthodes d'application pratiques pour l'échantillonnage préférentiel. Le travail contribue à proposer et à étendre les techniques d'approximation/d'estimation automatisées. D'une part, nous proposons une méthode qui imite un échantillonnage à partir de la mesure de la zéro-variance de l'échantillonnage préférentiel. D'autre part, nous proposons une autre méthode, qui elle estime le changement de mesure optimal de l'échantillonnage préférentiel dans une pré-simulation, et utilise ce changement de mesure pour l'échantillonnage préférentiel dans la simulation principale. Cela permettra de fournir une estimation précise (c'est-à-dire la réduction de la variance) des paramètres d'intérêt à un coût raisonnable (c'est-à-dire le temps de calcul). Les modèles mathématiques pris en considération dans ce travail sont à grande échelle à la fois pour les réseaux statiques et pour les systèmes dynamiques avec des contraintes logistiques (sous des hypothèses markoviennes). Ces modèles peuvent représenter les besoins de gestion des FDM de réels systèmes ferroviaires de voyageurs. Cette thèse tente de résoudre ces problèmes et comprend deux principales contributions.

Estimation de la fiabilité des réseaux statiques dans les cas des défaillances de nœuds

Le problème de fiabilité des réseaux statiques dont traite cette thèse, concerne l'estimation de la probabilité qu'un ensemble donné de nœuds dans un modèle graphique soient connectés [11]. Dans un tel cas, chaque composant individuel (lien ou nœud) pourrait être dans un état de fonctionnement ou de défaillance selon ses propres probabilités [11]. Le cas où les liens sont les éléments défaillants est essentiel dans de nombreuses applications et a été largement étudié [27]. Cependant, il existe un large éventail d'applications où les nœuds sont les composants défaillants, par exemple dans les modèles de survivalité du réseau [28]. Cela requiert une adaptation des méthodes existantes au cas des défaillances de nœuds [11]. Dans ce contexte, une défaillance de nœud signifie que le nœud devient non-fonctionnel et ses liens associés deviennent alors inutiles [11]. Dans le problème de fiabilité à deux terminaux ou source-à-terminal (s-t), deux nœuds du graphique sont fixes et la fiabilité du réseau est définie comme la probabilité d'obtenir un chemin entre ces deux nœuds [11]. Dans une telle analyse, lorsqu'un nœud est défaillant, un plus grand nombre de chemins s-t deviennent non-fonctionnels que dans le cas d'une défaillance de lien (selon le degré du nœud) [11]. Ainsi, la fiabilité d'un réseau serait affectée plus sévèrement dans le cas d'une défaillance de nœud [11].

Il faut également prendre en considération le fait que les défaillances de réseaux statiques si hautement fiables sont rares, et nous utilisons ainsi ici les techniques d'échantillonnage préférentiel. Dans le contexte des problèmes mentionnés ci-dessus, cette thèse contribue donc à l'estimation efficiente de la fiabilité/non-fiabilité des réseaux statiques des manières suivantes.

- **Prolongation de la méthode d'échantillonnage à partir de la mesure de la zéro-variance de l'échantillonnage préférentiel :** La méthodologie de l'approximation de l'estimateur à variance zéro dans l'échantillonnage préférentiel dans le cas de la défaillance des liens [29] est étendue au cas des défaillances de nœuds (au lieu des liens). Pour cette méthode, nous adaptons l'algorithme de *Ford-Fulkerson maxflow mincut* pour considérer le flux (*flow* en anglais) selon la capacité (calculée en fonction des probabilités de défaillance) des nœuds et obtenir les *mincuts* avec les probabilités maximales. La méthode estime la non-fiabilité entre deux nœuds (la source et le terminal). La connectivité entre le nœud source et le nœud terminal définit la défaillance complète du réseau statique dans le contexte d'événements rares. Pour les estimateurs, nous obtenons une erreur relative bornée (ErrRB) et dans certains cas des erreurs relatives disparaissant dans les régimes asymptotiques (en terme de rareté).

- **Application dans les systèmes réels :** Dans les systèmes ferroviaires de voyageurs (urbains) d'Alstom, le sous-système, dont l'objectif est d'assurer le fonctionnement de la communication principale, est appelé *Data Communication System* (DCS) [11]. Le DCS utilise deux réseaux parallèles qui permettent à des équipements situés dans différentes stations (ou sur les voies) de communiquer avec les équipements de contrôle centraux [11]. Le DCS est configuré pour que tous les équipements aient deux réseaux (rouge et bleu) en redondance active afin d'envoyer des messages simultanés sur les deux réseaux (rouge et bleu) [11]. Si l'un des deux réseaux est en panne, l'équipement peut utiliser l'autre réseau [11]. La fiabilité du DCS est la probabilité que tous les messages entre tous les équipements et contôleurs soient transmis avec succès [11]. Dans cette thèse, nous considérons la fiabilité ou la non-fiabilité entre la source et le terminal. Nous considérons un train à l'arrêt comme le nœud-source, et le contrôleur de la zone comme le nœud-terminal. Nous illustrons également l'utilité de l'algorithme proposé sur un modèle graphique de système DCS à 164 nœuds d'ALSTOM. Dans le cas de cette application réelle, nous observons également la propriété de l'erreur relative bornée (ErrRB).

Estimation de l'indisponibilité asymptotique des systèmes dynamiques avec contraintes logistiques

Dans cette partie, cette thèse contribue [30] à l'estimation de l'indisponibilité asymptotique (*steady-state unavailability* en anglais) pour les systèmes markoviens hautement fiables. Nous modélisons les systèmes sous forme de réseaux de Petri stochastiques (*Stochastic Petri Nets : SPNs* en anglais) Markovien. Ces derniers incluent des protocoles de logistique et de maintenance. Certains aspects logistiques importants inclus dans nos exemples sont : la disponibilité des pièces de rechange, une équipe de restauration dans un dépôt, une inspection minutée des composants pour toute défaillance et le temps de leur déplacement vers le site pour réparation/inspection.

La contribution principale est un algorithme de pré-simulation basé sur l'optimisation de distance de l'entropie croisée [31] afin d'approcher le changement de mesure optimal pour l'échantillonnage préférentiel appliqué aux transitions d'intérêt dans les modèles SPN (au sein de la même famille paramétrique). La densité de l'échantillonnage préférentiel est celle qui se rapproche le plus de la densité de la variance zéro, en terme de distance d'entropie croisée, c'est aussi la densité pour laquelle la variance asymptotique (en terme de rareté) de l'estimateur est minimale [32]. Nous exploitons également la structure régénérative [10, 33] des chaînes de Markov à temps continu (*Continuous Time Markov Chain : CTMC* en anglais) sous-jacentes des modèles SPN [34] markoviens. La simulation principale utilise les taux de l'échantillonnage préférentiel obtenus à partir de l'algorithme de pré-simulation pour estimer l'indisponibilité asymtotique. Les résultats pour différents exemples montrent un gain (quantifié par le ratio de la «work-normalized variance» [35] de la méthode Monte Carlo dans sa forme standard et de la méthode d'échantillonnage préférentiel proposée ici) et montrent également la réduction de la variance comparée aux simulations de la méthode Monte Carlo dans sa forme standard. Les contributions sont :

- **Evolution progressive de la rareté de chaque problème :** La première étape de l'algorithme proposé dans cette partie consiste à réduire le problème d'origine (donné par les taux de transitions d'un modèle SPN) à un sous-problème moins rare. Cela est possible en augmentant les taux de défaillance des composants afin de créer un système instable avec des défaillances non-rares. Ensuite la rareté du problème est progressivement augmentée à chaque étape de la pré-simulation (en réduisant le taux de défaillance des composants) jusqu'à ce que le problème d'origine soit atteint (c'est-à-dire les mêmes taux de défaillance des composants du problème d'origine). La topologie du modèle reste la même, cependant les dynamiques probabilistes intrinsèques impliquées sont changées à chaque étape. La répartition du problème d'origine est effectuée selon le nombre d'étapes (S) spécifiées pour la pré-simulation. Cela signifie que le problème d'origine est décomposé en sous-problèmes rares et facilement solubles (dans lesquels les contraintes sont de simuler à chaque étape le nombre de cycles) et définis par le choix de S, problèmes qui sont résolus à chaque étape. A chaque étape, ces sous-problèmes sont considérés comme la mesure de probabilité originale et les taux d'échantillonnage préférentiel sont ceux obtenus à l'étape précédente. Dans l'étape finale de pré-simulation, le problème d'origine est résolu en utilisant le vecteur des taux d'échantillonnage préférentiel obtenus lors de l'étape précédente.

- **Le choix de départ du changement de mesure pour l'échantillonnage préférentiel pour la pré-simulation :** Durant la première étape de simulation, le choix de départ du changement de mesure pour l'échantillonnage préférentiel (défini par le vecteur des taux de transitions dans les modèles SPN markoviens dans notre cas) est souvent spécifique à chaque problème dans les algorithmes d'optimisation de l'entropie croisée. La méthodologie que nous proposons ici crée une séquence de problèmes moins rares à résoudre (comme expliqué dans la contribution ci-dessus). Lors de la première étape, le vecteur de taux initial pour l'échantillonnage préférentiel (c'est-à-dire lorsque le changement de mesure est appliqué) pour les transitions d'intérêt est considéré comme étant le même que le vecteur de taux du problème devant être résolu à cette même étape. Cette approche fait de la première étape de la pré-simulation une méthode de simulation régénérative de Monte Carlo dans sa forme standard, où la fonction de vraisemblance des transitions d'intérêt respectives est égale à 1. L'équation principale de l'algorithme reflète alors la contribution de ces transitions respectives du modèle SPN markovien vers l'événement d'intérêt non-rare du problème devant être résolu à cette étape. Avec cette approche heuristique, il n'est pas nécessaire de spécifier le changement de mesure de l'échantillonnage préférentiel par l'utilisateur pour la première étape.

- **Facilité d'utilisation de l'algorithme :** La méthode que nous proposons permet également à l'utilisateur d'optimiser différentes transitions d'intérêt dans un modèle SPN markovien uniquement, ou dans des groupes. L'optimisation individuelle fournit des solutions sous la forme de taux d'échantillonnage préférentiel optimisés pour chaque transition uniquement. Une optimisation groupée fournit quant à elle une solution sous la forme d'une valeur commune de taux d'échantillonnage préférentiel pour les transitions spécifiques regroupées ensemble. Des groupes mutliples peuvent également être formés.

Cette approche de regroupement est intéressante dans le cas de très grands systèmes, par exemple si nous considérons un système avec plusieurs types de composants (type A, B, C, etc) qui peuvent mener le système à un état de défaillance (événement d'intérêt) dans un cycle régénératif. Dans un nombre limité de cycles de pré-simulation, il est possible que tous les composants de chaque type (A, B ou C) sur lesquels l'échantillonnage préférentiel est utilisé ne soient pas échantillonnés dans la simulation d'événements discrets (c'est-à-dire *Discrete Event Simulation : DES* en anglais). Cela peut donner comme solution une valeur de zéro indésirable à la méthode d'optimisation de l'entropie croisée pour les transitions de défaillance de certains composants. Cependant le regroupement peut aider à calculer une valeur commune de taux pour l'échantillonnage préférentiel, pour toutes les transitions de défaillance goupées ensemble. Dans ce cas, même si la transition d'un composant n'est pas échantillonnée, cela peut quand même fournir une valeur commune comme solution, grâce à l'échantillonnage possible des autres transitions du même groupe. L'approche de groupement aide également à la réduire le bruit statique dans des problèmes à grandes dimensions abordés dans ce travail. Un autre avantage du groupement constaté à partir des résultats empiriques des exemples de cette thèse, est la légère réduction du temps de calcul lorsqu'il est comparé individuellement à l'optimisation multidimensionnelle de chaque transition. Cette approche fournit une contribution plus pratique pour une facilité d'usage. La stratégie de regroupement peut être basée sur le jugement technique et la connaissance pratique d'un modèle SPN. Les différentes transitions de défaillance des composants peuvent être regroupées ensemble sur la base de similarité en terme de modes de défaillance, de types de composants, ou de niveaux hiérarchiques (composant, produit, sous-système, équipement, etc).

A partir des contributions de la thèse citées plus haut, nous pouvons conclure que cette thèse propose des méthodes qui peuvent être utilisées efficacement pour l'application automatisée de l'échantillonnage préférentiel aux modèles de fiabilité statique et dynamique (Markovien). Les résultats des exemples obtenus dans cette thèse montrent une réduction importante de la variance (ainsi que la propriété ErrRB souhaitée) dans le régime asymptotique (c'est-à-dire lorsque la probabilité d'une défaillance du système tend vers zéro). Cependant, comme la recherche scientifique est en évolution continue, il y a toujours une marge pour l'amélioration.

Pour les future travaux, un des aspects que nous considérons comme important est l'utilisation de l'entropie croisée proposée dans ce travail pour les SPNs non-markoviens, où les différentes distributions peuvent devenir n'importe quelle distribution générale (par exemple, Weibull, triangulaire, log-normal, etc.). Dans de tels cas, des «A-cycles» pourraient être utilisés pour représenter le ratio de la disponibilité asymptotique [17, 36]. La méthode de traitement par lots (c'est-à-dire Batch Means method en anglais) pourrait aussi être utilisée pour estimer la variance [17, 36]. Cette intégration de l'entropie croisée pour les SPNs non-markovien utilisant des «A-cycles» peut être très utile pour les praticiens de la FDM afin de modéliser et d'analyser les systèmes ferroviaires réels de voyageurs avec moins d'hypothèses.

La thèse est organisée selon les chapitres suivants (en anglais) : dans le chapitre 1, nous abordons les motivations de la thèse et les contributions de cette dernière. Dans le chapitre 2, nous montrons la nécessité de la simulation d'événements rares, nous décrivons l'échantillonnage préférentiel, les possibilités d'application de l'échantillonnage préférentiel, ainsi que les mesures de robustesse (en terme de précision, d'efficacité et de performance). Dans le chapitre 3, nous proposons notre méthodologie et ses résultats pour l'estimation de la fiabilité des réseaux statiques et des systèmes DCS réels. Dans le chapitre 4, nous illustrons notre méthode pour l'estimation de l'indisponibilité asymptotique via l'utilisation de l'optimisation de la distance d'entropie croisée pour les transitions (défaillances) dans les SPNs markoviens (avec contraintes logistiques). Dans le dernier chapitre 5, nous présentons nos conclusions ainsi que des perspectives de recherches futures.

Chapter 1

Introduction

1.1 Motivations

In the contemporary world, urban passenger rail systems are large-scale systems with highly redundant structures at different hierarchical levels. Rail system suppliers such as ALSTOM need to commit contractually to service-level-agreements (SLA), including stringent system availability targets, to remain competitive and advance in the market. A non-adherence to the performance levels often leads to penalties. To meet such strict contractual obligations, rail system suppliers need to minimize the cost of the offered solutions while also satisfying the high performance and dependability requirements [1]. A solution's life cycle cost (LCC) in these so-called *Performance-Based Contracts* is an appropriate measure based on which rail system suppliers can make purchasing decisions [1].

LCC can be defined as the total cost incurred during the life cycle [2] of a system. It can include, but not limited to, the costs of a rolling stock stopped due to signaling system failures, a corrective maintenance operation on a rail track that clogs the traffic, accidents that can cause serious injuries or be fatal, and so forth [3]. Reliability, Availability, Maintainability, and Safety (RAMS) parameters/metrics are essential to determine the LCC [3], and RAMS management is useful in system engineering projects to meet the high-performance targets [4]. RAM (Reliability, Availability, and Maintainability) analysis deals with the performance measures related to the dependability of rail systems (e.g., rolling stocks, communication networks, axle counters, etc.) and the factors affecting it [4]. Most importantly, RAM factors constitute a strategic approach for integration of reliability, availability, and maintainability, by use of methods, tools and engineering techniques to identify, quantify, and analyze equipment or system failures that prevent the achievement of RAM objectives [5]. Thus, RAM analysis is an integral part of assessing and meeting the contractual obligations effectively [3] and is also one of the most significant areas for profitability improvement [6]. To accurately study such highly reliable and complex systems, there are three main important aspects, namely, the choice of reliability metrics [7], a simple yet effective mathematical modeling (including the designing of a system) and an efficient analysis methodology [8].

The choice of reliability metrics can be useful for different situations and in some cases it is possible to deduct some metrics from the others [3]. This choice also requires putting into consideration if the main penalty or cost of the system failure depends on the total duration of failures or the frequency of failures [7]. There are several reliability metrics associated with RAM analysis, e.g., Mean Time To Failure (MTTF), Mean Time Between Failures (MTBF), Mean Time To Repair (MTTR), Mean Down Time (MDT), reliability (or unreliability), availability (or unavailability), etc. [3, 9, 10]. In this thesis, the focus is on computing the reliability/unreliability in case of static networks (where time plays no role) and the steady-state availability/unavailability for dynamic systems (under Markovian assumptions).

Another important aspect is an efficient modeling technique. It is needed to understand how a particular real system operates and the specific assumptions that can be made to model such a system mathematically [8]. The modeling technique needs to be simple yet sufficiently representing the real system [8] and most importantly, solvable. For static networks' reliability estimation, where time plays no role, graph modeling techniques provide simpler models that are easy to validate [11]. For dynamic systems, where the time factor intervenes, Petri Nets (PNs) modeling makes it possible to visualize and model complex behaviors comprising concurrency, synchronization, and resource sharing while also providing a condensed description [12].

Analysis of the respective mathematical models is the essential part of studying complex systems and is the main subject of this thesis. Techniques like direct computations (also called analytic techniques) and standard numerical analysis become quickly useless due to their stringent requirements regarding complexity and (or) assumptions on the model [14]. These methods also suffer from inefficiency as soon as the mathematical dimensions of the problem become large [14]. Standard Monte Carlo simulations (hereafter referred as standard MC simulation) are useful for simulating high dimensional mathematical models and use statistical approximation techniques [14], providing a more practical alternative. Specifically, multi-dimensional integrals fall into the category of problems that we can only evaluate numerically. For solving multi-dimensional integrals, MC methods are more practical than deterministic techniques [14, 15, 16]. However, even though standard MC simulations are easy to compute reliability metrics of interest (e.g., steady-state unavailability) for large models, but they also have limitations. Especially, when the *event of interest* (EOI) is rare (e.g., a system failure of a highly reliable system), they suffer from inefficiency [14]. In such cases, it is required to increase the sample size (and consequently the computation time) as the rarity of EOI increases (i.e., the probability of the occurrence of EOI decreases). *Rare events simulation* is an umbrella term covering the field of research for estimation of specific metrics when the probability of the EOI is very low.

Rare events, as the name itself indicates, are events that have very small probabilities of occurrence. The term small depends on the context and the application domain [14]. There are many fields where these EOI are rare but critically important. Some cases where the study of rare events finds use are: measures of dependability in reliability models [17], the buffer overflow probabilities in telecommunication networks [17, 18], the core damage frequency in risk and safety analysis of civilian nuclear power plants [19], etc. We associate these probabilities with the failure of particular critical systems or infrastructures that can lead to loss of essential services, catastrophic loss of human lives, financial instability, etc. Passenger rail systems typically constitute of heterogeneous very reliable components [1]. To satisfy the dependability requirements, rail system suppliers use redundancy at different hierarchical levels (component, item or subsystem levels) [1]. These aspects make the entire system highly reliable and a complex structure. As previously stated, standard MC simulations become inefficient for estimating reliability metrics in rare events context. There are several *acceleration techniques* (also known as variance reduction techniques) that have been proposed to increase the frequency of rare events under consideration. Otherwise, it may take unacceptably large sample sizes and computation time to get enough positive (useful) samples for estimation of any metrics related to the rare EOI [17]. The two main acceleration techniques that have received considerable focus in this context are Importance Sampling [14] and the *Splitting* technique [20].

The generic idea behind the *Splitting* method is to use a selection mechanism to favor the sample paths (sequence of states visited in a replication) deemed likely to lead to rare events [21]. The main idea is to decompose the sample paths leading to the rare events into smaller subpaths having a higher probability, encourage the realizations that follow these subpaths towards rare events by allowing them to reproduce and to discourage the realizations that do not follow these subpaths to rare events with some positive probability [21].

The other method for rare events simulation that has also received a considerable focus and is the subject of the current work is Importance Sampling (IS). The general idea behind IS is to change the probability laws (called as performing a change of measure) driving the model in a simulation to increase the occurrence of the rare EOI and thus the property is more likely to be seen [22]. The results are then used to calculate the target metric under the original probability laws by compensating for the differences. A correction factor, called as the *likelihood ratio*, is used to compensate the bias by multiplying the estimator with it, and thus obtaining an unbiased estimator of the target metric. The concept of the IS method originated from the work done on random sampling (i.e., the Monte Carlo method) by John von Neumann and Stanislaw Ulam in the Manhattan project during the 1940's and was used to solve problems in nuclear physics [23, 24].

IS is now an advanced class of variance reduction techniques that have been successfully applied in conjunction with MC simulations to obtain accurate estimations in a wide variety of fields, including queueing and reliability problems [10, 17, 25, 26]. However, the change of measure in IS used for the sampling during a simulation is unknown a priori, and the main difficulty is to determine a *good* IS change of measure. The ambiguous term *good* is the change of measure that reduces the variance of the final estimator, *optimal* one providing an estimator with zero variance [14]. The optimal change of measure is also unknown a priori, and even if it were known, it would be difficult to do the sampling from it [17]. An inappropriate selection of the IS change of measure can give an incorrect estimation. However, an optimal (or a good) IS change of measure allows a considerable improvement in terms of the computation time, variance reduction, and the accuracy of the estimator. In regard to this, the thesis makes specific contributions, as presented in the next section.

1.2 Thesis Contributions

In the wake of the above-mentioned discussions, the primary focus of the current work is to propose practically applicable IS methodologies. The work focuses on proposing and extending automated approximation/estimation techniques to form a zero-variance IS scheme or to obtain the optimal IS change of measure and use it in main simulations. We show the proposed methods provide accurate estimation (i.e., with variance reduction) of the reliability metrics at a reasonable cost (i.e., computation time). The mathematical models considered in the current work are large-scale static networks, as well as dynamic systems with logistics (under Markovian assumptions), that can resolve the needs of RAM management of real passenger rail systems. The thesis attempts to address these problems and makes the following contributions broadly classified into two main parts:

1. **Static Networks:** In this part, the thesis contributes [11] to efficiently estimate static network reliability (or contrarily the unreliability) and extends the work on approximate zero-variance IS where in this case, nodes in a graph model are considered to be components of failure. The problem of link failures has been extensively studied, however, the case of node failures is critical in our case. The specific contributions are as follows:

    1.1. **Extension:** We extend the approximate zero-variance IS methodology for link failure case as given in [29] to the case of node failures (instead of links) here. The adapted Ford-Fulkerson *maxflow-mincut* algorithm in this thesis considers flow through nodes (according to the capacity of nodes assigned as per the probability of failure of the respective nodes) and find the mincuts with maximal probability. The method estimates the source(s)-terminal(t) unreliability (i.e., the probability of s-t being disconnected) in context of rare events. We observe the bounded relative error (in some cases, vanishing relative error too) property from the empirical results. Thus, we also obtain a significant variance reduction and considerable gain, when comparing to standard MC simulations.

    1.2. **Application to a real system:** We also illustrate the usefulness of the proposed algorithm on a real system of ALSTOM, called as the *Data Communication System* (DCS) having 164 nodes in a graph model. For this real application also, we observe the bounded relative error property.

2. **Dynamic Systems with Logistics:** In this part, the thesis contributes [30] to estimate the cumulative steady-state unavailability of Highly Reliable Markovian Systems (HRMS). We model the systems as Markovian Stochastic Petri Nets (SPNs), and usually, also include complex protocols of logistics and maintenance. The main contribution is a pre-simulation algorithm based on Cross-Entropy (CE) optimization to approximate the optimal IS change of measure (i.e., the vector of rates for IS) for transitions of interest in the SPN models within the same parametric family. We also exploit the regenerative structure of the underlying continuous-time Markov chains (CTMC) of the Markovian SPN models, and apply IS on the failure transitions of the

model (i.e., the transitions of interest). The main simulation uses the IS rates obtained from the pre-simulation algorithm to estimate the cumulative steady-state unavailability. Results for different examples show a considerable gain and variance reduction compared to standard regenerative MC simulations. The specific contributions are as follows:

2.1. **Progressive rarity shift of the problem:** In the proposed algorithm for this part, the target problem is first reduced to a less rare sub-problem (by increasing the failure rates of components) to create an unstable system with non-rare failures. In the subsequent stages, the rarity of the sub-problem is progressively increased (by decreasing the failure rates of components) at each step of the pre-simulation until it reaches the original problem. The topology of the model remains the same; however, we shift the inherent probabilistic dynamics involved, at each stage. We perform this shifting of the original problem into different less rare sub-problems according to the number of stages S used for pre-simulation (within constraints of the number of regenerative cycles simulated at a stage). In the final pre-simulation stage, the original problem is solved using the set of IS rates obtained from the penultimate stage of pre-simulation.

2.2. **Starting choice of IS change of measure for pre-simulation:** The starting choice of IS change of measure (for example, parameter vector including rates in a Markov model) for the first pre-simulation stage is usually problem dependent in CE algorithms. The proposed methodology under this part forms a series of less rare sub-problems to be solved (as explained in contribution 2.1 above). At the first stage, the initial IS vector of rates (i.e., the change of measure applied) for the transitions of interest is considered to be the same as the vector of rates of the shifted problem (i.e., the sub-problem) to be solved at that stage. This approach makes the first pre-simulation stage as a standard regenerative MC simulation, where likelihood ratio of the respective transitions is equal to one. In this case, the algorithm's main equation captures the contribution of the respective transitions towards the EOI for the sub-problem solved at the first stage. With this heuristic, there is no requirement to specify the IS change of measure by the user for the first stage.

2.3. **Usability:** Our proposed methodology also allows the user to be able to optimize different transitions of interest in a Markovian SPN model individually or in groups together. Individual optimization provides the solution in the form of optimized IS rates for each transition uniquely. A grouped optimization provides the solution in the form of a common value of IS rates for the specific transitions in that group. It is also possible to form multiple groups.

     The approach of grouping is interesting in case of large systems. For example, let us consider a system has several types of components (Type A, B, C, etc.) that contribute towards the rare EOI in a sample path/regenerative cycle, and within each type, there are several components with possible failure transitions. In a limited number of cycles, it is possible that all components' failure transitions within each type A, B or C (on which IS change of measure is applied) are not sampled in the discrete event simulation (DES). Thus, it can give the undesired value of zero for those transitions as a solution to the CE stochastic approximation problem. However, grouping can help to compute a common value of IS rates, for all failure transitions in the specific group. In such a case, even if a component transition is not sampled, it would still provide a common value as a solution, thanks to the possible sampling of other transitions in the same group. Grouping also helps in reducing the statistical noise in the high dimensional problems considered in this work. Another aspect of grouping is that it also reduces the computation time slightly when comparing to the multi-dimensional optimization of each transition individually, as observed from the empirical results of the examples considered here. This approach provides a more practical contribution towards the ease of use. We can base the grouping strategy on the engineering judgment and knowledge of an SPN model. Also, we can group the various failure transitions of components by the similarity in modes of failure, types of components or the hierarchical levels (component, item or subsystem levels).

1.3 Thesis Organization

The introduction of the thesis here aims to emphasize the use of advanced, even if complex, simulation techniques for RAM analysis. The motivation of the entire thesis is to provide methodologies for highly efficient estimations of reliability metrics, which are mathematically sound and also easily applicable to real systems. The motivation/choice of using IS techniques is to analyze highly reliable large-scale systems where EOI are rare, and standard MC simulations do not guarantee accurate estimation. The thesis organization is as follows:

- **Chapter 2: A General Introduction to Rare Events Simulation**
  In this Chapter 2, we aim to provide the background for rare events simulation, where standard MC method is highly limited in terms of accuracy. It also provides a brief introduction to the two main variance reduction techniques, including the one focused in the current thesis (Importance Sampling). In the chapter, we also explain the general problems in the application of IS methods (unknown optimal change of measure) which the thesis specifically addresses. The chapter also attempts to provide brief background information on the possible methods to approximate/obtain the optimal change of measure. We also discuss in this chapter the different measures of accuracy generally used in rare events simulations.

- **Chapter 3: Static Network Reliability Estimation with Node Failures**
  In Chapter 3, we propose an adapted algorithm for approximate Zero-Variance IS methodology with the focus on node failures. Our proposed methodology is an adaptation of the approximate zero-variance IS methodology given in [29], where the focus was on links as failure components. We also propose an adaptation of the Ford-Fulkerson's maxflow-mincut algorithm for considering flow through nodes and

use it in our main algorithm to estimate s-t node unreliability using IS. The application is shown on benchmark networks and also a real DCS system of Alstom.

- **Chapter 4: Cross-Entropy Optimization of Transitions in Markovian SPNs**
  In this Chapter 4, we propose a multi-level CE optimization scheme for Markovian SPNs. The chapter illustrates the mathematical model of the CTMCs. We model the systems conveniently using SPNs from which we can also extract the underlying CTMC. The chapter gives a comprehensive description of the pre-simulation CE algorithm that we propose, to obtain optimized (in terms of CE distance) IS rates for transitions of interest in Markovian SPNs, while also including logistics aspects. Application and results of the proposed algorithm on various examples to estimate the cumulative steady-state unavailability is also shown.

- **Chapter 5: Conclusions**
  In this Chapter 5, we discuss the conclusions drawn from the various methodologies proposed in this thesis. The chapter also discusses the directions and possibilities for future research that we consider useful, for example, use of the CE pre-simulation algorithm in the context of Non-Markovian SPNs to estimate steady-state measures.

Chapter 2

A General Introduction to Rare Events Simulation

This chapter aims to give the background information about the problems encountered in the simulation of rare events by standard methods (i.e., standard Monte Carlo simulation). It briefly introduces the background information on the two main rare event simulation methods: the *Splitting* and the IS. The concepts behind the IS method are elaborated further and is the main conceptual idea focused upon in the current work. The chapter also provides the background information for the optimal change of measure in IS. It further discusses the various possibilities to find or approximate the optimal change of measure that subsequently can provide highly accurate estimates of reliability metrics and help in meeting the RAM objectives. Finally, we also discuss in the final section the measures of accuracy that are considered to quantify the effectiveness and efficiency of the IS methodologies in rare events simulation context.

2.1 Standard Monte Carlo Simulations & Limitations

The impracticality of analytic and numerical analysis to compute reliability metrics for large-scale highly reliable systems justify the use of simulation techniques [37]. Computer simulations provide a practically feasible alternative to study the behavior of real-life systems that are too difficult to examine analytically [37]. Simulations have found use in a wide variety of disciplines: engineering, operation research and management science, statistics, mathematics, physics, economics, biology, medicine, engineering, chemistry, and the social sciences [37]. Standard simulations are based on Monte Carlo (MC) simulation technique and are stochastic, that is, they include some randomness in the underlying model [37] and use a statistical approximation to provide point estimates [14].

To further illustrate standard MC simulations in a very generic setting as given in [17], let us consider that X is a real random variable (r.v.) having $f(x)$ as its probability density function (pdf). One wants to estimate the probability γ of some event A happening (i.e., $\gamma = \mathbb{E}[\psi(X)]$) [17], where $\psi(X)$ is an identity function such that $\psi(x) = 1$ if $x \in A$ or

$\psi(x) = 0$ if $x \notin A$. Then, the probability of A occurring is given by:

$$\gamma = \int_{-\infty}^{\infty} \psi(x)\, f(x)\, dx = \mathbb{E}[\psi(X)], \tag{2.1}$$

where $\mathbb{E}$ is the expectation under the density $f(x)$. Solving the above integral by standard MC simulation to estimate γ would require drawing n independent samples of X, i.e., $(X_1, ..., X_n)$ from the density $f(x)$. We assume that X_i is an independently and identically distributed (iid) random variable with mean μ and variance σ^2 [35, 37]. The estimator of the true value γ is then given by:

$$\hat{\gamma}_{MC} = \frac{1}{n}\sum_{i=1}^{n} \psi(X_i). \tag{2.2}$$

Here, $\hat{\gamma}_{MC}$ is an unbiased estimate of γ (i.e., $\mathbb{E}[\hat{\gamma}_{MC}] = \gamma$). As per the law of large numbers, $\hat{\gamma}_{MC}$ converges to γ as $n \to \infty$. However, to know the accuracy of the point estimate $\hat{\gamma}_{MC}$ (how close it is to the actual unknown value γ), one needs to provide not only the point estimate $\hat{\gamma}_{MC}$ but also a confidence interval (CI) with a given degree of confidence as well [37]. The CI is built around the estimator $\hat{\gamma}_{MC}$ and requires the variance of the estimator, as given by:

$$Var(\hat{\gamma}_{MC}) = {\sigma_n}^2 = \frac{\sigma^2}{n} = \frac{\gamma(1-\gamma)}{n},$$

where $\sigma^2 = \mathbb{E}[(\psi(X))^2] - (\mathbb{E}[\hat{\gamma}_{MC}])^2$. In practice, neither $\mathbb{E}[\hat{\gamma}_{MC}] = \gamma$ is known and nor the variance $Var(\hat{\gamma}_{MC})$ beforehand. Generally, we estimate σ^2 by the sample variance, given by [37]:

$$S^2_{MC} = \frac{1}{n-1}\sum_{i=1}^{n} \psi^2(X_i) \; - \; \frac{n}{n-1}\,(\hat{\gamma}_{MC})^2 .$$

According to the central limit theorem (CLT), a CI at level $(1-\alpha)$ for γ is approximately:

$$\left[\hat{\gamma}_{MC} \mp z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right] = \left[\hat{\gamma}_{MC} \mp z_{\alpha/2}\sqrt{\gamma(1-\gamma)/n)}\right],$$

where $z_{\alpha/2} = \Phi^{-1}(1-\alpha/2)$, with Φ being the standard normal cumulative distribution function $\mathcal{N}(0,1)$ (i.e., mean 0 and variance 1) [17, 22, 37]. The relative half-width of the CI is now given by:

$$z_{\alpha/2}\frac{\sqrt{S^2_{MC}/n}}{\mathbb{E}[\hat{\gamma}_{MC}]} = z_{\alpha/2}\left(\sqrt{\frac{1-\gamma}{\gamma n}}\right).$$

To explain the limitation of standard MC technique, in the above discussion, let us suppose we want the relative half-width for a 95% CI for γ to be less than 0.1, i.e., $1.96\left(\sqrt{\frac{1-\gamma}{\gamma n}}\right) \leq 0.1$. In such a case, we see that $n \approx 100 \times 1.96^2 \times (1-\gamma)/\gamma$ [17]. Thus, n is inversely proportional to γ and consequently, the smaller the γ is, the larger the n must be. Also, in rare event simulations, where the probabilities for the respective EOI are very low, *absolute error* ceases to be of interest [14]. In such situations, relative error (RE) for the point estimate $\hat{\gamma}_{MC}$ is used

[14]. We discuss these measures of accuracy for rare event simulations further in Section 2.5. The RE is the ratio of the standard deviation of the estimate and its expected value, given by:

$$RE(\hat{\gamma}_{MC}) = \sqrt{\frac{1-\gamma}{n\gamma}} \approx \frac{1}{\sqrt{n\gamma}}.$$

As evident from the above equation, for a fixed RE, n is inversely proportional to γ and the required $n \to \infty$ as $\gamma \to 0$ [14, 17]. In other words, for a fixed sample size n, RE is unbounded and would increase to ∞ as $\gamma \to 0$ [17]. Either way, the computational cost (in terms of simulation time or the number of samples n) is bound to increase in standard MC simulations for efficient estimation of rare event probabilities (when γ is too small). In highly reliable systems, this is the usual scenario where system failure is the EOI with a very low probability, and standard MC simulations become quickly inefficient.

2.2 Rare Events Simulation Techniques

To overcome the previously mentioned problems in the context of rare event simulations using the standard MC method, the two main methods of *Splitting* and Importance Sampling (IS) have been focused upon and developed in the literature. The primary goal of these methods is to accelerate the occurrence of the rare EOI and to obtain significant variance reduction that would not be practically possible by standard MC simulations.

The *splitting* method employs a sequential sampling strategy to decompose a "difficult" estimation problem into a sequence of "easy" problems [37]. Splitting is useful for various purposes, including, rare-event problems, Monte Carlo counting, randomized optimization, etc. [37]. This method is applied (in a generic sense) by creating an artificial drift towards the rare EOI in a twofold approach: terminate with some probability the trajectories that seem to go away from it and split (clone) those that are going towards the target EOI [20]. More specifically, the state space of a system is divided into intermediate subsets, also called as levels. Starting from a given level, the paths (also called as *trajectories* or *chains* or *particles*) that do not reach the next level will not reach the target set (i.e., the rare EOI); while those that reach the next level are again split (or cloned) into multiple copies and evolve thereafter [21]. This approach would create an artificial drift towards the target set by favoring the paths that are evolving in the direction of the target set [21]. We can obtain the unbiased estimator of the target metrics by multiplying the original estimator by an appropriate factor (which is 1 in some cases) [20]. This is also known as *multilevel splitting* [20, 21].

The two main aspects of the splitting method on which it's efficiency is dependent upon are the choice of the number of levels and the amount of splits per level [21]. The levels in the splitting method are defined via an *importance function* that aims to represent how close a state is from a rare set of state (i.e., the target state of EOI) [20, 21]. In the splitting method, finding this *importance function* is the main difficulty. Also, the amount of splitting when reaching a new level is also crucial. Too much splitting at a given level can result in an explosion of chains, while too little would make very few trajectories to go in the right direction [21]. There are several ways of doing the splitting, like, *fixed splitting*, the *fixed effort* method, the *fixed success* implementation or the *fixed probability of success* implementation [20, 21, 38]. Without going further into the details of the splitting method, in the current work, the focus is on the use of IS methods. Readers interested in the splitting method can find more detailed description in [20, 21, 37, 38, 39, 40, 41, 42].

The second method of interest is the IS method that owes its origins to the works of von Neumann, Ulam, Fermi, Kahn, Metropolis, and their colleagues [23, 24, 43, 44, 45] who paved the path for the IS method by employing random sampling to perform computations in nuclear physics during the 1940's. Since long, IS has been considered as a useful technique to increase the efficiency of MC simulation algorithms for numerical evaluation of integrals [17, 25]. IS is a variance reduction technique (like the splitting method also) that is useful in conjunction with MC simulations. In simulations, certain values of the input random variables have a higher impact on the target metrics we try to estimate as compared to others [46]. Emphasizing upon these "important" values (i.e., sampling more frequently) can result in a significant variance reduction [46]. Hence the name *Importance Sampling*. In IS, the dynamics of the system, in terms of the underlying probability distributions, are changed to increase the occurrence of the rare EOI and these new probability measures are called as the *change of measure* [14]. Since the simulation is done under a new probability law, it would result in a biased estimation if directly applied. The correction factor, called as the *likelihood ratios* corrects the bias of the simulation outputs. The likelihood ratio is the Radon-Nikodym derivative of the true underlying distribution with respect to the biased simulation distribution [22, 25]. However, the most crucial factor in IS is the proper selection of a change of measure. The choice of an appropriate change of measure is not straightforward and usually depends on the system we simulate, and an unsuitable change of measure can even increase the variance [47]. The following sections elaborate the IS method in more detail.

2.3 Importance Sampling: Basics

The IS method forms the basis of the current work and thus, discussed in more detail here. In Section 2.3.1, we introduce the basic notations for IS in the context of a continuous case, to maintain continuance with the general setting previously presented in Section 2.1. Similarly for discrete or static (time playing no role) cases, simply the probabilities could be used. In Section 2.3.2, we explain the definition of an optimal change of measure. We also discuss the different possibilities for application of IS techniques to obtain highly accurate estimates of target metrics in Section 2.4. Section 2.5 elaborates on the measures of accuracy that we generally consider in rare event simulations, and also use in the current work.

2.3.1 Notations

As previously explained, IS involves changing the probability laws to increase the occurrence of the rare EOI. In the general settings introduced in Section 2.1, let us consider the integral in Equation 2.1. Multiplying and dividing the integral by another pdf $g(x)$ we get:

$$\gamma = \int \psi(x)\, f(x)\, dx = \int \psi(x) \frac{f(x)}{g(x)} g(x) dx = \mathbb{E}_g \left[\psi(x) \frac{f(X)}{g(X)} \right] = \mathbb{E}_g \left[\psi(X) L(X) \right], \quad (2.3)$$

where $L(x) = f(x)/g(x)$ is the *likelihood ratio* on the set $\{x : \psi(x)f(x) > 0\}$ (g strictly positive), and by $L(x) = 0$ otherwise [14]. The sampling is done from the density $g(x)$ (i.e., performing the change of measure) and the expectation $\mathbb{E}_g$ is thus taken under $g(x)$. For IS, the Equation 2.3 is valid for any density $g(x)$ provided the following condition is met:

$$\textbf{Condition for IS:} \quad g(x) > 0\, \forall \psi(x) = 1 \ \text{ whenever } \psi(x)f(x) > 0, \quad (2.4)$$

meaning a non-zero possible sample under $f(x)$ must also be a non-zero possible sample under $g(x)$ [17]. The density $g(x)$ is sometimes called as the *importance density or the IS density or the IS change of measure* [17].

With the given formulation, sampling $X_1, ...X_n$ from $g(x)$, the IS estimator is given by:

$$\begin{aligned} \hat{\gamma}_{IS} &= \frac{\psi(X_1)L(X_1) + \cdots + \psi(X_n)L(X_n)}{n} \\ &= \frac{1}{n} \sum_{i=1}^{n} \psi(X_i)L(X_i). \end{aligned} \quad (2.5)$$

The variance of the IS estimator is $Var(\hat{\gamma}_{IS}) = \tilde{\sigma}_n^2 = \tilde{\sigma}^2/n$ and ,

$$\tilde{\sigma}^2 = \mathbb{E}_g \left[(\psi(X)L(X))^2 \right] - (\mathbb{E}\left[\psi(X)\right])^2, \quad (2.6)$$

which again is unknown and estimated by the sample variance:

$$S_{IS}^2 = \frac{1}{n-1} \sum_{i=1}^{n} \psi^2(X_i)L^2(X_i) - \frac{n}{n-1} \left(\hat{\gamma}_{IS}\right)^2. \quad (2.7)$$

By the CLT again, the CI at confidence level $(1-\alpha)$ for the IS estimator has the same form as before:

$$\left[\hat{\gamma}_{IS} \mp z_{\alpha/2}\tilde{\sigma}/\sqrt{n}\right].$$

Here, again it is assumed that the IS estimator has a normal distribution, which often can be a good approximation, but not always [22]. The goal of the IS method is to accurately estimate the target metric by its estimator $\hat{\gamma}_{IS}$, especially in the context of rare events (i.e., $\gamma \to 0$). Accurate estimation itself means the reduction of variance of the estimator. The next section elaborates the idea for the choice of the IS density $g(x)$ that can help in obtaining the desired accuracy (i.e., variance reduction).

2.3.2 Optimal Change of Measure

The importance density $g(x)$ in Equation 2.3 can be any density as long as it suffices the Condition 2.4. However, what would be the optimal importance density to choose for IS? An optimal importance density (i.e., the optimal change of measure) is the one that minimizes the variance of $\hat{\gamma}_{IS}$, and since $Var(\hat{\gamma}_{IS}) \geq 0$, the minimum variance possible is 0 [17]. If we consider a density $g(x) \equiv g^*(x)\ \forall\ \psi(x) > 0$ then:

$$g^*(x) = \frac{f(x)\psi(x)}{\gamma}, \tag{2.8}$$

where $L(x) = \gamma/\psi(x)$ whenever $\psi(x)f(x) > 0$ [14, 17]. Since $\psi(x)L(x)$ would be a constant in this case, the variance of a constant is zero, as given below:

$$Var(\hat{\gamma}_{IS}) = \frac{1}{n}Var_{g^*}(L(X)\psi(X)) = \frac{1}{n}Var_{g^*}[\gamma] = 0. \tag{2.9}$$

The optimal change of measure is thus the conditional density (the condition being the rare event occurs) and leads to a zero-variance estimator [14, 17]. In such a case, the simulation becomes a kind of *pseudo simulation* leading to the exact value in just one sample, i.e., the unbiased estimator with zero-variance [14]. The density $g^*(x)$ is sometimes referred to as the zero-variance IS density too. However, the problem with attempting to sample from the optimal importance density in Equation 2.8 is that it explicitly depends on the value of γ, the original problem we are trying to solve. If γ is already known, then there is no point in running simulations at all. Also, even if γ were known, it might be impractical to sample efficiently from $g^*(x)$ [17]. The observation in Equations 2.8 & 2.9, however, also gives two important possibilities: first, there is a possibility to find good IS densities that can reduce the variance; second, there are several possibilities to develop IS schemes leading to the best possible estimator by exploring in greater depth the optimal change of measure [14].

Since $\mathbb{E}_g[\psi(X)L(X)] = \gamma$ (see Equation 2.3) for any density $g(x)$ satisfying the Condition 2.4, reducing the variance of the estimator corresponds to selecting a density $g(\cdot)$ that reduces the second moment of $\psi(X)L(X)$ (the first term) of Equation 2.6. The second term $(\mathbb{E}[\psi(X)])^2$ is a constant. From this we can also say:

$$\begin{aligned}\mathbb{E}_g\left[\psi^2(X)L^2(X)\right] &= \int \psi(x)\left(\frac{f(x)}{g(x)}\right)^2 g(x)dx \\ &= \int \psi(x)\frac{f(x)}{g(x)}f(x)dx = \mathbb{E}\left[\psi(X)L(X)\right].\end{aligned} \tag{2.10}$$

Thus, in order to reduce the variance, the likelihood ratio $L(x)$ should be small on the target set A. It is evident that the event A is rare under density $f(x)$, i.e., $f(x)$ is small on set A. In order to make the $L(x)$ small on A, the selection of g should be such that $g(x)$ is large on A, thus making the event A more likely to occur [17]. Also, in the context of rare events, simulations are performed until the relative accuracy of the estimator (given by the ratio of the CI half-width and the quantity γ to be estimated) is below a certain threshold [22]. It requires that $\tilde{\sigma}^2/n$ is approximately proportional to γ^2 such that the number of samples needed is proportional to the variance of the estimator [22].

State dependency of the change of measure

The application of the IS change of measure can be considered in two ways, especially, in the context of simulations of Markov chains [22]:

1. State-Independent: We consider a change of measure as state-independent if it simply does not depend on the current state of the Markov chain.
2. State-Dependent: In this strategy, a new IS change of measure is used at each step of the Markov chain, and it takes into account the current state of the Markov chain.

In specific applications (e.g., queueing theory), the state-independent change of measure has been used [48]. However, it has been shown that the state-independent change of measure does not always work (i.e., not even asymptotically efficient) and the use of state-dependent change of measure can produce asymptotic efficiency [49, 50, 51]. The state-dependent change of measure, even though being more efficient, is also significantly more complex than a state-independent one for large state space models [49, 52].

2.4 Algorithmic Strategies for Importance Sampling

The general conclusion from the aforementioned discussions is the accurate estimation of the target metrics γ in the context of rare events, i.e., when $\gamma \to 0$. There are several possible strategies to meet this objective. We can define a good IS strategy as the one that leads to variance reduction of the final estimator, best case being a zero-variance estimator. There are several possible good IS strategies, and we can broadly classify them in two [22]: first, restricting a priori the change of measure to a parametric class and then trying to optimize the parameters; second, to directly approximate the optimal change of measure (i.e., the zero-variance one) via an approximation. In both these cases also, the choices can be based on simple heuristics, or via a known asymptotic approximation, or by adaptive methods that can learn the vector of parameters that meet a specific objective [22] (e.g., cross-entropy distance minimization or direct variance minimization). Some of these approaches are discussed in the next sections and also used in the current work.

2.4.1 Optimization within a Parametric Class

In case of large state spaces (the usual case for real systems), the best strategies include a priori restriction to a parametric class (either explicitly or implicitly) for the IS change of measures and to estimate the parameter vectors that can minimize the variance of the final estimator [22]. For optimization within a parametric class, we can consider a family of measures for IS $\{\tilde{\mathbb{P}}_{\tilde{\theta}}\ ,\ \tilde{\theta} \in \Theta\}$, where Θ is the parameter space. These family of measures may represent a family of densities $g_{\tilde{\theta}}(\cdot)$ or probability vectors $\tilde{p}_{\tilde{\theta}}$ for the discrete case, or probability measures associated with the transition matrix of a Markov chain, etc. [22]. For example, in a continuous case where g belongs to the exponential family, $\tilde{\theta}$ represents the vector of rates. It is noteworthy that the zero-variance change of measure is not always within the parametric family [18, 22]. Here, we wish to choose the IS change of measure g (as previously explained in Equation 2.3) such that the associated estimator is optimal in a well-defined sense [32] and also leads to variance reduction.

Let us consider a parametric family $\tilde{\mathbb{P}}_{\tilde{\theta}} = \{f(x;\tilde{\theta})\}$ indexed by a parameter vector $\tilde{\theta}$ that also contains the original density f [32]. Now, we can write $f(x) = f(x;\theta)$ for some parameter vector θ for the original density. The goal in optimization within a parametric class is to find the set of parameters $\tilde{\theta}$ over the set Θ that either minimizes the variance of the IS estimator (e.g., in case of Variance Minimization algorithms [31]) or some other measure of distance to the zero-variance measure (e.g., Cross-Entropy distance [31]). Clever selection of the parametric class is a key ingredient that inherently should include good IS strategies within that class [22]. The question arising now is how to find the good set of parameters $\tilde{\theta}$. Possibilities for this include selecting $\tilde{\theta}$ based on asymptotically valid approximations (e.g., Large-Deviations Theory) or learning them adaptively (e.g., Variance Minimization or Cross-Entropy algorithms) [22].

Large-Deviations Theory (LDT) has been considered for non-adaptive parameter selection [22] in the literature. The problem with the IS method as stated before is that an optimal change of measure is unknown. In [22], the popular idea of fixing $\tilde{\theta}$ based on an asymptotic analysis is explained for specific examples for binomially distributed random variables. Generally speaking, we can consider LDT as an extension of traditional limit theorems of probability theory [53]. The *weak* law of large numbers in its basic form states that certain probabilities converge to zero, while LDT focuses on the rate of convergence [53]. However, in [22] it has also been shown that the LDT cannot provide a bounded relative variance. The LDT method is beyond the scope of the current study and readers interested in LDT can find more detailed description and examples of selecting a good change of measure based on LDT in [17, 54, 55, 56, 57, 58, 59, 60].

Adaptive learning of $\tilde{\theta}$ for IS can be done in several ways, for example, using Variance Minimization (VM) algorithms or using Cross-Entropy (CE) algorithms to find an optimal $\tilde{\theta}^*$ that meets a certain optimality criteria [31]. For example, in the VM algorithms, the optimality criteria is to reduce variance directly; while, in case of the CE algorithms it is minimizing the CE distance. Obviously, the main goal of using IS is to obtain variance reduction (irrespective of using VM or CE algorithms) in the estimation of the target metrics. For further explanation, let us consider we restrict ourselves a priori to a parametric class with the IS density given by $f(x;\tilde{\theta})$. The objective is to find a $\tilde{\theta}$ within the same parametric class as the original one and reduce the variance. There are several strategies to obtain the mentioned $\tilde{\theta}$, for example, stochastic approximation or by using sample average approximation [22]. In the sample average approximation, one could write the variance of the second moment as a mathematical expression depending on $\tilde{\theta}$ [22]. The expectation can be thus replaced by a sample average function of $\tilde{\theta}$ and this sample function can be optimized with respect to $\tilde{\theta}$

via simulation [22]. In this approach, the sampling in the simulation is done under the IS measure $\{\tilde{\mathbb{P}}_{\tilde{\theta}}\}$ that may differ from the original measure $\mathbb{P}$ and not necessarily belong to the selected family. The solution of such an optimization approach should provide a $\tilde{\theta}^*$ that is optimized under certain optimality criteria. The next sections explain the two approaches that can be used to obtain an optimal $\tilde{\theta}^*$ based on VM (lets say $\tilde{\theta}^*_{vm}$) or CE optimization (lets say $\tilde{\theta}^*_{ce}$) techniques and their respective optimality criteria.

Variance-minimization

In the VM algorithms, the optimality criteria is to find a $\tilde{\theta}$ within the same parametric family $\tilde{\mathbb{P}}_{\tilde{\theta}}$, that reduces the variance, or actually the second moment as given in the Equation 2.10. Similar to Equation 2.10, the second moment is:

$$\begin{aligned}\mathbb{E}_{\tilde{\theta}}\left[\psi^2(X)L^2(X;\theta;\tilde{\theta})\right] &= \int \psi(x)\left(\frac{f(x;\theta)}{f(x;\tilde{\theta})}\right)^2 f(x;\tilde{\theta})dx \\ &= \int \psi(x)\frac{f(x;\theta)}{f(x;\tilde{\theta})}f(x;\theta)dx = \mathbb{E}\left[\psi(X)L(X;\theta;\tilde{\theta})\right].\end{aligned} \tag{2.11}$$

From above, the optimization problem aiming to reduce the second moment of the IS estimator is defined as:

$$\min_{\tilde{\theta}\in\Theta} \upsilon_{vm}(\tilde{\theta}) = \min_{\tilde{\theta}\in\Theta} \mathbb{E}_{\tilde{\theta}}\left[\psi^2(X)L^2(X;\theta;\tilde{\theta})\right], \tag{2.12}$$

where υ_{vm} is implicitly defined above. Let us consider the optimizer in this case to be $\tilde{\theta}^*_{vm}$. The optimal VM parameter vector is then:

$$\tilde{\theta}^*_{vm} = \arg\min_{\tilde{\theta}\in\Theta} \mathbb{E}_{\tilde{\theta}}\left[\psi^2(X)L^2(X;\theta;\tilde{\theta})\right]. \tag{2.13}$$

The optimization problem in Equation 2.12 is difficult to solve as the density with respect to which the expectation is computed depends on the decision variable $\tilde{\theta}$ [31]. To overcome this, let us consider a sampling density $f(x;\check{\theta})$ in the same parametric family [31]. Multiplying and dividing the integrand for the expectation in Equation 2.12, by $f(x;\check{\theta})$, the new expectation can be written as:

$$\min_{\tilde{\theta}\in\Theta} \upsilon_{vm}(\tilde{\theta}) = \min_{\tilde{\theta}\in\Theta} \mathbb{E}_{\check{\theta}}\left[\psi^2(X)\, L(X;\theta;\check{\theta})\, L(X;\theta;\tilde{\theta})\right]. \tag{2.14}$$

In the above Equation 2.14, the expectation is now taken under $f(x;\check{\theta})$ and $\check{\theta}$ is an arbitrary reference parameter [31]. The optimization problem of Equation 2.14 can be solved via sample average approximation obtained from simulations and thus the optimizer $\tilde{\theta}^*_{vm}$ in Equation 2.13 can be computed.

Cross-entropy optimization

The foundation of the CE method was an extension to the variance minimization problem in the context of rare event simulations [61], as introduced in [62, 63]. In conjunction with IS, the CE method is to be used as a pre-simulation methodology to optimize the parameters which we can later use as the IS change of measure for the final simulations [31]. A very generic description of the CE method is that it provides an easy and adaptive learning algorithm involving following two phases in stochastic simulations [31]:

- Generating a sample of random data (vectors, trajectories, etc.) as per a specific random mechanism [31].
- Updating the specifics (i.e., parameters) of the random mechanism for the next iteration based on the data and produce better samples in the next iteration [31].

In the current context of the work, let us consider the parameterized IS density $f(x;\tilde{\theta})$ as explained previously. The CE method aims to find the importance density that is closest in CE distance (*Kullback-Leibler divergence*) to the zero-variance importance density $g^*(x)$ as given in Equation 2.8 within the same parametric family f. Let us consider this IS density closest in CE distance to $g^*(x)$ is $f(\cdot;\tilde{\theta}^*_{ce})$, i.e., the optimal one. The CE distance between two distributions ($g^*(x)$ and $f(x;\tilde{\theta})$) is given by [18, 22, 31]:

$$\mathscr{D}(g^*(x), f(x;\tilde{\theta})) = \mathbb{E}_{g^*}\left[log \frac{g^*(x)}{f(x;\tilde{\theta})}\right], \tag{2.15}$$

where $\mathbb{E}_{g^*}$ is the expectation under $g^*(x)$ and $f(x;\tilde{\theta})$ is the density with the parameter vector $\tilde{\theta}$ [18, 22, 31]. Replacing $g^*(x)$ by its true value, it is equivalent to [22]:

$$\mathscr{D}(g^*(x), f(x;\tilde{\theta})) = \mathbb{E}\left[\frac{\psi(X)}{\mathbb{E}[\psi(X)]} log \left(\frac{\psi(X)}{\mathbb{E}[\psi(X)]} f(x;\theta)\right)\right] - \frac{1}{\mathbb{E}[\psi(X)]}\mathbb{E}\left[\psi(X)\; log\; f(x;\tilde{\theta})\right], \tag{2.16}$$

where except for the last term of expectation, all other terms are constants and denoted by their expectations. The goal of the CE method is to minimize the above equation which gives the CE distance between the two densities to obtain the $f(x;\tilde{\theta}^*_{ce})$ i.e., with the optimizer $\tilde{\theta}^*_{ce}$. As the last expectation only depends on $\tilde{\theta}$, we can instead maximize $\mathbb{E}\left[\psi(X)\; log\; f(x;\tilde{\theta})\right]$ to minimize the CE distance given above [22]. Now the problem is transformed from a minimization of CE distance to maximization of:

$$\max_{\tilde{\theta}\in\Theta} \upsilon(\tilde{\theta}) = \max_{\tilde{\theta}\in\Theta} \mathbb{E}[\psi(X)\; log\; f(x;\tilde{\theta})], \tag{2.17}$$

where υ is implicitly defined above [31].

Using IS with importance density for sampling, let us say $f(x;\check{\theta})$ with arbitrary reference parameter $\check{\theta}$ (as explained previously for the VM case), the above equation can be re-written as:

$$\max_{\tilde{\theta}\in\Theta} \upsilon(\tilde{\theta}) = \max_{\tilde{\theta}\in\Theta} \mathbb{E}_{\check{\theta}}\left[\psi(X)\; L(X;\theta;\check{\theta})\; log\; f(x;\tilde{\theta})\right]. \tag{2.18}$$

The expectation is taken under the importance density defined by $f(x;\check{\theta})$ and the likelihood ratio is again the ratio of the original density ($f(x;\theta)$) and the IS density used for sampling ($f(x;\check{\theta})$). Now, the optimal solution of Equation 2.18 can be written as [31]:

$$\tilde{\theta}^*_{ce} = \arg\max_{\tilde{\theta}\in\Theta} \mathbb{E}_{\check{\theta}}\left[\psi(X)\, L(X;\theta;\check{\theta})\, log\, f(X;\tilde{\theta})\right]. \tag{2.19}$$

The optimization problem is defined now by the right-hand side of Equation 2.18 to obtain the $\tilde{\theta}^*_{ce}$, and can be achieved by a sample average approximation. We can replace the expectation in Equation 2.18 by a sample average over simulations performed under $\check{\theta}$. However, the selection of $\check{\theta}$ is very crucial as we know that in case of rare events, it is difficult to know a priori the good value of $\check{\theta}$ to perform the simulations. The term "good" here means the distribution with parameter vector $\check{\theta}$ under which the optimizer of the sample average approximation does not have too much variance and is sufficiently reliable [22].

It is also noteworthy that restricting a priori to a specific parametric family would result in a lower dimensional (i.e., more restrictive) optimizer obtained via VM or CE strategies as explained above.

Comparison of VM and CE optimization techniques

The main goal of using IS in rare event simulations is to efficiently estimate the target metrics (e.g., γ here) with variance reduction. We discussed in previous sections how adaptive techniques of VM or CE could be useful for optimizing within the same parametric class. The two different approaches of VM and CE have the same goal to reduce the variance, however, a different optimality criteria. As explained previously, the VM method aims to reduce the second moment of the IS estimator and consequently the variance directly, and we obtain the optimizer $\tilde{\theta}^*_{vm}$. On the other hand, the CE method aims to reduce the CE distance between the zero-variance density $g^*(x)$ and the IS density $f(x;\tilde{\theta})$, providing the optimizer $\tilde{\theta}^*_{ce}$. Both the respective optimizing parameter vectors are well defined as per the respective approach of VM or CE.

However, the drawback in attempting to reduce the variance directly through the adaptive VM programs is that it has proved to be quite time-consuming and computationally burdensome in practice [31, 32]. Instead of directly trying to minimize the variance of the estimator, the CE method provides a simpler and faster adaptive procedure for estimating the optimal density. Also, another advantage of the CE method is that the solution of the optimization problem from which an optimal density can be obtained, often has a closed-form solution [32]. In the examples considered in [32], it is shown that the optimal VM and CE densities are asymptotically (in terms of rarity of EOI) identical or very close. The $f(\cdot;\tilde{\theta}^*_{ce})$ density which is closest to $g^*(x)$ in terms of the CE distance is also the one for which the asymptotic variance of the estimator is minimum [32]. In general, the optimization problems of VM and CE are difficult to be solved analytically, except in few specific cases [32]. To overcome this issue, a multi-level procedure (both for VM and CE) is used [31] and explained in later chapters. Thus, we consider that it could be easier to use a CE optimization technique in comparison to VM techniques to meet the objective of variance reduction in IS and find the good change of measure for that.

2.4.2 Learning Techniques and Heuristic Approximations

The previous Section 2.4.1 discussed optimization techniques based on optimization within a parametric class. However, a priori restriction to a parametric class also means a more restrictive optimization, as the zero-variance density may or may not be within the parametric class.

To overcome the limitations of the a priori restriction to a parametric class, in the literature, several heuristic approximation and learning techniques have been discussed [29, 55, 64, 65, 66, 67]. Learning techniques are based on adaptive learning of the target metrics (where we can consider it as a function of the states of a Markov chain) and plugging in a zero-variance approximation. One approach for this is called as *adaptive Monte Carlo* [66, 67]. Another approach is based on a stochastic approximation of the target metrics in the context of discrete-time finite-state Markov chains (DTMC) as proposed in [65], and is called as the *adaptive stochastic approximation*. Experimental results shown by [65] prove the effectiveness of these methods when the state-space is small. However, for practical cases where state-space is large, learning techniques become difficult to be applied and impractical [22].

Heuristic approximation techniques like zero-variance IS based on mincuts including the most likely paths to failure has been discussed in [29] for static network reliability estimation. In [64], heuristic methods to approximate the zero-variance IS for highly reliable Markovian systems are proposed and their effectiveness illustrated. In the next Chapter 3, we propose a zero-variance IS approximation based on mincuts with maximal probability for the case of static network reliability estimation, specifically, when nodes are the failing components.

2.5 Measures of Accuracy

The previous sections introduced how IS could be useful for estimation of target metrics in rare events context and the different techniques that help in obtaining an accurate estimator of the target metrics. We also discussed the second moment (and consequently the variance) of the IS estimator, the RE, and the CI. However, in the literature, several measures of accuracy are used to define and quantify the robustness and reliability of the estimators obtained via various IS techniques. In this section, we elaborate further on the topic of these measures of accuracy for IS estimators, especially in the rare event context (i.e., $\gamma \to 0$). These measures are used in the current work to quantify the benefits and efficiency of the IS techniques in the next chapters.

Theoretical analysis of rare event simulation usually involves the use of a rarity parameter "ε" [17, 35]. In such cases, the model is parameterized by a small and real ε ($\varepsilon > 0$) such that $\gamma = \gamma(\varepsilon) \to 0$ as $\varepsilon \to 0$. This parameterization means that the EOI occurs (in the original model) with a probability that converges to zero as the rarity parameter $\varepsilon \to 0$. We can formally say that $\lim_{\varepsilon \to 0^+} \gamma(\varepsilon) = 0$ [68]. In applications, $\gamma(\varepsilon)$ can be considered as a performance measure in the form of a mathematical expectation, and some model parameters can be defined as functions of ε [68]. For different models, different parameterization may specify different asymptotic regime [17, 55, 69]. For example, in reliability models, the failure rates of components can be parameterized by ε [17, 70, 71, 72] or in case of static network reliability estimation, the probabilities of failure of components can be parameterized by ε [29, 68].

In the context of rare event simulations, as the rarity of EOI increases (i.e., $\varepsilon \to 0$), the quality of the estimator with respect to accuracy and coverage needs to be controlled [35]. In [35], the two notions of *robustness* and *reliability* of an estimator are discussed. *Robustness* is concerned with the error itself (i.e., how far the estimator is from the true value), while *reliability* of the estimator considers the quality of error estimation (i.e., the CI coverage) as $\varepsilon \to 0$ [35]. The next sections elaborate the concepts of asymptotic robustness properties and efficiency measures that we use in the current work.

2.5.1 Asymptotic Robustness Properties

Asymptotic behavior of the IS estimator is usually studied on the basis of how the RE changes when the rarity parameter $\varepsilon \to 0$. Recall that in rare events context, *absolute error* is uninteresting and RE (the ratio of the standard deviation and the expected value of the target metrics) is considered [14]. The variance of the IS estimator is the same as before (see Equation 2.6) and the standard deviation is $\tilde{\sigma}_n = \sqrt{\tilde{\sigma}_n^2}$. The relative error RE is defined by the half-width CI and is given as:

$$\mathrm{RE}(\hat{\gamma}_{IS}(\varepsilon)) = z_\delta \frac{\sqrt{Var(\hat{\gamma}_{IS}(\varepsilon))}}{\gamma(\varepsilon)} = z_\delta \frac{\tilde{\sigma}_n}{\gamma(\varepsilon)}, \tag{2.20}$$

where we consider γ is parameterized by ε and z_δ is the $1-\delta/2$ quantile of the standard normal distribution ($z_\delta = \Phi^{-1}(1-\delta/2)$), Φ being the standard normal cumulative distribution) [35]. Here, the IS estimator is considered, but these measures can be computed for the estimator under original probability measure too. Now we consider three different asymptotic properties that define the robustness of the estimator obtained via IS.

- **Bounded Relative Error (BRE):** The typically desirable property in the asymptotic regime is the BRE [35]. BRE property is obtained if the RE of the estimator remains bounded as $\varepsilon \to 0$ [17]. Formally, it means that

$$\mathrm{RE}(\hat{\gamma}_{IS}(\varepsilon)) \leq C \quad \text{as} \quad \varepsilon \to 0, \tag{2.21}$$

  where C is some constant. For interpretation, estimation of a target metric $\gamma(\varepsilon)$ with a given relative accuracy can be achieved in a bounded number of replications even if $\varepsilon \to 0$ [35].

- **Logarithmic Efficiency (LE):** LE (also called as asymptotic optimality) property for an unbiased estimator, lets say $\hat{\gamma}_{IS}(\varepsilon)$ here, of γ is considered true with respect to the rarity parameter ε, if the following condition is met [17, 73]:

$$\lim_{\varepsilon \to 0} \frac{ln\, \mathbb{E}_g \left[\hat{\gamma}_{IS}^2(\varepsilon)\right]}{ln\, \gamma(\varepsilon)} = 2. \tag{2.22}$$

Generally, LE is a weaker property in comparison to BRE. The above quantity (under limit) is always positive and ≤ 2 because $Var(\hat{\gamma}_{IS}(\varepsilon)) \geq 0$ and so $\mathbb{E}_g \left[\hat{\gamma}_{IS}^2(\varepsilon)\right] \geq \gamma(\varepsilon)$ and then $ln\, \mathbb{E}_g \left[\hat{\gamma}_{IS}^2(\varepsilon)\right] \geq 2\, ln\, \gamma(\varepsilon)$ [35].

- **Vanishing Relative Error (VRE):** It is the strongest property in comparison to the BRE or LE and is formally defined for the RE as [68, 69]:

  $$\text{RE}(\hat{\gamma}_{IS}(\varepsilon)) \to 0 \quad \text{as} \quad \varepsilon \to 0, \tag{2.23}$$

  or equivalently if

  $$\limsup_{\varepsilon \to 0} \frac{\sigma(\varepsilon)}{\gamma(\varepsilon)} = 0. \tag{2.24}$$

  It means that the VRE property holds if the relative error also goes to zero as $\varepsilon \to 0$. Asymptotically, this would result in a zero-variance estimator.

Among the above asymptotic robustness measures, the most desirable property is the VRE followed by BRE and LE (note, BRE implies asymptotic optimality too [53]). There are several other robustness measures existing based on higher degree moments [68], on the Normal approximation [35] or the variance of the empirical variance, where BRE for the empirical variance was studied in [74]. In this thesis, we focus on the BRE and the VRE property. In the next section, we explain the efficiency measures that can quantify the advantage of using IS methods.

2.5.2 Efficiency Measures in Rare Event Simulations

Standard MC simulations have found use in a variety of fields due to their simpler algorithms. On the other hand, the use of IS (along with using methods for finding an optimal change of measure) results in more complex algorithms. These complex algorithms can generally lead to increased computation time (i.e., higher computational cost). However, using IS as explained previously can (under certain conditions) provide significant variance reduction and thus higher accuracy. This is a trade-off between accuracy and computation time, and therefore requires quantification of the gain obtained via IS methods when used in simulations. For this purpose, the product of the variance and expected computation time per replication has been defined [35], namely, the *work-normalized variance*.

Let us consider in the current context, using standard MC method, the variance of the estimator ($\hat{\gamma}_{MC}$) is ${\sigma_n}^2$ obtained from n replications in computation time t_n. We can define the work-normalized variance (let's say var_{wn}) for this by $var_{wn} = {\sigma_n}^2 t_n$ [35]. The IS estimator ($\hat{\gamma}_{IS}$) having variance $\tilde{\sigma}_n^2$ is let's say obtained in computation time $\tilde{t}_n$ (from n replications again) and has work-normalized variance ($v\tilde{a}r_{wn}$) equal to $\tilde{\sigma}_n^2 \tilde{t}_n$ [35]. This allows us to compare the two estimators obtained from two different techniques [35] and the better estimator is the one having lower work-normalized variance. In this context, using IS should mean that $\tilde{var}_{wn} < var_{wn}$ for the same sample size n. Also, using this, the gain by using IS technique can be quantified as:

$$\text{Gain} = \frac{var_{wn}}{\tilde{var}_{wn}} = \frac{\sigma_n{}^2 t_n}{\tilde{\sigma}_n^2 \tilde{t}_n}. \tag{2.25}$$

The "Gain" as defined above should be greater than one and $\tilde{\sigma}_n^2(\varepsilon) < \sigma_n{}^2(\varepsilon)$, for any IS method to be considered better in comparison to a standard MC simulation. It is obvious from this that the validation of a BRE property (or VRE) in an asymptotic regime would result in significant increase in gain as $\varepsilon \to 0$.

Thus, we discussed here the measures of accuracy as well as efficiency in the current context. However, when doing simulations, many times issues may occur that one is not aware of, or at worse, these issues are hidden. In the next section, we discuss some of these issues.

2.5.3 Issues in Empirical Results

As previously stated, the robustness measures in the asymptotic regime can be considered to capture the error accurately, however, reliability or the quality of the error estimation is also important. In the previous discussions on asymptotic robustness measures, we assumed that the coverage of the CI obtained based on the CLT is always valid [35]. It is noteworthy that the point estimates or variance are estimated via simulations and their exact values are unknown. For example, if the rare EOI does not occur ever in a simulation, one might obtain a CI of $(0,0)$ empirically. In [35], examples are presented where the RE seems bounded in empirical results but actually it is not. The validity of the CI coverage can be ascertained using the Normal approximation or the coverage function can be used to represent the actual coverage [35]. Diagnostic ideas based on expected value of likelihood ratio, observed relative error values or based on coverage function are also presented in [35].

In the current work, we consider the robustness measures of BRE (or VRE), the efficiency measures of work-normalized variance (and the gain), and the coverage of the computed CI for the true value (in smaller analytically solvable examples). However, one needs to be always careful with the empirical results obtained, in case the issues mentioned above occur.

Chapter 3

Static Network Reliability Estimation of Passenger Rail Systems

This chapter aims to focus on the estimation of reliability (or contrarily the unreliability in context of rare events) for static networks where time plays no role. The chapter illustrates and proposes an adapted version of the approximate zero-variance IS (Importance Sampling) method for estimation of static network reliability, and we also show the application of the proposed method on a real system, while also focusing on the measures of accuracy as explained in the previous chapter.

3.1 Motivation and Objectives

In the previous chapters, we discussed how RAM analysis could be done efficiently using simulations for highly reliable systems. The RAM metrics also help in determining the LCC (Life Cycle Costs) of the offered solutions and thus, help rail system suppliers, such as ALSTOM, to comply with contractual obligations and ensure system availability requirements [1]. In the current context, one of the main functions necessary for nominal operation is the communication of different signals between centrally localized computers and trackside/onboard equipment. In Alstom's urban metro solution the subsystem whose objective is to perform this communication function is called the *data communication system* (DCS). The DCS uses a dual-ring topology to communicate equipment located in different stations (or the track) with centrally located computers. It is configured so that all end communication equipment has a preferred ring through which it sends its messages (but is able to use the other ring if needed) and all messages are simultaneously transmitted on each ring separately. The availability of the DCS is the probability that all messages between all end-communication devices are successfully transmitted. We describe the model in more detail in Section 3.4.3.

The choice of the reliability metric here thus becomes straightforward as the reliability between specific components of the network of a subsystem like the DCS. Reliability of a system is defined as the probability of performing as required for the time interval (let us say

t_1,t_2), under given conditions [2]. In other words, as a metric, reliability is the probability that no failure occurs over a specified period of time [75].

Methods like Reliability Block Diagrams (RBD) and Fault Tree Analysis (FTA) are often used for evaluation of reliability or availability of passenger rail systems. For example, [76] discuss the use of RBD method for calculation of reliability or availability for non-repairable and repairable systems, respectively. A comparison between RBD and FTA methods is discussed by [77], where it is shown that the RBD method is more appropriate than the FTA method for availability assessment. Another approach to predict the availability of systems also involves the creation of a Markov model which characterizes the different failure paths of the network. Typically up to third-order failure paths are included. The selection of which paths to model is made by reliability modeling experts. Resulting models are hard to validate both by other experts and the end-user. Furthermore, it is not clear whether there exist relevant failure paths that have not been modeled. Modeling the communication network as a graph with communication equipment as nodes and communication paths as links overcomes both shortcomings: first, the model can be easily validated by the design expert and the client; and second, by defining successful communication as the existence of a path between the communicating devices, no modeling of failure paths is needed because path finding algorithms can be used to establish connectivity.

The static network reliability problem deals with the estimation of the probability that a given set of nodes in a graph model are connected when each individual component (link or node) is in an UP/ DOWN (working/ failed) state according to their respective probabilities. The case where links are the failing elements is essential in many applications and has been extensively studied [27]. However, there are a wide range of applications where nodes are the failing components, such as the DCS here, or models of network survivability [28]. This requires an adaptation of the existing methods to the case of node failures. Formally, a node failure means that the node becomes nonfunctional and its associated links useless. In the *2-terminal* or *source-to-terminal* reliability problem, two nodes of the graph are fixed and the reliability of the network is defined as the probability of having a path between those two nodes. In such analysis, a node failure causes a higher number of s-t paths to become nonfunctional as compared to a link failure (depending on the node's degree). Thus, the reliability of a network would be affected more severely in the case of node failures.

Computing the *unreliability* of highly reliable systems (e.g., the DCS) requires efficient simulation techniques. For large graphs, an exact computation of the unreliability u becomes an NP-hard problem that is impractical to be solved analytically [29]. As previously stated, standard MC methods can estimate u in its crude form (CMC) sampling n stochastically independent realizations of the graph and computing the proportion of these n realizations for which the s-t are not connected [29]. For rare events, when $u << 1$ here, the standard MC method (the crude MC) will suffer from inefficiency requiring unnecessarily large n and consequently increasing the computational cost, as previously detailed in Section 2.1. The accuracy and efficiency of the simulation process is captured by the RE and the gain (based on work-normalized variance) when $u \to 0$, see Section 2.5.

In this case of estimating the static network unreliability u for highly reliable networks, the use of IS methods is justifiable. However, again for using IS, if the IS probabilities which lead to frequent failure are not properly selected, the likelihood ratio may have a huge variance resulting in a bad estimation, even if the failure event is not rare anymore [22]. Recall, the optimal change of measure is unknown (see Section 2.3.2) and to find/approximate such a change of measure is our goal here.

The current chapter aims to propose and adapt the dynamic importance sampling method based on MC simulations as described by [29], considering node failures, and to prove its application on an existing example of a communication network (i.e., the DCS). We propose an approximation of the zero-variance IS method based on minimal cuts having relatively high failure probability in the subgraph that remains after removing the nonfunctional nodes and their associated links (irrespective of being functional or not, if one of the associated node is failed), while enforcing the states of the nodes which are functional, at each step of a Markov chain [29]. These cuts approximate the u conditional on the current state, at each step. The networks are analysed as a graph model and the Ford-Fulkerson *maxflow-mincut* algorithm [78] is adapted for considering flow through nodes. The estimators obtained via simulations here adhere to the measures of accuracy, as explained previously in Section 2.5. We observe the BRE property in general as node reliability increases, and the VRE property under additional conditions (as proved by [29] for link failure case). The usefulness of the proposed scheme is proved using a quantified measure of work normalized variance (var_{wn}).

The next sections are as follows: in Section 3.2, we explain the mathematical model for considering node failures and the inefficiency of crude Monte Carlo (CMC) methods with respect to rare event analysis for static network unreliability. In Section 3.3, continuing with the basic idea of IS, an approximate Zero-Variance IS method is illustrated (from a theoretical perspective). In Section 3.4, we propose the adaptation of the Ford-Fulkerson maxflow-mincut algorithm and the approximate zero-variance IS method based on mincuts with maximal probability while considering node failures. The analysis of the method on various networks, including a case study on an existing network of DCS and its results showing BRE or VRE properties are illustrated in Sections 3.4.2 & 3.4.3. Conclusions of the current study are drawn at the end.

3.2 Mathematical Model for Static Networks

Let us consider an undirected connected graph $\mathscr{G} = (\mathscr{N}, \mathscr{L})$, where $\mathscr{N} = \{1, ..., m\}$ is the set of nodes, and $\mathscr{L}$ is the set of links. The model is static, that is, time is not considered. Links are assumed to always work, but the nodes are subject to (independent) failures. Node $i \in \mathscr{N}$ fails with a probability q_i, where $0 < q_i < 1$. A configuration of the graph [79] is given by the random vector $X = (X_1, ..., X_m)$, where for all $i \in \mathscr{N}$, $X_i = 1$ or 0 representing the working or failed state of a node i, respectively. Retaining only the functional nodes $\mathscr{N}'$, a random partial graph $\mathscr{G}' = (\mathscr{N}', \mathscr{L})$ of $\mathscr{G}$ is obtained.

To estimate the probability u that two nodes named s and t (for source and terminal respectively) are not connected in the random graph $\mathscr{G}$, a structure function can be defined as $\psi(X)$ equal to 1 if s and t are not connected in $\mathscr{G}'$ (or equivalently the configuration X), else as 0 [29]. The expectation $u = \mathbb{E}[\psi(X)]$ or the s-t unreliability is given by [29]:

$$u = \mathbb{E}[\psi(X)] \;=\; \sum_{x\in\{0,1\}^m} \psi(x)\mathbb{P}[X = x] \;=\; \sum_{x\in\{0,1\}^m} \psi(x)\prod_{i=1}^{m}(q_i(1-x_i)+(1-q_i)x_i),$$

where $x = (x_1,...,x_m)$, $\mathbb{P}$ is the original probability law of the network and $\mathbb{E}$ is the expectation under $\mathbb{P}$.

The state space having 2^m possible configurations will require an exponentially increasing time to calculate the u from the above formula. The exact evaluation is an NP-hard problem in general [80], so approximation techniques like MC simulations are required in such cases. The performance of the proposed methodology is studied by parameterizing q_i (under the condition $q_i \longrightarrow 0$) as a polynomial function of a rarity parameter $\varepsilon \ll 1$. As explained in [29], for each $i \in \mathcal{N}$, there are independent constants $a_i > 0$ and $b_i \geq 0$ such that $q_i = a_i\varepsilon^{b_i}$. The overall u is a finite sum of products of such possibilities. It is then a polynomial in ε and

$$u = u(\varepsilon) = \Theta(\varepsilon^c), \tag{3.1}$$

for a constant $c \geq 0$ and Θ is a mathematical notation. The different mathematical notations for the asymptotic analysis here are:

- For Θ: $f(\varepsilon) = \Theta(g(\varepsilon))$, if $f(\varepsilon) = \underline{O}(\varepsilon^d)$ and $f(\varepsilon) = O(\varepsilon^d)$,
- For $\underline{O}$: $f(\varepsilon) = \underline{O}(g(\varepsilon))$ if $|f(\varepsilon)| \geq c_2 g(\varepsilon)$ for some constant $c_2 > 0$ for all ε sufficiently small,
- For O: $f(\varepsilon) = O(g(\varepsilon))$ if $|f(\varepsilon)| \leq c_1 g(\varepsilon)$ for some constant $c_1 > 0$ for all ε sufficiently small.

For estimation of u using the CMC method, independent samples of X are generated to form an unbiased estimator (similar to the one for the continuous case in Equation 2.2) for which the s and t is disconnected [29]. In this case it is given by :

$$\overline{U}_{MC}^{(n)} = \frac{1}{n}\sum_{j=1}^{n}\psi(X^{(j)}). \tag{3.2}$$

The accuracy of the estimator $\overline{U}_{MC}^{(n)}$ is measured by its empirical variance (lower value means better accuracy):

$$(S_{MC}^{(n)})^2 = \frac{\overline{U}_{MC}^{(n)}(1-\overline{U}_{MC}^{(n)})n}{(n-1)}, \tag{3.3}$$

and the CI on the estimation of u is given by:

$$\left[\overline{U}_{MC}^{(n)} - c_\alpha S_{MC}^{(n)}/\sqrt{n},\, \overline{U}_{MC}^{(n)} + c_\alpha S_{MC}^{(n)}/\sqrt{n}\right]. \tag{3.4}$$

The relative half-width of the CI:

$$c_\alpha \frac{((S_{MC}^{(n)})^2/n)^{1/2}}{\mathbb{E}[\psi(X)]} = c_\alpha \left(\frac{1-u}{un}\right)^{1/2},$$

for a confidence level α increases to ∞ when $u \longrightarrow 0$ (i.e., rare-event) for a fixed n [14, 29], as previously explained in Section 2.1. Thus, it is required to have a more efficient technique than the CMC method for rare-event analysis of static networks as considered here.

3.3 Zero-Variance Importance Sampling Approximation

The generic idea of IS methods as presented in Section 2.3 was to change the probability laws driving the model to increase the occurrence of the rare event (here s-t nodes being disconnected) and recover the bias by multiplying the estimator with the likelihood ratio.

Similarly here, the original probabilities $\mathbb{P}$ of the 2^m possible configurations of X are replaced by a new probability $\widetilde{\mathbb{P}}$ which gives

$$u = \mathbb{E}[\psi(X)] = \sum_{x\in\{0,1\}^m} \psi(x)\mathbb{P}[X=x] = \sum_{x\in\{0,1\}^m} \psi(x)L(x)\widetilde{\mathbb{P}}[X=x],$$

where the likelihood ratio $L(x) = \mathbb{P}[X=x]/\widetilde{\mathbb{P}}[X=x]$. The condition $\widetilde{\mathbb{P}}[X=x] > 0$, when $\mathbb{P}[X=x] > 0$ must be met when $\psi(x) > 0$.

The unreliability is now $u = \widetilde{\mathbb{E}}[\psi(X)\ L(X)]$ and the unbiased estimator obtained from IS takes the form of:

$$\overline{U}_{IS}^{(n)} = \frac{1}{n}\sum_{j=1}^{n} \psi(X^{(j)})L(X^{(j)}),$$

where $X^{(1)},\ldots,X^{(n)}$ are s-independent copies of X distributed according to $\widetilde{\mathbb{P}}$. The CI over u under the new probability law $\widetilde{\mathbb{P}}$ can be obtained from Equation 3.4, by replacing the sample mean $\overline{U}_{MC}^{(n)}$ with $\overline{U}_{IS}^{(n)}$, and the variance $(S_{MC}^{(n)})^2$ with the sample variance $(S_{IS}^{(n)})^2$ of $\psi(X^{(j)})\ L(X^{(j)})$.

As explained in [29], if $\widetilde{\mathbb{P}}$ is the optimal probability (i.e., the optimal change of measure as explained for the continuous case in Equation 2.8) for which the variance is reduced to zero (i.e., ideal zero-variance estimator), all the probabilities are inflated by a factor proportional to $\psi(x)$ and:

$$\widetilde{\mathbb{P}}[X=x] = \frac{\psi(x)\mathbb{P}[X=x]}{u},$$

for all the configurations of $x \in \{0,1\}^m$. This above equation means that, for the realizations for which the system does not fail (i.e., $\mathbb{P}[X=x]$), the sampling $\widetilde{\mathbb{P}}[X=x] = 0$, while for the other realizations where system fails (i.e., $u > 0$), the original $\mathbb{P}$ is to be divided by u to obtain the optimal $\widetilde{\mathbb{P}}$. However, this method is impractical because it requires the knowledge of u, the value that is to be computed actually.

Under the zero-variance IS method, as described in [29], but considering the sampling of nodes instead of links, node states are sampled sequentially given the state of previously sampled nodes. Formally, if $q_i = \mathbb{P}[X_i = 0] = 1 - \mathbb{P}[X_i = 1]$ under $\mathbb{P}$, then q_i is changed at each step depending on the previously generated values $X_1, ..., X_{i-1}$. The unreliability of the graph $\mathscr{G}'$, conditional on the already sampled nodes 1 to $i-1$ is given by:

$$u_i(x_1, ..., x_{i-1}) = \mathbb{E}[\psi(X) \mid X_1 = x_1, ..., X_{i-1} = x_{i-1}],$$

which also means

$$u_i(x_1, ..., x_{i-1}) = q_i \; u_{i+1}(x_1, ..., x_{i-1}, 0) + (1 - q_i) \; u_{i+1}(x_1, ..., x_{i-1}, 1),$$

and the overall unconditional unreliability of the graph can be written as $u = u_1(\emptyset)$[29].

If for $i = (1, ..., m)$, q_i is replaced by

$$\widetilde{q}_i \overset{def}{=} \widetilde{\mathbb{P}}[X_i = 0 | X_1 = x_1, ..., X_{i-1} = x_{i-1}] \; = \; q_i \frac{u_{i+1}(x_1, ..., x_{i-1}, 0)}{u_i(x_1, ..., x_{i-1})}, \tag{3.5}$$

as shown in [29], following the same arguments for the proof, this sequential IS gives a zero-variance estimator. However, it (again) requires the exact knowledge of all the functions of u_i and specifically $u_1(\emptyset) = u$, which is not practical.

Following [29], it is proposed to replace $u_i(.)$ in Equation 3.5 by an approximation $\hat{u}_i(.)$, which gives

$$\widetilde{q}_i = \widetilde{\mathbb{P}}[X_i = 0] = \frac{q_i \hat{u}_{i+1}(x_1, ..., x_{i-1}, 0)}{q_i \hat{u}_{i+1}(x_1, ..., x_{i-1}, 0) + (1 - q_i) \hat{u}_{i+1}(x_1, ..., x_{i-1}, 1)}. \tag{3.6}$$

If $\hat{u}_{i+1}(.)$ is not too far from $u_{i+1}(.)$ for each i, then the variance would be reduced by a large factor. It is important to note that the network unreliability u will not change according to the order in which the nodes (or vertices) are numbered in the graph but the change of measure would depend on the ordering in the proposed algorithm. In the analysis, it is found that certain enumeration of nodes does vary the estimated unreliability $\hat{u}$ by a very small factor and so does the value of RE. However, we do not have yet a robust heuristic to choose the ordering of the nodes which could evaluate the optimum unreliability estimate $\hat{u}$ possible, or a correlation between the estimation or RE with the ordering.

Using this IS scheme, we can prove in an asymptotic regime where $\varepsilon \longrightarrow 0$, while the graph topology is fixed, that some condition on the approximation $\hat{u}_i(.)$ guarantees that BRE or even VRE are satisfied. Recall that BRE means the standard deviation of the estimator divided by the mean value σ/u is kept bounded as $\varepsilon \to 0$; in other words the sample size to get a predefined RE is independent of the rarity parameter ε. VRE means that σ/u tends to zero with $\varepsilon \longrightarrow 0$: asymptotically the estimator is perfect. The conditions are the same as in [29], with nodes considered instead of links but again the arguments are exactly the same. The impact of considering nodes is more in the computation of the chosen approximation that will be described later.

- Let us suppose that for each i and $(x_1,...,x_i) \in \{0,1\}^i, 1 \le i \le m$, there is a constant $a_{i+1}(x_1,...,x_i)$ independent of ε such that

$$\hat{u}_{i+1}(x_1,...,x_i) \quad = \quad a_{i+1}(x_1,...,x_i)u_{i+1}(x_1,...,x_i) + o(u_{i+1}(x_1,...,x_i)). \tag{3.7}$$

If this condition is satisfied, then BRE holds.

- Let us define

$$S_1 = \{x \in \{0,1\}^m : \psi(x) = 1 \quad \text{and} \quad \widetilde{\mathbb{P}}[X = x] = \Theta(1)\},$$

and

$$S_0 = \{x \in \{0,1\}^m : \psi(x) = 1 \quad \text{and} \quad \widetilde{\mathbb{P}}[X = x] = o(1)\}.$$

The union $S_0 \cup S_1$ is the set of possible configurations where the system fails. The configurations in S_1 are not rare under IS, while the ones in S_0 are still rare. The required additional condition for VRE involves $x \in S_1$ only. Assuming the assumptions defined for BRE hold, VRE property holds if at least one of the following three conditions is satisfied $\forall x = (x_1,.....,x_m) \in S_1$ and for each i:

$$\frac{\hat{u}_{i+1}(x_1,...x_{i-1},1)}{u_{i+1}(x_1,...,x_{i-1},1)} = \frac{\hat{u}(x_1,...,x_{i-1},0)}{u_{i+1}(x_1,...,x_{i-1},0)} + o(1), \tag{3.8}$$

or

$$\begin{gathered} x_i = 0,\ a_{i+1}(x_1,...,x_i) = 1, \quad \text{and} \\ (1-q_i)\,\hat{u}_{i+1}(x_1,...,x_{i-1},1) = o(q_i\,\hat{u}_{i+1}(x_1,...,x_{i-1},0)), \end{gathered} \tag{3.9}$$

or

$$\begin{gathered} x_i = 1,\ a_{i+1}(x_1,...,x_i) = 1, \quad \text{and} \\ q_i\,\hat{u}_{i+1}(x_1,...,x_{i-1},0) = o((1-q_i)\,\hat{u}_{i+1}(x_1,...,x_{i-1},1)). \end{gathered} \tag{3.10}$$

3.4 Static Network Reliability Simulations: Application and Results

In this section we illustrate the various topologies considered in our analysis along with the DCS structure. Also, we show the results obtained from the application of the approximate zero-variance IS method based on mincuts. For the purpose of our study, we modified the maxflow-mincut algorithm proposed by Ford-Fulkerson [78] for considering flow through nodes. The algorithm used thereafter is explained in the next Section 3.4.1, where the computation based on node failures is presented.

3.4.1 Approximation based on Mincuts computed from Ford-Fulkerson Maxflow-Mincut Algorithm

The proposed approximation of $\hat{u}_i$ is to consider the probability of a *mincut with a maximal probability* where nodes (and associated links) sampled as failed are removed from the graph, and nodes sampled operational are compacted. Recall that a cut of a graph is defined as the partition of nodes of the graph into two disjoint subsets of $\mathscr{G}$ while a mincut $\mathbb{C}$ is a cut whose capacity is minimum over all the cuts of $\mathscr{G}$. A mincut with a maximal probability is a mincut whose probability that all nodes are failed, is computed as the product of the failure probabilities of those nodes. With such an approximation, the condition for BRE are always satisfied (similarly to [29]), and VRE can be satisfied in some cases.

The question is now, how to compute such an approximation in the case of nodes? As explained in Section 3.2 by Equation 3.1, we parameterize the system unreliability in an asymptotic regime with respect to a rarity parameter $\varepsilon \longrightarrow 0$ such that $q_i \rightarrow 0 \; \forall i$. Define $\varepsilon = \max_i q_i$ such that $\forall i$,

$$q_i = \varepsilon^{c_i}, \tag{3.11}$$

with $c_i = \frac{\log q_i}{\log \varepsilon} \geq 1$. Calling c_i the *capacity* of node i, and c the capacity of the graph obtained from maxflow-mincut algorithm proposed by Ford-Fulkerson based on links, we get $c = \sum_{i \in \mathscr{C}} c_i$ for $\mathscr{C}$, a mincut with a maximal probability, and the corresponding probability is $q_{\mathscr{C}} = \varepsilon^c = \prod_{i \in \mathscr{C}} \varepsilon^{c_i} = \prod_{i \in \mathscr{C}} q_i$. The trick is to use the log to switch from the sum of capacities to the product of probabilities.

The Ford-Fulkerson algorithm [78] adapted for nodes differs from the case of links such that: if a node i fails (i.e., $x_i = 0$), all its associated links are useless and removed, and if a node i is considered working (i.e., $x_i = 1$) then it is removed from the graph model and its associated nodes are mutually linked to each other. This makes the algorithm for the node failure case more complex compared to link failure where a failed link can be just removed and the connecting nodes of a perfectly working link are merged. The Algorithm 1 proposed illustrates the adapted Ford-Fulkerson algorithm.

Algorithm 1 Adapted Ford-Fulkerson Maxflow-Mincut Algorithm

```
use nodelist (list) to store node names of any random graph $\mathscr{G}'$ or $\mathscr{G}$
for $i = 1$ to size.nodelist() do
    assign capacity $c_i$ to node $i$ using Equation 3.11
end for
initialize cap; initialize apath (array) and flow (array) for all nodes $i$
use Breadth First Search (BFS) to find a path apath between $s$ and $t$ passing apath and flow argument
Ensure: BFS finds path only where flow can be assigned and return a value $P > 0$ if found a path
while $P > 0$ do
    find node with minimum $c_i$ (mincap) in apath
Ensure: remaining capacity in all the nodes in apath is greater than mincap
    assign flow equal to the mincap to all nodes in apath
    $cap += mincap$
    initialize apath and use BFS to find another path where flow can be assigned and return $P > 0$
end while
return cap
```

The mincut with maximal probability problem is $\Theta(u)$ as explained in [29] and as a consequence the BRE property is satisfied in the case of node failures too. With more stricter conditions (one of the Equations 3.8 or 3.9 or 3.10) satisfied, the VRE property is also observed as illustrated in [29] for link failures. The overall procedure is illustrated by Algorithm 2.

Algorithm 2 Approximate Zero-Variance IS using Ford-Fulkerson adapted algorithm

```
 1: $L \longleftarrow 1$; Starting with original graph $\mathscr{G}$
 2: for $i = 1$ to $m$ until $s$ are directly connected or completely disconnected i.e., BFS find direct path or no path
 3:     Compress node $i$ corresponding to $\mathscr{G}(x_1, ..., x_{i-1}, 1)$ of Step 2 and mutually connect its neighbors in $\mathscr{G}'$
 4:     find $cap$ (mincut of maximal probability in the set $\mathscr{C}_i$) using Algorithm 1
 5:     $\hat{u}_{i+1}(x_1, ..., x_{i-1}, 1) \longleftarrow \mathbb{P}[E(cap)]$;
 6:     Erase node $i$ corresponding to $\mathscr{G}(x_1, ..., x_{i-1}, 1)$ of Step 2 and remove it everywhere in $\mathscr{G}'$
 7:     find $cap$ (mincut of maximal probability in the set $\mathscr{C}_i$) using Algorithm 1 again
 8:     $\hat{u}_{i+1}(x_1, ..., x_{i-1}, 0) \longleftarrow \mathbb{P}[E(cap)]$;
 9:     compute $\widetilde{q}_i$ via Equation 3.6
10:     generate $U_i$ a random variate over $(0, 1)$;
11:     if $U_i < \widetilde{q}_i$ then
12:         $x_i \longleftarrow 0$;   $L_i \longleftarrow q_i / \widetilde{q}_i$;
13:     else
14:         $x_i \longleftarrow 1$;   $L_i \longleftarrow (1 - q_i)/(1 - \widetilde{q}_i)$;
15:     end if
16:     $L \longleftarrow L \times L_i$;
17: end for
18: return $Y = \psi(x_1, ..., x_m) \times L$
```

3.4.2 Numerical Results

We considered four topologies for illustrating the mincut-maxprob approximation. In all the studied topologies, the nodes are sampled by order of their numbers from s to t. The nodes are homogeneous with unreliability $q_i = \varepsilon$ for $i = 1, 2, 3, \ldots, m$, where $\varepsilon \in \mathbb{R}$. For three examples, we also show the application of the methodology for a heterogeneous case where the unreliability of all the nodes in the graph model is not the same and are selected according to random heuristics. The metric of interest is the probability that s and t are disconnected in the following networks.

Example 1: First we consider a graph with 21 nodes and 36 links as shown in Figure 3.1. The two cases considered for this example are as follows:

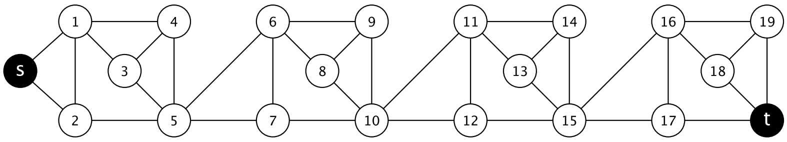

Figure 3.1 Graph with 21 nodes.

- *Homogeneous case:* In this case, all the nodes in the graph model of Figure 3.1 have the same unreliability. It means, $q_i = \varepsilon\ \forall i$ ($i = 1,2,3,...,m$ and $\varepsilon \in \mathbb{R}$). Computing recursively, its exact solution of the unreliability is:

$$u = 1-(1-\varepsilon)^7[(1-\varepsilon)^3 - 3(1-\varepsilon)^2 + 3 - 2\varepsilon]^4,$$

where u is the exact unreliability of the graph.

Results from simulations (with $n = 10^6$) using the IS scheme and the CMC method are presented in Tables 3.1 & 3.2, respectively.

Table 3.1 Static Network: Simulation results from IS scheme for 21 Nodes graph (Homogeneous).

ε	Exact Soln	Estimate	95 % CI	STD	**R.E.**	Time (s)
10^{-3}	3.0010×10^{-3}	3.0005×10^{-3}	$(2.9986 \times 10^{-3}, 3.0025 \times 10^{-3})$	1.01×10^{-3}	**0.34**	147.21
10^{-5}	3.0000×10^{-5}	2.9995×10^{-5}	$(2.9975 \times 10^{-5}, 3.0015 \times 10^{-5})$	1.00×10^{-5}	**0.33**	147.54
10^{-7}	3.0000×10^{-7}	2.9995×10^{-7}	$(2.9975 \times 10^{-7}, 3.0014 \times 10^{-7})$	1.00×10^{-7}	**0.33**	145.41
10^{-9}	3.0000×10^{-9}	2.9995×10^{-9}	$(2.9975 \times 10^{-9}, 3.0014 \times 10^{-9})$	1.00×10^{-9}	**0.33**	144.02
10^{-11}	3.0000×10^{-11}	2.9995×10^{-11}	$(2.9975 \times 10^{-11}, 3.0014 \times 10^{-11})$	1.00×10^{-11}	**0.33**	147.22

Table 3.2 Static Network: Simulation results from CMC for 21 Nodes graph (Homogeneous).

ε	Estimate	95 % CI	STD	R.E.	Time (s)
10^{-3}	2.9720×10^{-3}	$(2.8653 \times 10^{-3}, 3.0787 \times 10^{-3})$	5.44×10^{-2}	18.32	4.87
10^{-5}	2.9000×10^{-5}	$(1.8445 \times 10^{-5}, 3.9555 \times 10^{-5})$	5.39×10^{-3}	185.69	5.00
10^{-7}	0.0	(0.0,0.0)	0.0	–	4.96

- *Heterogeneous case:* In this case, we consider the unreliability of nodes in Figure 3.1 as: $q_i = \varepsilon$ (for $i = 1,6,11,16$), $q_i = \varepsilon^{1.15}$ (for $i = 2,7,12,17$), $q_i = \varepsilon^{1.35}$ (for $i = 3,8,13,18$), $q_i = \varepsilon^{1.5}$ (for $i = 4,9,14,19$) , $q_i = \varepsilon^{0.85}$ (for $i = 5,10,15$) and $\varepsilon \in \mathbb{R}$. Again computing recursively the exact unreliability u of the graph, we have:

$$u = 1 - \left[\left(1-\varepsilon(\varepsilon^{1.15}+\varepsilon^3-\varepsilon^4)\right)^4(1-\varepsilon^{0.85})^3\right].$$

Results from simulations using the IS scheme and the CMC method are shown in Tables 3.3 & 3.4 for $n = 10^6$ iterations in all cases.

Table 3.3 Static Network: Simulation results from the IS scheme for 21 Nodes graph (Heterogeneous).

ε	Exact Soln	Estimate	95 % CI	STD	**R.E.**	Time (s)
10^{-3}	8.4327×10^{-3}	8.4313×10^{-3}	$(8.4258 \times 10^{-3}, 8.4368 \times 10^{-3})$	2.81×10^{-3}	**0.33**	156.44
10^{-5}	1.6869×10^{-4}	1.6866×10^{-4}	$(1.6855 \times 10^{-4}, 1.6877 \times 10^{-4})$	5.62×10^{-5}	**0.33**	154.27
10^{-7}	3.3661×10^{-6}	3.3655×10^{-6}	$(3.3633 \times 10^{-6}, 3.3677 \times 10^{-6})$	1.12×10^{-6}	**0.33**	160.66
10^{-9}	6.7162×10^{-8}	6.7150×10^{-8}	$(6.7106 \times 10^{-8}, 6.7194 \times 10^{-8})$	2.24×10^{-8}	**0.33**	156.34
10^{-11}	1.3401×10^{-9}	1.3398×10^{-9}	$(1.3389 \times 10^{-9}, 1.3407 \times 10^{-9})$	4.47×10^{-10}	**0.33**	158.55

The IS scheme adapted for the case of node failures suggest that the BRE property holds while we also obtain a tight confidence interval over the estimation of u as $\varepsilon \longrightarrow 0$, as shown in Tables 3.1 & 3.3, for both the homogeneous and heterogeneous cases, respectively. The exact analytical solution (for both the cases) is also bounded in the 95% CI we obtain from the simulations, thus proving the accuracy of the adapted IS scheme.

Table 3.4 Static Network: Simulation results from CMC for 21 Nodes graph (Heterogeneous).

ε	Estimate	95 % CI	STD	R.E.	Time (s)
10^{-3}	8.4170×10^{-3}	$(8.2379 \times 10^{-3}, 8.5961 \times 10^{-3})$	9.14×10^{-2}	10.85	5.02
10^{-5}	1.6000×10^{-4}	$(1.3521 \times 10^{-4}, 1.8479 \times 10^{-4})$	1.26×10^{-2}	79.05	5.35
10^{-7}	1.0000×10^{-6}	$(-9.6000 \times 10^{-7}, 2.9600 \times 10^{-6})$	1.00×10^{-3}	1000.00	5.46
10^{-9}	0.0	$(0.0, 0.0)$	0.0	–	5.16

For the homogeneous case, comparing it with CMC method (see Table 3.2), for the same topology the proposed IS scheme average simulation time was 1.4628×10^{-4} seconds per iteration while for the CMC method it was 4.9433×10^{-6} seconds per iteration. However, the CMC method didn't record any failure event for $\varepsilon < 10^{-6}$. Also, the work normalized variance ($\tilde{var}_{wn}$) (variance multiplied by the expected computing time per iteration) in the IS scheme is much lower and reduces much rapidly as $\varepsilon \longrightarrow 0$ for any value of ε as compared to the one obtained from CMC simulations. For example, for a rare event $\varepsilon = 10^{-7}$, $(S_{MC}^{(n)})^2 = \hat{u}(1-\hat{u})$ and CMC method's var_{wn} is 1.4827×10^{-12}. But with zero-variance approximation based on mincuts, $\tilde{var}_{wn}$ for same case is 1.4628×10^{-18} which is much lower. The gain in such a case (as given in Equation 2.25), is the ratio given by $var_{wn}/\tilde{var}_{wn}$ and in this case would be approximately 1.01 million times. This means for this particular example when $\varepsilon = 10^{-7}$, the IS scheme is 1.01 million times more efficient than the CMC method.

For the heterogeneous case, the results are similar in terms of BRE property and accuracy as $\varepsilon \rightarrow 0$. The average time for the IS scheme was 1.5725×10^{-4} seconds per iteration while for the CMC method (see Table 3.4) it was 5.2470×10^{-6} seconds per iteration. For a very rare event with $\varepsilon = 10^{-9}$, with $(S_{MC}^{(n)})^2 = \hat{u}(1-\hat{u})$, the CMC method has a $var_{wn} = 3.5234 \times 10^{-13}$. On the other hand, for proposed IS scheme the work-normalized variance is $\tilde{var}_{wn} = 7.8813 \times 10^{-20}$. The gain with the IS method, again using Equation 2.25, in such a case would be approximately 4.47 million times. Also, similar to the homogeneous case, here also the $\tilde{var}_{wn}$ from the IS scheme decreases faster as compared to the CMC method's as $\varepsilon \rightarrow 0$.

It is to be noted that in Table 3.4, for $\varepsilon = 10^{-7}$, even with $n = 10^6$ iterations, we are able to get a point estimate because the unreliability $u = 3.3661 \times 10^{-6}$ for this case (see Table 3.3, $\varepsilon = 10^{-7}$). In such a case, there is always a possibility of a single or more failure events occurring in $n = 10^6$ samples. Also, the lower bound of the 95% CI in it is shown as negative, as the CI is built around the point estimate (see Equation 3.4) and the values can go in the negative region when the point estimate itself is inaccurate.

Example 2: We now take a Dodecahedron topology having 20 nodes and 30 links, as shown in Figure 3.2, which is often used as a benchmark for network reliability estimation techniques [27].

For this example also, both the homogeneous and heterogeneous cases are considered. In the homogeneous case, all nodes are assigned same probability of failure, i.e., $q_i = \varepsilon \, \forall i$. In the heterogeneous case for this example, we considered $q_i = \varepsilon$ for all the even numbered nodes and $q_i = \varepsilon^{0.65}$ for all the odd numbered nodes in the Figure 3.2, respectively. Results are

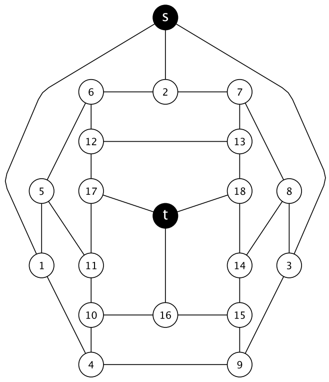

Figure 3.2 Dodecahedron.

presented in Tables 3.5 & 3.6 for both the homogeneous and heterogeneous cases respectively, with $n = 10^6$ in all cases.

Table 3.5 Static Network: Simulation results from the IS scheme for a Dodecahedron (Homogeneous).

ε	Estimate	95 % CI	STD	**R.E.**	Time (s)
10^{-3}	2.0061×10^{-9}	$(2.0058 \times 10^{-9}, 2.0064 \times 10^{-9})$	1.50×10^{-10}	$\mathbf{7.45 \times 10^{-2}}$	269.17
10^{-5}	2.0001×10^{-15}	$(2.0000 \times 10^{-15}, 2.0001 \times 10^{-15})$	1.31×10^{-17}	$\mathbf{6.56 \times 10^{-3}}$	270.39
10^{-7}	2.0000×10^{-21}	$(2.0000 \times 10^{-21}, 2.0000 \times 10^{-21})$	2.00×10^{-28}	$\mathbf{1.00 \times 10^{-7}}$	270.06
10^{-9}	2.0000×10^{-27}	$(2.0000 \times 10^{-27}, 2.0000 \times 10^{-27})$	2.00×10^{-36}	$\mathbf{1.00 \times 10^{-9}}$	268.21
10^{-11}	2.0000×10^{-33}	$(2.0000 \times 10^{-33}, 2.0000 \times 10^{-33})$	2.00×10^{-44}	$\mathbf{1.00 \times 10^{-11}}$	269.44

Table 3.6 Static Network: Simulation results from the IS scheme for a Dodecahedron (Heterogeneous).

ε	Estimate	95 % CI	STD	**R.E.**	Time (s)
10^{-3}	1.3916×10^{-7}	$(1.3904 \times 10^{-7}, 1.3927 \times 10^{-7})$	5.78×10^{-8}	$\mathbf{4.16 \times 10^{-1}}$	118.76
10^{-5}	3.2204×10^{-12}	$(3.2193 \times 10^{-12}, 3.2216 \times 10^{-12})$	5.86×10^{-13}	$\mathbf{1.82 \times 10^{-1}}$	75.50
10^{-7}	7.9720×10^{-17}	$(7.9706 \times 10^{-17}, 7.9734 \times 10^{-17})$	7.05×10^{-18}	$\mathbf{8.84 \times 10^{-2}}$	66.45
10^{-9}	1.9967×10^{-21}	$(1.9965 \times 10^{-21}, 1.9968 \times 10^{-21})$	8.93×10^{-23}	$\mathbf{4.47 \times 10^{-2}}$	63.30
10^{-11}	5.0125×10^{-26}	$(5.0123 \times 10^{-26}, 5.0127 \times 10^{-26})$	1.14×10^{-27}	$\mathbf{2.27 \times 10^{-2}}$	63.90

We considered different ways of ordering of nodes in our analysis for the homogeneous case in the graph model of Figure 3.2 and in all the cases of enumerations the VRE property was observed, as is also shown in Table 3.5 for this particular enumeration. Also, for the heterogeneous case, VRE property is also observed, as is shown in Table 3.6. However, it is noticeable in both Tables 3.5 & 3.6 that the rate of decrease of the RE is slower in the specific heterogeneous case tried here as compared to the homogeneous case. For both the cases, the CMC method failed to record a single failure event for $\varepsilon < 10^{-2}$, hence yielding a useless $(0,0)$ empirical 95% CI.

Comparing the results with CMC method, the average per unit computation time for the proposed IS scheme in the homogeneous case is 2.6945×10^{-4} seconds, while CMC method took approximately 4.7000×10^{-6} seconds per iteration. From the estimated $\hat{u}$, for example for a rare event $\varepsilon = 10^{-7}$, var_{wn} obtained from CMC method is 9.4000×10^{-27} and with the proposed IS scheme is much lower ($\tilde{var}_{wn} = 1.0778 \times 10^{-59}$). As $\varepsilon \longrightarrow 0$, var_{wn} of the IS scheme decreases much faster than the CMC method.

Similarly, in the heterogeneous case, the proposed IS scheme took 7.7582×10^{-5} seconds per iteration while the CMC method took 5.1000×10^{-6} seconds per iteration. Now, for $\varepsilon = 10^{-7}$, the var_{wn} obtained from the CMC method is 4.0657×10^{-22} and for the IS scheme it is $\tilde{var}_{wn} = 3.8563 \times 10^{-39}$. It is clear that $\tilde{var}_{wn} << var_{wn}$ and along with the VRE property being observed, the proposed IS scheme works very well to estimate rare event probabilities here.

Example 3: The third network considered is a much larger network [29] where three dodecahedrons of Figure 3.2 are juxtaposed in a parallel configuration as shown in Figure 3.3. The source s and terminal t of Figure 3.2 are merged and represented by a single node, s and t respectively. The topology has 56 nodes and 90 links.

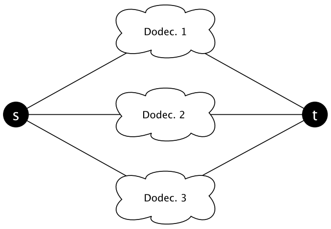

Figure 3.3 Three dodecahedrons connected in parallel.

For the purpose of conciseness of the document here, only the homogeneous case is considered here. It is assumed that $q_i = \varepsilon$ for all nodes in the Figure 3.3, and the goal is to compute the unreliability that the s and t are not connected. The unreliability obtained here is a cube of the unreliability of a single dodecahedron for the case of node failures [29]. Results from simulations using the proposed IS scheme based on mincuts is presented in Table 3.7.

Table 3.7 Static Network: Simulation results from the IS scheme for three Dodecahedrons in parallel configuration (Homogeneous).

ε	Estimate	95 % CI	STD	**R.E.**	Time (s)
10^{-3}	8.0739×10^{-27}	$(8.0714 \times 10^{-27}, 8.0764 \times 10^{-27})$	1.27×10^{-27}	$\mathbf{1.57 \times 10^{-1}}$	5406.23
10^{-5}	8.0005×10^{-45}	$(8.0003 \times 10^{-45}, 8.0007 \times 10^{-45})$	1.03×10^{-46}	$\mathbf{1.28 \times 10^{-2}}$	5572.21
10^{-7}	8.0000×10^{-63}	$(8.0000 \times 10^{-63}, 8.0000 \times 10^{-63})$	8.00×10^{-66}	$\mathbf{1.00 \times 10^{-3}}$	5585.42
10^{-9}	8.0000×10^{-81}	$(8.0000 \times 10^{-81}, 8.0000 \times 10^{-81})$	1.39×10^{-89}	$\mathbf{1.73 \times 10^{-9}}$	5492.28
10^{-11}	8.0000×10^{-99}	$(8.0000 \times 10^{-99}, 8.0000 \times 10^{-99})$	1.38×10^{-109}	$\mathbf{1.72 \times 10^{-11}}$	5617.49

It is observable that the unreliability estimates in the case of node failures, as shown in Table 3.7 are of the same order of magnitude for both Example 2 (single dodecahedron) and Example 3 (three parallel dodecahedrons) as it is for the case of link failures obtained by [29]. The empirical results from Table 3.7 show that the VRE property holds for this example. The CMC method again failed to record any failure event for all $\varepsilon \leq 10^{-1}$. Comparing with CMC method, for example for $\varepsilon = 10^{-7}$ rare event, var_{wn} for CMC method is 1.1770×10^{-67} (average per run computation time $= 1.4713 \times 10^{-5}$ seconds). For the same case, with the IS scheme $\tilde{var}_{wn}$ is 3.5422×10^{-133} (average per run computation time $= 5.5347 \times 10^{-3}$ seconds). The $\tilde{var}_{wn}$ for IS scheme decreases (as $\varepsilon \longrightarrow 0$) much rapidly compared to the one that is obtained from CMC method and with the VRE property observed here, the gain $\rightarrow \infty$ with the IS scheme as compared to the CMC method. Also, the order of magnitude of the $\hat{u}$ for all $\varepsilon \leq 10^{-3}$ are unrealistically low if practicality is considered, however, it also proves that the IS scheme here would be able to estimate a very high accuracy in relatively lesser number of samples than $n = 10^6$ here.

3.4.3 Practical Case Study: Data Communication System

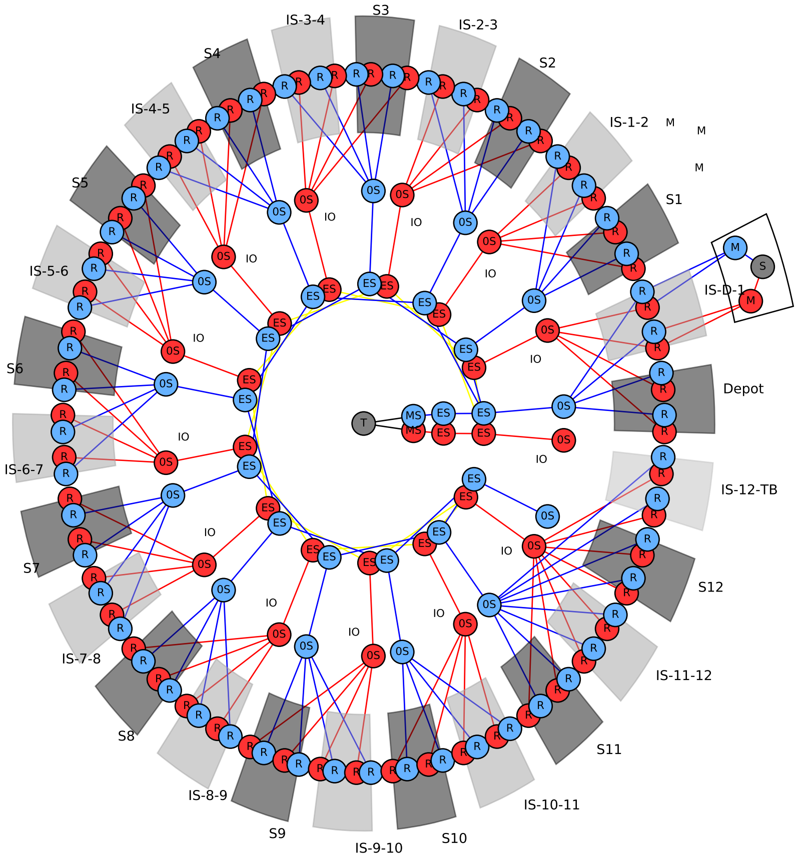

Figure 3.4 The DCS structure with the outermost and the innermost node representing the train (source) and the Zone Controller (terminal) respectively.

We now consider a real Data Communication System (DCS), a part of a large scale passenger rail system *Communication Based Train Control* (CBTC). The role of a DCS is to carry without hindrance the data between various other rail systems ensuring end-to-end communication. It consists of reliable and redundant communication paths, as shown by RED and BLUE in Figure 3.4. More detailed description is provided in the following section.

Description of the DCS system

In Figure 3.5, the idea behind the structure of the DCS is represented for a small section between two stations and comprising of subsystems. In both the Figures 3.4 & 3.5, the links represent wired or wireless communication channels between ground-to-ground or ground-to-rolling stock (train), respectively. The nodes represent ethernet or electrical switches, routers, servers, radio equipment, modems, etc. The red and blue components (links and nodes) are in pairs, and are in UP state all the time (active redundancy) such that there is no switching of functioning if one of them fails. This also adds a complexity of undetected failure. However, the redundancy of red and blue makes certain that there are two independent communication paths available all the time.

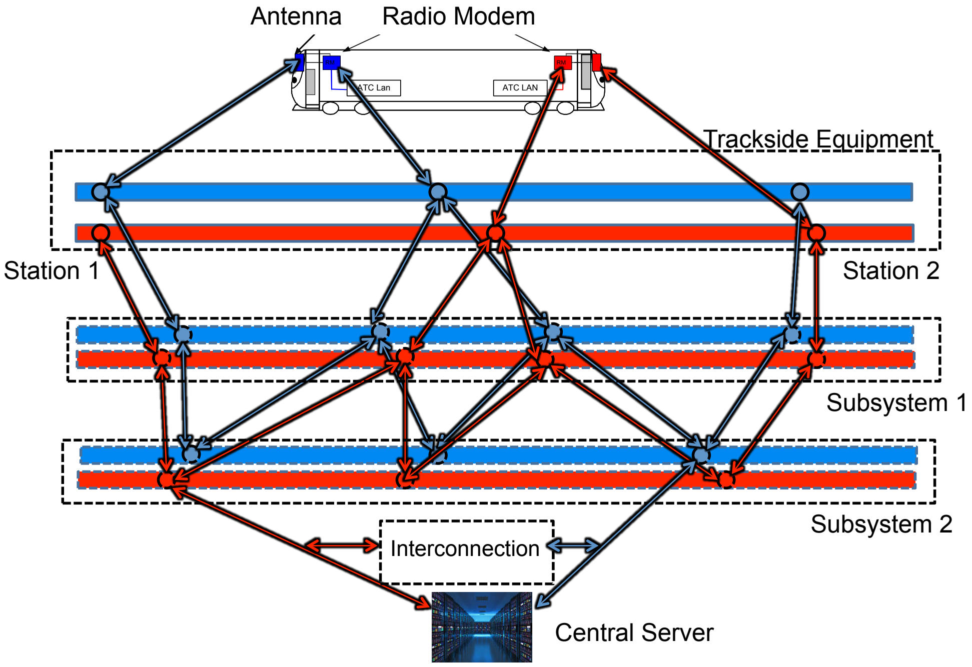

Figure 3.5 Schematic representation of a DCS section.

For this case study, the train is considered as the source s (outermost node) and the Zone Controller server is considered as the terminal t (innermost node), as shown in Figure 3.4. The circular topology is for the purpose of simplicity for the readers. The nodes are enumerated start from the source (s) as 0 up to the terminal (t) as 163, counter-clockwise for each ring. In the DCS, the outer circle of nodes represent the trackside equipment which communicates directly with the train through overlapping wireless radio access points' coverage, as represented in Figure 3.5. A failure of more than three consecutive pairs of red and blue nodes will make the s and t disconnected. Also, another important characteristic of it is that there is an interconnection between the red and blue rings by connecting the two respective red and blue nodes just before the terminal together, as also shown in Figure 3.5. This way, a red ring can also use the blue ring and vice-versa in case of failure of a particular ring.

In the analysis, for simplicity, there is only one train and it is considered to be at a fixed position as shown in Figure 3.4. Practically there are more than one train present (i.e., multiple sources s) along the outer circle. Also, both the homogeneous and heterogeneous cases are considered, where in the first case all the nodes have the same unreliability ε while in the second case, their unreliability is not the same. There are 164 nodes and 169 links in the graph model.

Results of the case study

The empirical results (with $n = 10^6$ iterations for each simulation) for the homogeneous case using the proposed IS scheme are shown in Table 3.8 and the results from simple CMC simulations are presented in Table 3.9. Results from Table 3.8 show that the BRE property holds when $\varepsilon \longrightarrow 0$, and we obtain tight bounds over the estimated $\hat{u}$. The results from this static model give the steady-state availability of the system. From the estimated $\hat{u}$, for example when $\varepsilon = 10^{-7}$, the var_{wn} for CMC method is 5.6367×10^{-18} (average per run computation time being 3.5300×10^{-5} seconds, see Table 3.9). For the proposed IS scheme, the var_{wn} is 8.4406×10^{-28} (average per run computation time being 7.3882×10^{-3} seconds, in Table 3.8), which is much lower and also decreases much rapidly as $\varepsilon \longrightarrow 0$.

Table 3.8 Static Network: Simulation results from the IS scheme for the DCS (Homogeneous).

ε	Estimate	95 % CI	STD	**R.E.**	Time (s)
10^{-3}	1.5954×10^{-5}	$(1.5888 \times 10^{-5}, 1.6020 \times 10^{-5})$	3.38×10^{-5}	**2.12**	7327.19
10^{-5}	1.5968×10^{-9}	$(1.5902 \times 10^{-9}, 1.6034 \times 10^{-9})$	3.38×10^{-9}	**2.11**	7417.07
10^{-7}	1.5968×10^{-13}	$(1.5902 \times 10^{-13}, 1.6034 \times 10^{-13})$	3.38×10^{-13}	**2.11**	7433.32
10^{-9}	1.5968×10^{-17}	$(1.5902 \times 10^{-17}, 1.6034 \times 10^{-17})$	3.38×10^{-17}	**2.11**	7378.22
10^{-11}	1.5968×10^{-21}	$(1.5902 \times 10^{-21}, 1.6034 \times 10^{-21})$	3.38×10^{-21}	**2.11**	7385.41

Table 3.9 Static Network: Simulation results from CMC for the DCS (Homogeneous).

ε	Estimate	95 % CI	STD	R.E.	Time (s)
10^{-3}	7.0000×10^{-6}	$(1.8143 \times 10^{-6}, 1.2186 \times 10^{-5})$	2.65×10^{-3}	377.96	35.64
10^{-5}	0.0	(0.0, 0.0)	0.0	–	34.96

The analysis for the heterogeneous case assumes the following for the unreliability of nodes:

- Nodes that are enumerated as multiples of 5 are considered to have unreliability $\varepsilon^{1.15}$.
- Nodes that are enumerated as multiples of 11 and are not a multiple of 5 are considered to have unreliability $\varepsilon^{0.85}$.
- Rest all other nodes are considered to have unreliability of $\varepsilon^{0.65}$.

The nodes of the graph model DCS network are assigned unreliabilities as per the above assumptions and the results from simulations (with $n = 10^6$ iterations for each simulation) are shown for the IS scheme (in Table 3.10) and the CMC simulations (in Table 3.11).

Table 3.10 Static Network: Simulation results from the IS scheme for the DCS (Heterogeneous).

ε	Estimate	95 % CI	STD	**R.E.**	Time (s)
10^{-3}	1.1530×10^{-3}	$(1.1497 \times 10^{-3}, 1.1564 \times 10^{-3})$	1.69×10^{-3}	**1.47**	7428.98
10^{-5}	2.8523×10^{-6}	$(2.8442 \times 10^{-6}, 2.8603 \times 10^{-6})$	4.13×10^{-6}	**1.45**	7994.68
10^{-7}	7.1495×10^{-9}	$(7.1292 \times 10^{-9}, 7.1697 \times 10^{-9})$	1.03×10^{-8}	**1.44**	7432.22
10^{-9}	1.7956×10^{-11}	$(1.7905 \times 10^{-11}, 1.8006 \times 10^{-11})$	2.59×10^{-11}	**1.44**	7976.21
10^{-11}	4.5102×10^{-14}	$(4.4974 \times 10^{-14}, 4.5229 \times 10^{-14})$	6.51×10^{-14}	**1.44**	7762.18

Table 3.11 Static Network: Simulation results from CMC for the DCS (Heterogeneous).

ε	Estimate	95 % CI	STD	R.E.	Time (s)
10^{-3}	1.0940×10^{-3}	$(1.0292 \times 10^{-3}, 1.1588 \times 10^{-3})$	3.31×10^{-2}	30.22	35.26
10^{-5}	0.0	$(0.0, 0.0)$	0.0	–	34.35

The empirical results for this heterogeneous case using the IS scheme show that the BRE property holds (see Table 3.10) as $\varepsilon \to 0$ while a tight 95% CI is obtained for each ε too. Comparing the results with the CMC method (see Table 3.11), the CMC method took on average 3.4805×10^{-5} seconds per iteration while the IS scheme proposed here took on average 7.7189×10^{-3} seconds per iteration. Now as for previous examples, if we consider a possible rare event with $\varepsilon = 10^{-7}$, the IS scheme has $\tilde{var}_{wn} = 8.2322 \times 10^{-19}$, while on the other hand, the CMC method would have $var_{wn} = 2.4884 \times 10^{-13}$. Again, $\tilde{var}_{wn} << var_{wn}$ here too and the $\tilde{var}_{wn}$ decreases much rapidly as compared to var_{wn} of the CMC method as the EOI becomes rarer (i.e., $\varepsilon \to 0$).

From the results presented for *Examples 1-4* for both the homogeneous and the heterogeneous cases, it can be concluded that the proposed zero-variance IS scheme based on mincuts with maximal probability efficiently estimates rare event probabilities. The IS scheme here also adheres to the measures of accuracy (BRE property and in some cases VRE also), as previously discussed in Section 2.5. In terms of the quantified measure of efficiency using the work-normalized variance, the proposed IS scheme is highly efficient as compared to a standard CMC simulation when $\varepsilon \to 0$.

3.5 Conclusions from the Chapter

The motivation of the work (as discussed in Section 3.1) in the first place was to be able to estimate the unreliability u efficiently in static networks using IS for rare event probabilities. Considering the optimal change of measure in IS is unknown, an approximate zero-variance IS scheme based on mincuts is extended here for the case of node failures using the basis laid out in [29] for the case of link failures originally. The sequential sampling of nodes, as done by [29], reduces the variance by a large factor. We prove that the methodology explained here and by [29], works for the case of node failures also and illustrate its efficiency on a real network, a Data Communication System used in urban train control.

It is also important to observe that the zero-variance IS scheme is more computationally burdensome compared to CMC methods, as it needs to find two mincuts with maximal probability at each step of the sampling process [29] using a Ford-Fulkerson adapted algorithm. However, the method estimates the unreliability u with a higher accuracy (variance reduction) at the expense of increased computation time. This is a trade off between choosing a more precise estimate or a faster estimate with huge variance. With respect to rare event analysis, the quantified efficiency measure of work normalized variance var_{wn}, which gives the estimate of variance reduction with respect to the cost (i.e., time), the proposed zero-variance IS scheme is highly efficient compared to the CMC method.

Chapter 4

Cross-Entropy Application to Highly Reliable Markovian Systems

In this chapter, the focus is on the estimation of the steady-state unavailability of Highly Reliable Markovian Systems (HRMS). The chapter illustrates the use of Stochastic Petri Nets (SPNs) to represent complex systems conveniently, while also explaining how Markovian SPNs represent the underlying continuous-time Markov Chains (CTMCs). In the examples considered here, we also include complex logistic aspects for representing real passenger rail systems closely. The optimization technique based on Cross-Entropy (CE) minimization is proposed to optimize/find the IS rates of Markovian SPNs' failure transitions and efficiently estimate the steady-state unavailability via simulations.

4.1 Motivation and Objectives

In the previous chapter, we discussed the application of the approximate zero-variance IS scheme based on mincuts for static networks, where time plays no role. The earlier assumptions considered that the components of the systems could only be in *working* or *failed* states and their respective states are independent of time. In this chapter, we examine dynamic systems where the state of the system changes over time.

In dynamic systems, contrary to static networks, components can be repaired (or restored) to operational states, while protocols of logistics, maintenance, etc., can also play a role. These systems are repairable such that the internal components or even the entire system, after undergoing a failure can be restored to fully satisfactory performance by a method other than the replacement of the entire system [81, 82]. In addition to this, in literature, maintenance actions which aim at servicing the systems for better performance are also included [82]. Other practical aspects in such systems that can be considered are: timed inspections, maintenance actions (preventive or corrective), availability of spares and (or) repair personnel in the depot, travel time for on-site operations of maintenance/repair/inspection, etc. All these practical aspects, when considered together, make the system under consideration (and the resulting mathematical models) very complex.

In the analysis of the complex systems with repair or maintenance possibilities, the choice of an appropriate reliability metric is essential. As we discussed before, for rail system suppliers (such as Alstom), the chosen reliability metric should be helpful in determining the LCC (Life Cycle Cost) of the offered solutions. The reliability metric can also help in making well-informed purchasing decisions for performance-based contracting, as previously discussed [1]. When we consider the case where the total duration of failures is crucial, the choice of the appropriate reliability metric for RAM analysis would be the *availability* of a system [7]. Also, the metric of availability (or unavailability) is very helpful in estimating costs associated with the loss of income due to the outage of a system [3]. Thus, here we can consider the availability of a system as a useful metric of interest in determining the LCC.

The definition of availability from a qualitative point of view is the ability of a component/system to be operational when required for sure [75]. From a quantitative point of view, it is a probability of finding a component/system in the operational state at an arbitrary point in time [75]. Some well-known availability metrics of interest are steady-state availability, time-dependent availability, mission availability, overall availability, etc. In the current work, the focus is on the estimation of the cumulative steady-state unavailability (or contrarily the availability) which is the equilibrium behavior of the system. In mathematical terminology, the steady-state unavailability (let's say U) of a system is the long-run fraction of time the system is in the down (i.e., failed) state [10], such that:

$$U = \lim_{t\to\infty} \frac{1}{t} \int_0^t 1\{X(t) \in \mathscr{D}\} dt,$$

where $X(t)$ represents a specific state of the system at a given time t and $1\{\cdot\}$ is an indicator function when $X(t)$ (i.e., the system) is in a failed state $\mathscr{D}$. We explain this definition in more detail in the later sections.

In the current work, we consider highly reliable complex systems with Markovian assumptions (i.e., HRMS), where exponential laws govern the distribution of holding times in any state. The exponential distributions are memoryless, the quintessential property of any Markov chain. The approach of Markov modeling can overcome the limitation of dependencies encountered when using RBDs or FTAs. However, Markov modeling also suffers from a significant drawback: the largeness of their state space [83]. Generalized Stochastic Petri Nets (GSPNs) can be used in such cases to generate a large underlying Markov process automatically starting from a concise description [83]. Also, Petri Nets (PNs) in general are useful to model and visualize different behaviors [12]. For this purpose, Markovian SPN models are used in the current work to conveniently represent the complex systems and their respective underlying CTMCs [84].

When studying/analyzing HRMS models, the system failures are rare events and justify the use of IS techniques, where the optimal change of measure is unknown. For this purpose, in the current chapter, we propose a multi-level CE optimization scheme (as previously mentioned for optimization within a parametric class in Section 2.4.1). The idea is to exploit the regenerative structure of the underlying CTMC and our proposed method aims to determine the optimal IS rates for rare event simulations. The CE scheme is used in a pre-simulation and applied to failure transitions of the Markovian SPN models only. The proposed CE method divides a rare problem into a series of less rare sub-problems by first increasing the failure rates of the components at the first stage, and thus, creating an unstable system. In the subsequent stages, we decrease the failure rates of the first sub-problem, forming new and gradually rarer sub-problems, until we reach the original rare problem. During the first stage, we perform a standard regenerative simulation for the non-rare system failures, where IS rates are same as the rates of the first sub-problem (likelihood ratio being one in this case). The CE update equation proposed here captures the contributions of the respective transitions towards the system failure (i.e., the EOI). At each subsequent stage, we progressively increase the rarity as mentioned above, while using the IS rates of transitions obtained from the previous sub-problem in the current stage. The final pre-simulation stage provides a vector of IS rates that are optimized, and we use them in the main simulation. In general, empirical results show the BRE property as the rarity of the original problem increases, and as a consequence a considerable variance reduction and gain also.

The next sections explain the idea and specifics of our proposed method in more detail. Section 4.2 describes the mathematical model for the HRMS considered here, while also explaining the reliability metric of steady-state unavailability in the current context. Also, it briefly illustrates the use of regenerative IS simulations for estimation of the steady-state unavailability here and lays the foundation of the present work in terms of Cross-Entropy optimization. In Section 4.3, we discuss the use of SPNs for a compact representation of complex systems and the underlying CTMCs, while also explaining the computation of the likelihood ratio for different cases of the IS change of measure. The section also presents the update equation used for the multi-level CE optimization scheme and the algorithm that we propose for this work. We present the numerical results obtained for different examples in Sections 4.4-4.6, and finally draw the conclusions of the chapter in Section 4.7.

4.2 Mathematical Model of HRMS

The mathematical model for the HRMS considered here are in the form of discrete space CTMCs. We specifically use the regenerative structure of the CTMC models here to perform the stochastic simulations [25]. In a Markov chain, the associated random variable is a function of the sample path of a Markov chain [22].

Let us consider a system with c types of components with a total n_l number of components of each type. The total number of components are $\mathbb{N} = \sum_{l=1}^{c} n_l$. A system can fail if sufficient combination of components of each type fail [85, 86]. The system can be modeled as a CTMC where the state of the chain at time t is given by the vector:

$$X(t) = (X_1(t), X_2(t), X_3(t), \ldots, X_c(t)).$$

Here, $X(t)$ is a vector of the number of failed components of type $l = 1, \ldots, c$ at time t. The states are denoted by c-dimensional vectors $x = (x_1, x_2, x_3, \ldots, x_c)$. The perfect state is $\mathbf{\{0\}} = (0,0,0,0,\ldots,0)$ representing all components of all types are working and zero failed components [85, 86, 87]. We also assume that the finite state space $\mathscr{K}$ is divided into set of UP states $\mathscr{U}$ and set of DOWN (failed) states $\mathscr{D}$, such that $\mathscr{K} = \mathscr{U} \cup \mathscr{D}$, $\mathscr{D} \neq \emptyset$ and $\mathscr{U} \cap \mathscr{D} = \emptyset$. This is a general formulation of a CTMC model as given in [10, 85, 86, 87]. Some other general assumptions are:

A.1 There is no failure propagation.

A.2 The CTMC $\{X(t) : t \geq 0\}$ is irreducible over its state space.

A.3 From the state $\{\mathbf{0}\}$, there is at least one failure transition and from any other state of $\mathscr{K}$ except $\{\mathbf{0}\}$, there is at least one repair transition with a positive probability.

A.4 From any UP state $\mathscr{U}$ of the system, there is a positive probability to have a failure transition.

From the above formulation, the continuous-time stochastic process given by $\{X(t) : t \geq 0\}$, evolves in continuous-time over the finite state space $\mathscr{K}$. It is obvious that every state of a CTMC is a regenerative state due to the memoryless property of exponential distributions of holding times in a given state. As per the regenerative process theory [10, 33], the evolution of the process from $\{\mathbf{0}\}$ and back to $\{\mathbf{0}\}$ is called a *cycle*. The stochastic evolution of the system is independent of its past, has the same distribution as if the system actually started in the state $\{\mathbf{0}\}$ and we can obtain independent and identically distributed (iid) samples [17]. Thus, assuming the system to return to $\{\mathbf{0}\}$ infinitely often [17] and visited the most (i.e., more regeneration in highly reliable systems), we base our choice of $\{\mathbf{0}\}$ as the regeneration state. This is the basic notion behind regenerative simulations.

Let us also consider:

$$X_i = X(t_i^+), \text{ and } \{i = 0, 1, 2, ..., \tau - 1\}, \tag{4.1}$$

where the process is assumed to be right continuous and we consider the embedded discrete times of the CTMC $\{X(t) : t \geq 0\}$. The state is X_i at a given time and t_i is the time of the $i - th$ change of state. We adopt the convention that $t_0^+ = 0$ [88]. From this, a single cycle (or the sample path) of this CTMC can be written as:

$$\omega = (X_0, V_0), (X_1, V_1), ..., (X_{\tau-1}, V_{\tau-1}).$$

Here, X_i is as given in Equation 4.1 denoting the state of the CTMC at a given time, starting always from $\{\mathbf{0}\}$, and $\tau - 1$ corresponds to the last change of state before re-entering back into $\{\mathbf{0}\}$. V_{x_i} is the sojourn time in a given state X_i. The last state (X_τ) is in fact the regeneration state $X_\tau = \{\mathbf{0}\}$, so the change of state from $X_{\tau-1} \longrightarrow X_\tau = \{\mathbf{0}\}$ would be caused by only a repair action of a failed component at $t_{\tau-1}^+$. For any CTMC, we also consider the system to be balanced or unbalanced in terms of the probability of a sample path from the initial state $\{\mathbf{0}\}$ to the failure set of state $\{\mathscr{D}\}$ of the system. The formal definition is given in the following section.

4.2.1 Balanced or Unbalanced Systems

In an asymptotic regime, let us consider that the rarity of the EOI is increased by a rarity parameter λ (i.e., $\lambda \to 0$). Let us consider a system starting from the initial state $\{\mathbf{0}\}$ has two different paths to the system failure state $\{\mathscr{D}\}$. The probability of the two paths is a function of the rarity parameter given by failure rates of components λ. Formally, we can say $p_1(\lambda), p_2(\lambda) : (0,1) \to \mathbb{R}$. We define below a system to be balanced or unbalanced in terms of the relative growth of the functions $p_1(\lambda)$ and $p_2(\lambda)$ asymptotically (i.e., $\lambda \to 0^+$).

- For a balanced system: $p_1(\lambda) = O(p_2(\lambda))$ if $|p_1(\lambda)| \leq c_1 p_2(\lambda)$ for some constant $c_1 > 0$ or $p_1(\lambda) = \underline{O}(p_2(\lambda))$ if $|p_1(\lambda)| \geq c_2 p_2(\lambda)$ for some constant $c_2 > 0$. For a balanced system, we can say $p_1(\lambda) = \Theta(p_2(\lambda))$, where $p_1(\lambda) = O(p_2(\lambda))$ and $p_1(\lambda) = \underline{O}(p_2(\lambda))$ both for $c_1, c_2 > 0$. (Here $\Theta(\cdot)$ is not to be confused with the parameter space Θ used before).
- For an unbalanced system: $p_1(\lambda) = o(p_2(\lambda))$ if $\lim_{\lambda \to 0^+} p_1(\lambda)/p_2(\lambda) = 0$. This means the probability of a path given by $p_1(\lambda)$ decreases much faster as compared to $p_2(\lambda)$.

The above description of a balanced or unbalanced system is slightly different from the one given in literature. For example, in [47, 87, 89] systems are considered to be balanced (or unbalanced) if the failure rates of components are of the same order of magnitude (or not). Another description of balanced (or unbalanced) systems is given in [10], where balanced systems are those in which components have the same amount of redundancy, i.e., same number of components of a particular type must fail for a system failure (such as 1oo2 of one type or 2oo3 of another type). In the current work, the term balanced or unbalanced are used for systems as per the formal definition given above.

4.2.2 Steady-State Measure

The goal here is to compute the steady-state measure, namely the *cumulative steady-state unavailability* (U). Steady-state measures are independent of the starting state of the system (here it is specifically **{0}**) and the system unavailability is the long run fraction of time the system is in the down state [10]. Formally:

$$U = \lim_{t \to \infty} \frac{1}{t} \int_0^t 1\{X(t) \in \mathscr{D}\} dt,$$

where $1\{\cdot\}$ is an indicator function when $X(t)$ is in the failure set $\mathscr{D}$. Let us define another random variable Z which is a (measurable) function of $X(t)$, given as:

$$Z = \int_0^\tau 1(X(t) \in \mathscr{D}) dt,$$

and means the time spent by $X(t)$ in the set $\mathscr{D}$ during a cycle [87].

If W is the length of a cycle, then $\mathbb{E}[W]$ is the expected (or average) time (or length) of a cycle. Then U is a ratio of two expectations [10]:

$$U = \frac{\mathbb{E}[Z]}{\mathbb{E}[W]} = \frac{\int_0^\tau 1(X(t) \in \mathscr{D})dt}{\mathbb{E}[W]}. \tag{4.2}$$

A.5 Let us assume that $\mathbb{E}[Z] < \infty$ and $\mathbb{E}[W] < \infty$ where $\mathbb{E}$ is the expectation under the original measure $\mathbb{P}$ having a generator matrix Q [10, 17].

4.2.3 Steady-State Unavailability Estimation of HRMS

The focus of the study is to estimate the steady-state unavailability U of HRMS via simulations. As already explained previously (see Sections 2.1 & 2.3), the following sections re-introduces the background in the context of regenerative process theory. Also, we discuss the use of standard regenerative MC method and the regenerative IS for simulations of CTMCs.

Standard regenerative MC simulations of CTMC

The iid structure of the regenerative processes and the Equation 4.2 together form the basis of the regenerative method [17]. The standard estimator of the U from a regenerative standard MC simulation is given by:

$$\hat{U} = \frac{\hat{Z}_n}{\hat{W}_n},$$

where $\hat{Z}_n$ and $\hat{W}_n$ are the respective averages over n cycles [10]. As per the law of large numbers, if n is large enough, then $\hat{U} \to U$ as $n \to \infty$. Cycles are of particular interest to build CI (due to their stochastic independence) which in this case using the CLT [17] is given as:

$$\sqrt{n}\frac{\hat{U}_n - U}{\hat{\sigma}/\hat{W}_n},$$

and is asymptotically distributed as a normal distribution $\mathscr{N}(0,1)$ with mean 0 and variance 1 [14]. The steady-state simulation is to efficiently estimate U by its estimator $\hat{U}$ and to develop the associated CI [90].

However, when a system is highly reliable (i.e., failures are very rare), then the standard MC simulation's estimator $\hat{U}$ would not be an accurate estimate as the occurrence of rare event (e.g., $1(X(t) \in \mathscr{D})$ would not happen, and $Z = 0$ in most cycles for this case. This lack of useful samples could result in a huge variance (sometimes even bad or no estimation at all) or an increasingly high RE (relative error) for the estimator (recall the measures of accuracy in rare event simulations, see Section 2.5). The denominator $\mathbb{E}[W]$ is the expected time of the regenerative cycles which is easier to estimate by a standard simulation even.

IS is a viable option in such cases to obtain an alternative estimator of $\mathbb{E}[Z]$ by changing the original probability measure $\mathbb{P}$ with a new one $\tilde{\mathbb{P}}$ for sampling and increasing the occurrence of rare failure events in the HRMS during simulations. We present the basics of the regenerative IS simulations in the next section.

Importance sampling simulations of CTMC

In the introductory Section 2.3.1, we explained the concept of IS in a general context. Let us consider here that the density of the sample path of the CTMC defined here is $f(x;\theta)$, with parameter vector θ under the original measure $\mathbb{P}$. The vector θ is in the parameter space Θ, and for CTMCs, it is composed of the transition rates between different states of the CTMC. The density (or likelihood) of a sample path of the CTMC for regenerative simulation can be written as [22]:

$$f(x;\theta) = \prod_{i=0}^{\tau-1} q_{x_i,x_{i+1}} \exp\left[-\sum_{i=0}^{\tau-1} q_{x_i} v_{x_i}\right]. \tag{4.3}$$

Here, the term q denotes the rate. The probability of moving from a generic state $x_i \longrightarrow x_{i+1}$ is given by the ratio of the jump rate from x_i to x_{i+1} and the sum of all the departure rates from state x_i. The density for moving from state x_i to x_{i+1} would be: $q_{x_i} \exp[-q_{x_i} \cdot v_{x_i}]$, where v_{x_i} is the time spent in the particular state x_i for each transition. The density of the entire sample path is thus the product of density of each transition from $i = 0$ to $\tau - 1$.

For application of IS within the CTMC formulation given here, we generally replace the original probability measure $\mathbb{P}$ by a new one $\tilde{\mathbb{P}}$. Here, we apply the IS change of measure within the same parametric family (f) by changing the parameter vector from θ (under $\mathbb{P}$) to $\tilde{\theta}$ (under $\tilde{\mathbb{P}}$). The condition that any non-zero sample under $f(x;\theta)$ also remains a non-zero possibility under $f(x;\tilde{\theta})$ must be met. The general equation for the likelihood ratio of a sample path ω is then:

$$L(\omega) = \frac{f(x;\theta)}{f(x;\tilde{\theta})} = \prod_{i=0}^{\tau-1} \frac{q_{x_i,x_{i+1}}}{\tilde{q}_{x_i,x_{i+1}}} \exp\left[\sum_{i=0}^{\tau-1} (\tilde{q}_{x_i} - q_{x_i}) v_{x_i}\right]. \tag{4.4}$$

The density of each transition and the entire sample path (cycle) under the change of measure $f(x;\tilde{\theta})$ is computed the same way as for the original measure (as given in Equation 4.3). The likelihood ratio $L(\omega)$ is computed as the ratio of the original and new densities at each state change [17, 22].

The expectation under IS is now given as:

$$\mathbb{E}_{\tilde{\theta}}[Z\, L_{\omega}] = \mathbb{E}_{\theta}[Z] = Z, \tag{4.5}$$

under the assumption that $\mathbb{E}_{\tilde{\theta}}[W] < \infty$ under $\tilde{\mathbb{P}}$. For clarity, the expectation operator is suffixed with the parameter vector under which the expectation is taken. Here, $\mathbb{E}_{\theta}$ is the expectation under the original density $f(x;\theta)$ (under probability measure $\mathbb{P}$) and $\mathbb{E}_{\tilde{\theta}}$ is under the IS density $f(x;\tilde{\theta})$ (under new probability measure $\tilde{\mathbb{P}}$). Obviously, the goal of using IS is to obtain variance reduction in the final estimation of U, which here depends on accurate estimation of Z, the numerator in Equation 4.2. This leads us to the possibility of a zero-variance estimator of Z also.

Zero-Variance Estimator: The theoretical optimal change of measure (i.e., the zero-variance density), as explained by Equation 2.8 previously, in this case is:

$$g^*(x) = \frac{f(x;\theta)\cdot|Z|}{Z}. \tag{4.6}$$

This $g^*(x)$ is the conditional density given the rare event occurs (i.e., $|Z| > 0$) but again requires the knowledge of Z, the original problem to be estimated accurately.

To solve this problem of approximating the optimal change of measure, we use the idea of minimizing the CE distance between the zero-variance density and the IS density used, as explained in the following section.

4.2.4 Cross-Entropy for HRMS

The CE method, as previously discussed in a general context in Section 2.4.1, could be utilized to find a density closest to the zero-variance IS density $g^*(x)$ in Equation 4.6. The main objective is to reduce the variance of the final estimator of U, by accurately estimating the numerator Z using IS. Previously in Section 2.4.1, we discussed that the IS density closest to $g^*(x)$ in terms of CE distance (let us say with the optimizing parameter vector $\tilde{\theta}^*_{ce}$) is also the one for which the asymptotic variance of the estimator is minimum [32]. We employ this idea here.

The same parametric family as the original measure (represented by the notation f) is used with the idea to minimize the CE distance between $g^*(x)$ and the IS density $f(x;\tilde{\theta})$. Following the same analogy as previously presented (in Section 2.4.1, see Equations 2.15–2.19), the CE distance between the two densities here is given by [18, 22, 31]:

$$\mathscr{D}(g^*(x), f(x;\tilde{\theta})) = \mathbb{E}_{g^*}\left[log\frac{g^*(x)}{f(x;\tilde{\theta})}\right]. \tag{4.7}$$

Replacing $g^*(x)$ by its true value, it is equivalent to [22]:

$$\mathscr{D}(g^*(x), f(x;\tilde{\theta})) = \mathbb{E}_\theta\left[\frac{|Z|}{\mathbb{E}_\theta[|Z|]}log\left(\frac{|Z|}{\mathbb{E}_\theta[|Z|]}f(x;\theta)\right)\right] - \frac{1}{\mathbb{E}_\theta[|Z|]}\underbrace{\mathbb{E}_\theta\left[|Z|\;log\,f(x;\tilde{\theta})\right]}_{\text{to maximize}}, \tag{4.8}$$

where the expectations are taken with respect to the density $f(\cdot;\theta)$. To minimize the CE distance between $g^*(x)$ and $f(x;\tilde{\theta})$ in the above Equation 4.8, it means to maximize the term $\mathbb{E}_\theta\left[|Z|\;log\,f(x;\tilde{\theta})\right]$ (all other terms are constants) depending on the $f(\cdot;\tilde{\theta})$. The optimizing parameter vector, lets say $\tilde{\theta}^*_{ce}$, forms the density $f(x;\tilde{\theta}^*_{ce})$ that is closest to $g^*(x)$ in CE distance [22]. The problem is now transformed to a maximization problem, as shown below:

$$\max_{\tilde{\theta}\in\Theta} \upsilon(\tilde{\theta}) = \max_{\tilde{\theta}\in\Theta}\mathbb{E}_\theta\left[|Z|\;log\,f(x;\tilde{\theta})\right], \tag{4.9}$$

where υ is implicitly defined.

For the purpose of sampling, let us consider an IS density within the same parametric family, $f(x;\check{\theta})$ with the arbitrary reference parameter vector $\check{\theta}$ [31]. Now, the above Equation 4.9 can be re-written as:

$$\max_{\tilde{\theta}\in\Theta} \upsilon(\tilde{\theta}) = \max_{\tilde{\theta}\in\Theta} \mathbb{E}_{\check{\theta}} \left[|Z| \, L(X;\theta;\check{\theta}) \, log \, f(x;\tilde{\theta}) \right]. \tag{4.10}$$

The expectation is now taken under the IS density $f(\cdot;\check{\theta})$ and the likelihood ratio is the ratio of the respective densities $f(\cdot;\theta)$ and $f(\cdot;\check{\theta})$. Similar to Equation 2.19, the optimizer $\tilde{\theta}^*_{ce}$ is given by [31]:

$$\tilde{\theta}^*_{ce} = \arg\max_{\tilde{\theta}\in\Theta} \mathbb{E}_{\check{\theta}} \left[|Z| \, L(X;\theta;\check{\theta}) \, log \, f(x;\tilde{\theta}) \right]. \tag{4.11}$$

The above Equation 4.11 can not be solved analytically [31]. However, the vector $\tilde{\theta}^*_{ce}$ can be estimated by sample average approximation (the stochastic part) [31], as given below:

$$\tilde{\theta}^*_{ce} = \arg\max_{\tilde{\theta}\in\Theta} \frac{1}{n} \sum_{m=1}^{n} \left(Z_m(\omega_m) \, L(\omega_m;\theta;\check{\theta}) \, log \, f(\omega_m;\tilde{\theta}) \right). \tag{4.12}$$

As long as the above equation is convex and differentiable with respect to $\tilde{\theta}$ and $Z_m > 0$, the above equation can be solved through the following system of equations [31]:

$$\frac{1}{n} \sum_{m=1}^{n} \left(Z_m(\omega_m) \, L(\omega_m;\theta;\check{\theta}) \, \frac{\partial log \, f(\omega_m;\tilde{\theta})}{\partial \tilde{\theta}} \right) = 0. \tag{4.13}$$

Necessity of multi-level CE schemes

The CE optimization solution given by Equation 4.12 is applicable in case of non-rare event problems when $Z_m > 0$ under $\check{\theta}$ [31] via sample average approximation. In case of rare event problems, Z_m would be zero in most cycles for a small n. Also, since the CE scheme is supposed to be used in a pre-simulation to find the optimal IS rates that can be used for main simulations, the number of samples n can not be too large. Another important issue is the selection of the density $f(x;\check{\theta})$ with the parameter vector $\check{\theta}$. As previously discussed in Section 2.4.1, a good choice of $\check{\theta}$ is the one which leads to a sufficiently reliable optimizer of Equation 4.12 with less variance.

In order to overcome the above problem, we can use a *multilevel* CE scheme [22, 31]. The idea of the multi-level CE scheme is usually applied in an iterative manner, where at the start a model is solved that does not suffer from rare-event problems and subsequently the rarity is increased [22]. In a multilevel scheme (involving many pre-simulation stages), a sequence of $\check{\theta}^{(j)} = \check{\theta}^{(j=1,2,..)}_{ce}$ are chosen for sampling (where $\check{\theta}^{(j)}_{ce} \subset \Theta$) in several stages (e.g., j number of stages) and Equation 4.12 can be used to obtain:

$$\check{\theta}^{(j+1)}_{ce} = \arg\max_{\tilde{\theta}\in\Theta} \frac{1}{n_j} \sum_{m=1}^{n_j} Z_{j,m}(\omega_{j,m}) \, L(\omega_{j,m};\theta;\check{\theta}^{(j)}_{ce}) \, log \, f(\omega_{j,m};\tilde{\theta}). \tag{4.14}$$

At each subsequent stage $(j+1)$, n_j number of cycles $(\omega_{j,m}, m = 1,2,...,n_j)$ are simulated using $\check{\theta}_{ce}^{(j+1)}$ obtained from the current stage j as the parameter vector for the IS density. At each stage j, the sample size n_j is supposed to be smaller than N (the number of cycles to be simulated for the main simulation), but it should be large enough so that Equation 4.14 can be solved [86].

Our goal is to propose an algorithm based on the multi-level CE scheme given by Equation 4.14. Next sections discuss the modeling of the HRMS models using Markovian SPNs that represent large complex systems conveniently, and the use of the multi-level CE scheme (using the general Equation 4.14 as the basis) to find optimal IS rates.

4.3 Stochastic Petri Nets (SPN) Application

We previously discussed the mathematical formulation of CTMC models, in Section 4.2, that we use in the current work. RBDs and FTAs are the two main frameworks widely considered at the modeling phase for quantitative estimation of reliability metrics [91]. However, they can not represent dependencies occurring in real systems [83, 92]. To overcome this, Markov modeling approach is capable of capturing different kind of dependencies occurring in complex systems [83, 93, 94], but they also suffer from largeness of the state space [83] when used for even slightly large models. Since passenger rail systems are complex and large scale, it would be practically unfeasible to model such systems using Markov modeling approach. Also, when considering testing of various protocols for logistics and maintenance and its effect on the system unavailability, Markov modeling approach becomes too cumbersome. Therefore, for the purpose of accurate, concise and convenient representation of the complex models as well as the underlying CTMCs, as discussed in Section 4.2, we use Markovian Stochastic Petri Nets (SPNs) here. The objective is to first model complex systems using Markovian SPNs that also comprise of the underlying CTMC models and then to use the CE pre-simulation scheme (of Section 4.2.4) to find the optimal IS rates for the transitions of Markovian SPN models. The optimal IS rates are then used in a main simulation to estimate steady-state unavailability U.

In the next sections, we introduce various constructs of SPNs that make them very useful as a modeling tool. Also, the discussion focuses on the computation of the likelihood ratio when using regenerative IS method for simulations, and specifically when applying IS on the failure transitions of components in Markovian SPN models only, as we propose here. Finally, the update equations for the multi-level CE scheme are given and we propose an algorithm based on them.

4.3.1 SPN modeling for CTMC models

A Petri Net (PN) is an abstract and formal model of information flow, and is a powerful method for describing and analyzing the flows of information and controls in a system [95]. A PN is a directed graph whose nodes are partitioned into two disjoint sets, *places* (represented by circles) and *transitions* (represented by bars) [34]. *Arcs* are used for connecting places to transitions (input arcs) and transitions to places (output arcs) [34]. Places can contain *tokens* that are represented by black dots within places [95]. The state of a PN is given by the configuration of tokens in different places, where each specific configuration is called a *marking* [34]. A SPN is obtained from a PN model by assigning probability distribution function to the firing time of each transition [34]. In Markovian SPNs, there are only transitions with exponential distributions (timed transitions), whereas an extension of such SPNs is to have transitions with distributions having zero holding time (immediate transitions), called as Generalized SPNs (GSPN) [83]. To avoid any confusion, we use the term SPN for both here.

SPNs also have other important constructs such as marking dependencies, arc cardinality, guards (enabling conditions for transitions), etc., that make modeling of complex systems easier. For example, in SPNs, places can be used to represent an individual component (one token in the place) or multiple components of a sub-system (multiple tokens in a place). Similarly, transitions can be assigned firing rates (according to the exponential distributions) for a component failure or a system failure, depending on the model. Also, guards can be used to control the firings of transitions.

In SPNs, the *reachability graph* is the graph representing all the reachable markings, i.e., the various configurations of tokens representing the state of a system [34]. For Markovian SPNs, the reachability graph can be directly mapped to represent the underlying stochastic processes (here the CTMC). In GSPNs, as there are both timed and immediate transitions, the reachability graph is called as the *Extended Reachability Graph* (ERG), where the vanishing markings (markings in which the process spends zero time due to the firing of an immediate transition) are eliminated to obtain the underlying CTMC.

In literature, GSPNs have been shown to be the same as the underlying CTMCs and steady-state solutions have also been proved [83]. Simulations of SPNs also overcome the problem of largeness of the reachability graphs (they are not generated in such cases) and steady-state measures can be thus estimated using regenerative simulations [34]. These Markovian SPNs make it easier for a practitioner to model complex systems with relative ease. Due to this, we consider Markovian SPN models for simulations in the current work. Without going into further details of SPNs, interested readers can consult [34, 91, 92, 95, 96] for a deeper insight. In the current work, we use the Stochastic Petri Nets Package (SPNP) developed at the Duke University [34, 83], for analysis and modeling. An introduction of the tool is also given in the Appendix A.

4.3.2 Regenerative IS simulations in Markovian SPNs

The general formulation of the CTMC model in Section 4.2 considered a parametric density (under original measure $\mathbb{P}$) as $f(x;\theta)$ with a parameter vector θ containing the transitions rates among different states of the CTMC. Since we model the HRMS using Markovian SPNs here, let us consider that the vector θ contains the rate parameters of different transitions in a SPN. Each transition of the SPN can represent a single transition of a CTMC state change or a family of CTMC transitions too. The likelihood of a sample path ω in a cycle is still computed the same way as given in Equation 4.3 at the firing of each transition (i.e., state change in the CTMC).

Let us consider all the transitions are represented by a set F with their respective rates forming the parameter vector θ. We consider F to be a superset of transitions of interest (i.e., where the original distribution is to be replaced by a change of measure in IS later on) and any other transitions. Therefore,

$$\mathsf{F} = \mathscr{F} \cup \mathscr{R} \quad \text{and} \quad \{\theta = \theta_{\mathscr{F}} \cup \theta_{\mathscr{R}}\}. \tag{4.15}$$

Now, F is the set of all transitions of the model (with parameter vector θ), $\mathscr{F}$ is the subset of all transitions of interest (with subset parameter vector $\theta_{\mathscr{F}}$) and $\mathscr{R}$ is the subset of all other transitions not in $\mathscr{F}$ (with subset parameter vector $\theta_{\mathscr{R}}$).

Let there be a total finite number tr of transitions in the subset $\mathscr{F}$, such that:

$$\mathscr{F} = \{\underbrace{TR_1}_{\text{k=1}}, \underbrace{TR_2}_{\text{k=2}}, \underbrace{TR_3}_{\text{k=3}}, ..., \underbrace{TR_{tr}}_{\text{k=tr}}\} \quad \text{with vector of rates} \quad \theta_{\mathscr{F}} = \{\lambda_1, \lambda_2, \lambda_3, ..., \lambda_{tr}\}, \tag{4.16}$$

comprising of their respective rate parameters and each transition is indexed uniquely by $k = \{1, ..., tr\}$, as shown by Equation 4.16. Here k is considered as an identification parameter for *grouping* of the transitions in subset $\mathscr{F}$ (consequently in $\theta_{\mathscr{F}}$ also), that is explained later on.

From any given state x_i, let $\lambda(i)$ be the total rate out of that state. The total rate out of the state x_i is then $\lambda(i) = \sum_{k=1}^{tr} \lambda_k(i) + \sum R(i)$, where $\sum R(i)$ is the sum of the rates of transitions in subset $\mathscr{R}$ that are possible out of x_i and $\sum_{k=1}^{tr} \lambda_k(i)$ is the sum of rates of transitions in subset $\mathscr{F}$ that are possible out of x_i.

Now, we define two important parameters that are later on used in the current work:

- Parameter a_k: Let a_k be the number of times a k-th transition from subset $\mathscr{F}$ *occurs* (i.e., fires in the SPN model) in a cycle. Here, a_k is a random integer counter ≥ 0 for each transition that is updated each time a transition fires over the entire cycle.

- Parameter $b_{x_i,k}$: Let $b_{x_i,k}$ be the number of transitions from group k in subset $\mathscr{F}$ that are *possible* (i.e., enabled in the SPN model) at a given state x_i. Here, $b_{x_i,k}$ is also a random integer counter that can be only 1 or 0 (for this case) at a given state x_i for k-th transition (recall unique identification parameter k). A specific transition can be only enabled (1) or disabled (0) in this case. This parameter is updated at each state change (i.e., firing of a transition in the SPN)

From the above explanation, we can rewrite the Equation 4.3 of the likelihood of a sample path as:

$$f(x;\theta) = \prod_{k=1}^{tr} (\lambda_k)^{a_k} \cdot \Gamma \cdot \exp\left[-\sum_{i=0}^{\tau-1} \left(b_{x_i,1} \cdot \lambda_1 + ... + b_{x_i,tr} \cdot \lambda_{tr} + \sum R_{x_i} \right) v_{x_i} \right], \tag{4.17}$$

where Γ is the product of rates of transitions belonging to the subset $\mathscr{R}$ that have occurred (fired) in a sample path and ΣR_{x_i} is the sum of the rates for transitions from subset $\mathscr{R}$ enabled at a given state x_i. Separating the terms, the likelihood (or density) of a sample path is now:

$$f(x;\theta) = \prod_{k=1}^{tr} (\lambda_k)^{a_k} \cdot \Gamma \cdot \exp\left[- \sum_{i=0}^{\tau-1} (b_{x_i,1} \cdot \lambda_1 + ... + b_{x_i,tr} \cdot \lambda_{tr}) v_{x_i} \right] \cdot \exp\left[- \sum_{i=0}^{\tau-1} (\Sigma R_{x_i}) v_{x_i} \right]. \tag{4.18}$$

As explained in Section 4.2.3, we want to apply IS within the same parametric family and thus, by changing the parameter vector from θ (under $\mathbb{P}$) to $\tilde{\theta}$ (under $\tilde{\mathbb{P}}$). This means we have the original density $f(x;\theta)$ and the importance density $f(x;\tilde{\theta})$. In the current context, we discussed applying IS only on a subset of transitions of the model (i.e., $\mathscr{F} \in \digamma$) by changing the parameter vector $\theta_{\mathscr{F}}$ to $\tilde{\theta}_{\tilde{\mathscr{F}}}$. This means that the rates of the transitions in the original subset $\mathscr{F}$ (e.g., $\theta_{\mathscr{F}} = \{\lambda_1, \lambda_2, ...\lambda_{tr}\}$) are changed to form a new IS subset $\tilde{\mathscr{F}}$ with new IS rates vector (e.g., $\tilde{\theta}_{\tilde{\mathscr{F}}} = \{\tilde{\lambda}_1, \tilde{\lambda}_2, ...\tilde{\lambda}_{tr}\}$).

However, a question arises regarding the reason behind the application of IS only on the transitions of the subset $\mathscr{F} \in \digamma$, and not on the entire set $\digamma$. There are two logical reasons involved that we consider important in this context, and are discussed below.

1. The issue of *likelihood ratio degeneracy* could be avoided by this application. In [37, 97], it has been discussed to not use IS for high dimensional problems due to the likelihood degeneracy issue. This is due to the reason that in high-dimensional simulation models, the CE optimization schemes becomes useless, as the likelihood ratio term becomes the product of a large number of marginal likelihoods and the IS estimator based on the likelihood ratio would degenerate (i.e., having a large variance) [97]. This would mar the entire objective of using IS to obtain variance reduction. Particularly, the CE method (and also the VM approach) is susceptible to likelihood ratio degeneracy issue [31, 97]. The *screening method* as proposed in [97], focuses on reducing the dimension of the likelihood ratio by application of IS only on a subset of $\digamma$ (i.e., the *bottleneck parameters*) to form the subset $\mathscr{F}$.

2. We assume to apply IS only to the failure transitions of a Markovian SPN, selecting the subset $\mathscr{F}$ out of $\digamma$ based on failure transitions (or not failure transition for subset $\mathscr{R}$). The notion behind it is that increasing the failures of individual components would also increase the probability of a system failure (i.e., the target event to estimate Z using IS) in a cycle. For moderately sized problems, we can choose $\mathscr{F} = \digamma, \mathscr{R} = \emptyset$ to contain all the transitions of the model.

From the aforementioned discussions, using parameters a_k and $b_{x_i,k}$ we apply IS within the same parametric family on a subset $\mathscr{F} \in \digamma$ of transitions of the Markovian SPN. Note, the subsets in the original density and the IS density are equivalent, except the values. The likelihood ratio being the ratio of the two densities, $L(\omega) = f(x;\theta)/f(x;\tilde{\theta}) = f(x;\theta_{\mathscr{F}})/f(x;\tilde{\theta}_{\tilde{\mathscr{F}}})$

here, the general equation (see Equation 4.4) can be re-written as:

$$L(\omega) = \frac{f(x;\theta_{\mathscr{F}})}{f(x;\tilde{\theta}_{\tilde{\mathscr{F}}})} = \prod_{k=1}^{tr}\left(\frac{\lambda_k}{\tilde{\lambda}_k}\right)^{a_k} \cdot \exp\left[\sum_{i=0}^{\tau-1}\Big(b_{x_i,1}(\tilde{\lambda}_1-\lambda_1)+...+b_{x_i,tr}(\tilde{\lambda}_{tr}-\lambda_{tr})\Big)v_{x_i}\right]. \tag{4.19}$$

In the above equation, the ratio of the terms from subset $\mathscr{R}$ would be by default one, as $\theta_{\mathscr{R}}$ remains unchanged in $f(x;\theta)$ and $f(x;\tilde{\theta})$.

Grouping/ One-dimensional change of measure

Let us consider a specific case where a particular group of components in the Markovian SPN model have similar behavior in a sample path towards the target set (i.e., rare system failure). In such a case, the transition rates of that particular group (in subset $\mathscr{F}$) could be replaced by common IS rates in subset $\tilde{\mathscr{F}}$ for that particular group. This approach of grouping is specifically dependent on the model under consideration and the knowledge of the model for the practitioner. However, such a grouping of transitions is of particular interest as in the case of large models having many transitions in the subset $\mathscr{F}$, when IS is applied individually on each transition by changing their rates, it is possible that some transitions are not sampled even for a large number of cycles simulated. This would be undesirable when used for CE optimization, as such non-sampled transitions would provide a zero value. Also, in practice, it could be simpler to find a common value for IS rates of the transitions grouped together. We consider two cases here: first, when we apply IS on the transitions of the subset $\mathscr{F}$ by dividing them in groups within the subset $\tilde{\mathscr{F}}$; second when we apply IS on the subset $\mathscr{F}$ by using a single common IS value for all the transitions in $\tilde{\mathscr{F}}$.

Let us suppose that the transitions in subsets $\tilde{\mathscr{F}}$ and $\mathscr{F}$, are grouped in g number of groups (having tr_k number of transitions ($l = 1,...,tr_k$) in group $k = 1,...,g$). Here, we consider that each group has a unique k and for each group we have a common IS rate $\tilde{\lambda}_k$ for all the transitions within that group. This is explained by the following equation:

$$\begin{aligned}
\text{Original subset:}\quad & \mathscr{F} = \{\underbrace{TR_1,TR_2}_{\text{k=1}},\underbrace{TR_3,TR_4,TR_5}_{\text{k=2}},...,\underbrace{TR_{tr}}_{\text{k=g}}\}\\
\text{with}\quad & \theta_{\mathscr{F}} = \{\lambda_{1,1},\lambda_{2,1},\lambda_{3,2},\lambda_{4,2},\lambda_{5,2},...,\lambda_{tr,g}\}\ \text{and}\\
\text{IS subset:}\quad & \tilde{\mathscr{F}} = \{\underbrace{TR_1,TR_2}_{\text{k=1}},\underbrace{TR_3,TR_4,TR_5}_{\text{k=2}},...,\underbrace{TR_{tr}}_{\text{k=g}}\}\\
\text{with}\quad & \tilde{\theta}_{\tilde{\mathscr{F}}} = \{\underbrace{\tilde{\lambda}_1,\tilde{\lambda}_1}_{\text{k=1}},\underbrace{\tilde{\lambda}_2,\tilde{\lambda}_2,\tilde{\lambda}_2}_{\text{k=2}},...,\underbrace{\tilde{\lambda}_{tr,g}}_{\text{k=g}}\}.
\end{aligned} \tag{4.20}$$

Here, the original subset $\mathscr{F}$ is grouped and the same grouping is followed for the IS subset $\tilde{\mathscr{F}}$. For example, the transitions TR_3, TR_4, TR_5 are in group $k = 2$ above, having three transitions ($tr_k = 3$) and are replaced by a common value of IS rate $\tilde{\lambda}_2$ for all the transitions within the group.

In this case, the Equation 4.15 still holds true, however, the elements in subsets $\mathscr{F}$ and $\tilde{\mathscr{F}}$ are grouped. Each transition of a particular group has a common IS rate value $\tilde{\lambda}_k$. Now, the likelihood ratio given in Equation 4.19 can be written as:

$$L(\omega)=\frac{f(x;\theta_{\mathscr{F}})}{f(x;\tilde{\theta}_{\tilde{\mathscr{F}}})}=\left[\prod_{k=1}^{g}\prod_{l=1}^{tr_k}\left(\frac{\lambda_{l,k}}{\tilde{\lambda}_k}\right)^{a_{l,k}}\right]\cdot\exp\left[\sum_{i=0}^{\tau-1}\left(\sum_{k=1}^{g}\sum_{l=1}^{tr_k}b_{x_i,l,k}(\tilde{\lambda}_k-\lambda_{l,k})\right)v_{x_i}\right], \tag{4.21}$$

where $\lambda_{l,k}$ are the original rates of the transitions in subset $\mathscr{F}$ and similarly in subset $\tilde{\mathscr{F}}$ as per their respective groups identified by $k=1,...,g$. The number of transitions within each group k are $l=1,...,tr_k$.

Another possibility is a one-dimensional change of measure such that all the rates of transitions in subset $\mathscr{F}$ are replaced by a single common IS rate $\tilde{\lambda}$ for all transitions in subset $\tilde{\mathscr{F}}$ as explained in Equation 4.22 below.

$$\begin{array}{rr}
\text{Original subset:} & \mathscr{F}=\{\underbrace{TR_1,TR_2,TR_3,TR_4,TR_5,...,TR_{tr}}_{\text{k=1}}\} \\
\text{with} & \theta_{\mathscr{F}}=\{\lambda_1,\lambda_2,\lambda_3,\lambda_4,...,\lambda_{tr}\}\text{ and} \\
\text{IS subset:} & \tilde{\mathscr{F}}=\{\underbrace{TR_1,TR_2,TR_3,TR_4,TR_5,TR_{tr}}_{\text{k=1}}\} \\
\text{with} & \tilde{\theta}_{\tilde{\mathscr{F}}}=\{\tilde{\lambda},\tilde{\lambda},\tilde{\lambda},\tilde{\lambda},\tilde{\lambda},...,\tilde{\lambda}\}.
\end{array} \tag{4.22}$$

In such a case, there is only one group and $k=g=1, tr_k=tr$ and likelihood ratio is written as:

$$\begin{aligned}L(\omega)=\frac{f(x;\theta_{\mathscr{F}})}{f(x;\tilde{\theta}_{\tilde{\mathscr{F}}})}&=\left[\prod_{l=1}^{tr}\left(\frac{\lambda_l}{\tilde{\lambda}}\right)^{a_l}\right]\cdot\exp\left[\sum_{i=0}^{\tau-1}\left(\sum_{l=1}^{tr}b_{x_i,l}\left(\tilde{\lambda}-\lambda_l\right)v_{x_i}\right)\right]\\
&=\prod_{l=1}^{tr}\left(\frac{\lambda_l}{\tilde{\lambda}}\right)^{a_l}\cdot\exp\left[\sum_{i=0}^{\tau-1}\left(b_{x_i,1}(\tilde{\lambda}-\lambda_1)+...+b_{x_i,tr}(\tilde{\lambda}-\lambda_{tr})\right)v_{x_i}\right],\end{aligned} \tag{4.23}$$

where l is the corresponding index in subset $\mathscr{F}$, and all the transitions in both the subsets $\mathscr{F}$ & $\tilde{\mathscr{F}}$ are considered to be in the same group (i.e., $k=1$). If there is no grouping such that rate of each transition in subset $\mathscr{F}$ is replaced by unique IS rates in $\tilde{\mathscr{F}}$, then we can consider that there are tr number of groups and $g=1, tr_k=tr, k=l$ in Equation 4.21. The likelihood ratio is then given as:

$$L(\omega)=\frac{f(x;\theta_{\mathscr{F}})}{f(x;\tilde{\theta}_{\tilde{\mathscr{F}}})}=\prod_{k=1}^{tr}\left(\frac{\lambda_k}{\tilde{\lambda}_k}\right)^{a_k}\cdot\exp\left[\sum_{i=0}^{\tau-1}\left(b_{x_i,1}(\tilde{\lambda}_1-\lambda_1)+...+b_{x_i,tr}(\tilde{\lambda}_{tr}-\lambda_{tr})\right)v_{x_i}\right]. \tag{4.24}$$

The above Equation 4.24 is the same as Equation 4.19. Equations 4.21 and 4.23 are specific cases of Equation 4.19 in which the transition rates for IS are grouped in specific number of groups or one whole group (one-dimensional change of measure). In all the cases, the expectation of Z is still computed by $\mathbb{E}_{\tilde{\theta}}\left[Z\, L_{\omega}\right]$. In the next section, we explain how CE optimization can be performed as per grouping of transitions (or no grouping).

4.3.3 CE Optimization of SPN Transitions: Update Equation

The problem presented in Section 4.2 was to estimate the cumulative steady-state unavailability U, where IS is required to solve the problem of estimating the numerator Z efficiently in the Equation 4.2. The expectation of Z under IS change of measure is given by Equation 4.5. The IS density closest to the zero-variance importance density $g^{*}(x)$ (see Equation 4.6) regarding CE distance is possible to be approximated by the CE optimization technique. The update equation for the multi-level CE optimization scheme can be obtained from Equation 4.13.

In the context of the current work, we propose update equations for the multi-level CE scheme in two cases: first, when all the transitions of interest are to be optimized separately; second, transitions are grouped (either in multiple groups or a single one). In the following sections, we propose the update equations for these cases and also a general equation that we use in the multi-level CE scheme later on.

Multi-dimensional optimization

In the case of multi-dimensional optimization, the intent is to find the best rate parameters for IS application for each specific transition of interest individually, and form the CE optimized IS parameter vector $\tilde{\theta}^{*}_{ce}$ with those values (i.e., solution of Equation 4.12). This is the general case considered here for multi-dimensional optimization, where transitions are not grouped within subset $\tilde{\mathscr{F}}$ (or contrarily each single transition is a group made of itself having a single element). Let us consider Equation 4.13 again here:

$$\frac{1}{n}\sum_{m=1}^{n} Z_m(\omega_m)\, L(\omega_m;\theta;\check{\theta})\, \frac{\partial log\, f(\omega_m;\tilde{\theta})}{\partial\tilde{\theta}} = 0. \tag{4.25}$$

We can solve the system of equations given by Equation 4.25 above for each transition within the subset $\tilde{\mathscr{F}}$. This can be done by solving Equation 4.25 for each transition (separately in this case of multi-dimensional optimization) with respect to the IS change of measure $f(\omega_m;\tilde{\theta})$ (with vector $\tilde{\theta}$). Recall that $\{\tilde{\theta} = \tilde{\theta}_{\tilde{\mathscr{F}}} \cup \theta_{\mathscr{R}}\}$. The vector $\tilde{\theta}_{\tilde{\mathscr{F}}}$ comprises the IS rates of the transitions of interest (in the subset $\tilde{\mathscr{F}}$) and $\theta_{\mathscr{R}}$ comprises of unchanged rates (in the subset $\mathscr{R}$).

For the above purpose, we can take the partial derivative of $log\, f(\omega_m;\tilde{\theta})$ with respect to a single element (i.e., the rate of a specific single transition as shown by Equation 4.16) from the subset vector $\tilde{\theta}_{\tilde{\mathscr{F}}}$. By plugging in the partial derivative in Equation 4.25 and equating it to zero for that particular transition, we can obtain the CE optimized IS rate for that transition. In order to form the optimized IS vector $\tilde{\theta}^{*}_{ce}$ (the solution of the problem given in Equation 4.12), this computation is done for each transition in the subset $\tilde{\mathscr{F}}$.

Let us consider the density of a sample path given by Equation 4.18, where there is no grouping of transitions. Similarly, the density of a sample path under a change of measure $f(\cdot;\tilde{\theta})$, with parameter vector $\tilde{\theta}$, such that (s.t.) $\{\tilde{\theta} = \tilde{\theta}_{\tilde{\mathscr{F}}} \cup \theta_{\mathscr{R}}\}$ is given by:

$$f(\cdot;\tilde{\theta}) = \prod_{k=1}^{tr} \left(\tilde{\lambda}_k\right)^{a_k} \cdot \Gamma \cdot \exp\left[-\sum_{i=0}^{\tau-1} (b_{x_i,1}\cdot\tilde{\lambda}_1 + ... + b_{x_i,tr}\cdot\tilde{\lambda}_{tr}) v_{x_i}\right] \cdot \exp\left[-\sum_{i=0}^{\tau-1} (\Sigma R_{x_i}) v_{x_i}\right]. \tag{4.26}$$

Taking log of $f(\cdot;\tilde{\theta})$ and simplifying:

$$= log\left[\prod_{k=1}^{tr} \left(\tilde{\lambda}_k\right)^{a_k}\right] + log\,\Gamma + log\left[\exp\left[-\sum_{i=0}^{\tau-1} (b_{x_i,1}\cdot\tilde{\lambda}_1 + ... + b_{x_i,tr}\cdot\tilde{\lambda}_{tr}) v_{x_i}\right]\right]$$
$$+ log\left[\exp\left[-\sum_{i=0}^{\tau-1} (\Sigma R_{x_i}) v_{x_i}\right]\right]$$

$$= \left[log\left(\tilde{\lambda}_1\right)^{a_1} + log\left(\tilde{\lambda}_2\right)^{a_2} + log\left(\tilde{\lambda}_3\right)^{a_3} + ... + log\left(\tilde{\lambda}_{tr}\right)^{a_{tr}}\right] + log\,\Gamma$$
$$+ \left[-\sum_{i=0}^{\tau-1} (b_{x_i,1}\cdot\tilde{\lambda}_1 + ... + b_{x_i,tr}\cdot\tilde{\lambda}_{tr}) v_{x_i}\right] + \left[-\sum_{i=0}^{\tau-1} (\Sigma R_{x_i}) v_{x_i}\right].$$

Now, partially differentiating the above equation with respect to one of the IS rate (e.g., $\tilde{\lambda}_1$ of let's say a transition TR_1 in subset $\tilde{\mathscr{F}}$, as shown in Equation 4.16):

$$\frac{\partial log(f(\cdot;\tilde{\theta}))}{\partial\tilde{\lambda}_1} = \frac{a_1}{\tilde{\lambda}_1} - \left[\sum_{i=0}^{\tau-1} b_{x_i,1}\cdot v_{x_i}\right]. \tag{4.27}$$

Similarly for any other transition of interest (e.g., a failure transition) indexed by k (as shown in Equation 4.16), the partial derivative would be :

$$\frac{\partial log(f(\cdot;\tilde{\theta}))}{\partial\tilde{\lambda}_k} = \frac{a_k}{\tilde{\lambda}_k} - \left[\sum_{i=0}^{\tau-1} b_{x_i,k}\cdot v_{x_i}\right]. \tag{4.28}$$

Substituting Equation 4.28 in Equation 4.25:

$$\frac{1}{n}\sum_{m=1}^{n} Z_m(\omega_m)\, L(\omega_m;\theta;\check{\theta}) \left(\frac{a_k^{(m)}}{\tilde{\lambda}_k} - \left[\sum_{i=0}^{\tau-1} b_{x_i,k}^{(m)}\, v_{x_i}\right]\right) = 0 \tag{4.29}$$

$$\tilde{\lambda}_k = \frac{\sum_{m=1}^{n} \left[Z_m(\omega_m)\, L(\omega_m;\theta;\check{\theta})\ a_k^{(m)} \right]}{\sum_{m=1}^{n} \left[Z_m(\omega_m)\, L(\omega_m;\theta;\check{\theta}) \left[\sum_{i=0}^{\tau-1} b_{x_i,k}^{(m)}\, v_{x_i} \right] \right]}. \tag{4.30}$$

The above equation solves the system of equations given by Equation 4.25 with respect to a single transition's IS rate, where there are no multi-levels of CE used. As CE optimization needs a multi-level scheme (see Section 4.2.4), let us consider j number of stages of optimization, where in each stage n_j number of cycles are simulated and use the above Equation 4.30 for a multi-level scheme. We use the IS density $\check{\theta}_{ce}^{(j)}$ at each stage for sampling, where it is also divided in subsets $\tilde{\mathscr{F}}$ and $\mathscr{R}$ as done for $\tilde{\theta}$. Plugging the above value of the solution of the maximization problem for a single transition's IS rate in Equation 4.14, where $\check{\lambda}_{ce,k}^{(j+1)} = \tilde{\lambda}_k^{(j)}$, we have the updating equation as:

$$\check{\lambda}_{ce,k}^{(j+1)} = \frac{\sum_{m=1}^{n_j} \left[Z_{j,m}(\omega_{j,m})\, L(\omega_{j,m};\theta;\check{\theta}_{ce}^{(j)})\ a_k^{(j,m)} \right]}{\sum_{m=1}^{n_j} \left[Z_{j,m}(\omega_{j,m})\, L(\omega_{j,m};\theta;\check{\theta}_{ce}^{(j)})\ \left[\sum_{i=0}^{\tau-1} b_{x_i,k}^{(j,m)} v_{x_i} \right] \right]}. \tag{4.31}$$

Here, the likelihood ratio is $L(\omega_{j,m};\theta;\check{\theta}_{ce}^{(j)})$ at a given stage j, with the original measure under θ and the IS change of measure under $\check{\theta}_{ce}^{(j)}$. It is the computed the same way as given by Equation 4.19. For optimizing multiple transitions having different k indexes, the new parameter vector $\check{\theta}_{ce}^{(j+1)}$ is constructed for each subsequent stage $j+1$ by using Equation 4.31 separately for each transition at the current stage j. Thus from this, we can construct an optimized CE parameter vector $\tilde{\theta}_{ce}^{*}$ in j number of stages where $\tilde{\theta}_{ce}^{*} = \check{\theta}_{ce}^{(j+1)}$ obtained after the final pre-simulation stage.

Grouping/ One-dimensional optimization

In this section, we explain the special case presented in Section 4.3.2 for grouped or one-dimensional change of measure. Recall that in this case, the transitions of interest (i.e., in subset $\mathscr{F}$ and consequently in subset $\tilde{\mathscr{F}}$) have their transition rates replaced by either a common value for a group of transitions (grouped change of measure) or a single value (one-dimensional change of measure) in the vector $\tilde{\theta}_{\tilde{\mathscr{F}}}$. Let us consider the two cases separately in the following text.

Case 1 (Grouped Optimization): Here we assume that the transitions are grouped in the subsets $\mathscr{F}$ and $\tilde{\mathscr{F}}$ as shown by Equation 4.20. There are g number of groups indexed by k, tr_k number of transitions within each group. The likelihood ratio $L(\omega)$ is given by Equation

4.21, as the ratio of the original density (with vector θ containing $\theta_{\mathscr{F}}$) and the IS density (with vector $\tilde{\theta}$ containing $\tilde{\theta}_{\mathscr{F}}$).

The density of the sample path under a change of measure ($f(\cdot;\tilde{\theta})$) is given by:

$$f(\cdot;\tilde{\theta}) = \left[\prod_{k=1}^{g} (\tilde{\lambda}_k)^{\Sigma_{l=1}^{tr_k} a_{l,k}}\right] \cdot \Gamma \cdot \exp\left[-\sum_{i=0}^{\tau-1}\left(\sum_{k=1}^{g}\left(\tilde{\lambda}_k \sum_{l=1}^{tr_k} b_{x_i,k,l}\right)\right) v_{x_i}\right] \cdot \exp\left[-\sum_{i=0}^{\tau-1}(\Sigma R_{x_i}) v_{x_i}\right], \tag{4.32}$$

where $f(\cdot;\tilde{\theta})$ is with parameter vector $\{\tilde{\theta} = \tilde{\theta}_{\mathscr{F}} \cup \theta_{\mathscr{R}}\}$, s.t. $\tilde{\theta}_{\mathscr{F}}$ contains IS rates in grouped form as shown in Equation 4.20.

Now taking the log of $f(\cdot;\tilde{\theta})$ again and simplifying:

$$\begin{aligned} &= log\left[\prod_{k=1}^{g} (\tilde{\lambda}_k)^{\Sigma_{l=1}^{tr_k} a_{l,k}}\right] + \ log\,\Gamma + log\left[\exp\left[-\sum_{i=0}^{\tau-1}\left(\sum_{k=1}^{g}\left(\tilde{\lambda}_k \sum_{l=1}^{tr_k} b_{x_i,k,l}\right)\right) v_{x_i}\right]\right] \\ &+ log\left[\exp\left[-\sum_{i=0}^{\tau-1}(\Sigma R_{x_i}) v_{x_i}\right]\right] \\ &= \left[log(\tilde{\lambda}_1)^{\left(\Sigma_{l=1}^{tr_1} a_{l,1}\right)} + log(\tilde{\lambda}_2)^{\left(\Sigma_{l=1}^{tr_2} a_{l,2}\right)} + \ldots + log(\tilde{\lambda}_g)^{\left(\Sigma_{l=1}^{tr_g} a_{l,g}\right)}\right] + \ log\,\Gamma \\ &+ \left[-\sum_{i=0}^{\tau-1}\left(\sum_{k=1}^{g}\left(\tilde{\lambda}_k \sum_{l=1}^{tr_k} b_{x_i,k,l}\right)\right) v_{x_i}\right] + \left[-\sum_{i=0}^{\tau-1}(\Sigma R_{x_i}) v_{x_i}\right]. \end{aligned}$$

In this case, the partial derivative is taken in the above equation, with respect to a group k's common element $\tilde{\lambda}_k$. Let's suppose we take the partial derivative with respect to a group $k = 1$ (i.e., with respect to the element $\tilde{\lambda}_1$) in above the equation and we get:

$$\frac{\partial log(f(\cdot;\tilde{\theta}))}{\partial \tilde{\lambda}_1} = \frac{\Sigma_{l=1}^{tr_1} a_{l,1}}{\tilde{\lambda}_1} - \left[\sum_{i=0}^{\tau-1}\left(\sum_{l=1}^{tr_1} b_{x_i,1,l}\right) v_{x_i}\right]. \tag{4.33}$$

Similarly for any other group indexed by k, the partial derivative is:

$$\frac{\partial log(f(\cdot;\tilde{\theta}))}{\partial \tilde{\lambda}_k} = \frac{\Sigma_{l=1}^{tr_k} a_{l,k}}{\tilde{\lambda}_k} - \left[\sum_{i=0}^{\tau-1}\left(\sum_{l=1}^{tr_k} b_{x_i,k,l}\right) v_{x_i}\right]. \tag{4.34}$$

Again, substituting Equation 4.34 in Equation 4.25:

$$\begin{aligned} &\frac{1}{n}\sum_{m=1}^{n} Z_m(\omega_m)\, L(\omega_m;\theta;\check{\theta}) \left(\frac{\Sigma_{l=1}^{tr_k} a_{l,k}^{(m)}}{\tilde{\lambda}_k} - \left[\sum_{i=0}^{\tau-1}\left(\sum_{l=1}^{tr_k} b_{x_i,k,l}^{(m)}\right) v_{x_i}\right]\right) = 0 \\ &\tilde{\lambda}_k = \frac{\Sigma_{m=1}^{n} Z_m(\omega_m)\, L(\omega_m;\theta;\check{\theta}) \ \left(\Sigma_{l=1}^{tr_k} a_{l,k}^{(m)}\right)}{\Sigma_{m=1}^{n} Z_m(\omega_m)\, L(\omega_m;\theta;\check{\theta}) \left[\Sigma_{i=0}^{\tau-1}\left(\Sigma_{l=1}^{tr_k} b_{x_i,k,l}^{(m)}\right) v_{x_i}\right]}. \end{aligned} \tag{4.35}$$

The above equation solves the system of equations formed by the Equation 4.25 with respect to a group's multiple elements indexed by k (see grouping as shown in Equation 4.20), having a common value of IS rate $\tilde{\lambda}_k$. As previously done for multi-dimensional optimization, let us consider for a multi-level CE scheme j number of stages of optimization where in each stage, n_j number of cycles are simulated and the above equation is used for each group at each stage. We use the IS density given by parameter vector $\breve{\theta}_{ce}^{(j)}$ at stage j for sampling, where it is also divided in subsets $\tilde{\mathscr{F}}$ and $\mathscr{R}$. Plugging the above value of Equation 4.35 in Equation 4.14, for $\breve{\lambda}_{ce,k}^{(j+1)} = \tilde{\lambda}_k^{(j)}$ (i.e., common IS rate for a single group of transitions), we have:

$$\breve{\lambda}_{ce,k}^{(j+1)} = \frac{\displaystyle\sum_{m=1}^{n_j} \left[Z_{j,m}(\omega_{j,m})\, L(\omega_{j,m};\theta;\breve{\theta}_{ce}^{(j)}) \;\; (\textstyle\sum_{l=1}^{tr_k} a_{l,k}^{(j,m)}) \right]}{\displaystyle\sum_{m=1}^{n_j} \left[Z_{j,m}(\omega_{j,m})\, L(\omega_{j,m};\theta;\breve{\theta}_{ce}^{(j)}) \left[\textstyle\sum_{i=0}^{\tau-1} \left(\sum_{l=1}^{tr_k} b_{x_i,k,l}^{(j,m)} \right) v_{x_i} \right] \right]}. \tag{4.36}$$

Here, $\breve{\lambda}_{ce,k}^{(j+1)}$ for the next stage (or $\breve{\lambda}_{ce,k}^{(j)}$ at the current stage), is a common value of the rate for all transitions in group k in subset $\tilde{\mathscr{F}}$. The $\breve{\theta}_{ce}^{(j+1)}$ is now constructed by using the above Equation 4.36 for each group k having multiple transitions within it. The parameters a in the sum $(\sum_{l=1}^{tr_k} a_{l,k}^{(j,m)})$ is computed if any of the transition in a group k (having tr_k number of transitions in that group) has fired in sample path m at stage j. Similarly, the parameter b in the sum $\sum_{l=1}^{tr_k} b_{x_i,k,l}^{(j,m)}$ is also computed if any transition of group k is enabled at state x_i in the sample path m simulated at stage j. The final stage j allows us to obtain the CE optimized IS rates $\breve{\lambda}_{ce,k}^{(j+1)}$ to form the vector $\tilde{\theta}_{ce}^{*} = \breve{\theta}_{ce}^{(j+1)}$.

Case 2 (One-Dimensional Optimization): In the last special case, the transitions in the subset $\mathscr{F}$ of $\mathcal{F}$ are to be replaced by a single value of IS rates as explained by Equation 4.22. There is only a single group $k = g = 1$ containing $tr_k = tr$ number of transitions. The likelihood ratio is as given by Equation 4.23 with the original density (with vector θ containing $\theta_{\mathscr{F}}$) and IS density (with vector $\tilde{\theta}$ containing $\tilde{\theta}_{\tilde{\mathscr{F}}}$). The density of the sample path $(f(\cdot;\tilde{\theta}))$ is now given by:

$$f(\cdot;\tilde{\theta}) = \tilde{\lambda}^{(\sum_{l=1}^{tr} a_l)} \cdot \Gamma \cdot \exp\left[-\sum_{i=0}^{\tau-1} \left(\sum_{l=1}^{tr} b_{x_i,l} \right) \tilde{\lambda} v_{x_i} \right] \cdot \exp\left[-\sum_{i=0}^{\tau-1} (\Sigma R_{x_i}) v_{x_i} \right], \tag{4.37}$$

where $f(\cdot;\tilde{\theta})$ is defined by parameter vector $\{\tilde{\theta} = \tilde{\theta}_{\tilde{\mathscr{F}}} \cup \theta_{\mathscr{R}}\}$, s.t. $\tilde{\theta}_{\tilde{\mathscr{F}}}$ contains IS rates with a single common value for all transitions in $\tilde{\mathscr{F}}$ (as shown in Equation 4.22). As done previously for the multi-dimensional and grouped optimization cases, we first take the log of

$f(\cdot;\tilde{\theta})$ and simplify, as shown below.

$$
\begin{aligned}
&= log\left[\tilde{\lambda}^{(\Sigma_{l=1}^{tr} a_l)}\right] + log\,\Gamma + log\left[\exp\left[-\sum_{i=0}^{\tau-1}\left(\sum_{l=1}^{tr} b_{x_i,l}\right)\tilde{\lambda} v_{x_i}\right]\right] \\
&+ log\left[\exp\left[-\sum_{i=0}^{\tau-1}(\Sigma R_{x_i})v_{x_i}\right]\right] \\
&= log\left[\tilde{\lambda}^{(\Sigma_{l=1}^{tr} a_l)}\right] + log\,\Gamma - \left[\sum_{i=0}^{\tau-1}\left(\sum_{l=1}^{tr} b_{x_i,l}\right)\tilde{\lambda} v_{x_i}\right] - \left[\sum_{i=0}^{\tau-1}(\Sigma R_{x_i})v_{x_i}\right].
\end{aligned}
$$

Since there is only one group that contains only a common value $\tilde{\lambda}$ in the subset $\tilde{\mathscr{F}}$, the partial derivative of the above equation is taken with respect to $\tilde{\lambda}$ for all the transitions in subset vector $\tilde{\theta}_{\tilde{\mathscr{F}}}$:

$$
\frac{\partial log(f(\cdot;\tilde{\theta}))}{\partial\tilde{\lambda}} = \frac{\Sigma_{l=1}^{tr} a_l}{\tilde{\lambda}} - \left[\sum_{i=0}^{\tau-1}\left(\sum_{l=1}^{tr} b_{x_i,l}\right)v_{x_i}\right]. \tag{4.38}
$$

Substituting the above in Equation 4.25:

$$
\begin{aligned}
&\frac{1}{n}\sum_{m=1}^{n} Z_m(\omega_m)\,L(\omega_m;\theta;\breve{\theta})\left(\frac{\Sigma_{l=1}^{tr} a_l^{(m)}}{\tilde{\lambda}} - \left[\sum_{i=0}^{\tau-1}\left(\sum_{l=1}^{tr} b_{x_i,l}^{(m)}\right)v_{x_i}\right]\right) = 0 \\
&\tilde{\lambda} = \frac{\Sigma_{m=1}^{n} Z_m(\omega_m)\,L(\omega_m;\theta;\breve{\theta})\left(\Sigma_{l=1}^{tr} a_l^{(m)}\right)}{\Sigma_{m=1}^{n} Z_m(\omega_m)\,L(\omega_m;\theta;\breve{\theta})\left[\Sigma_{i=0}^{\tau-1}\left(\Sigma_{l=1}^{tr} b_{x_i,l}^{(m)}\right)v_{x_i}\right]}.
\end{aligned}
\tag{4.39}
$$

The above Equation 4.39 would provide a single common value of IS rate $\tilde{\lambda}$ when all the transitions in $\mathscr{F}$ and consequently in $\tilde{\mathscr{F}}$ are grouped together in a single group and it solves the Equation 4.25. For a multi-level CE scheme, having j number of stages, n_j number of cycles simulated at stage j, we can use the above equation. Let us consider again we use IS density given by $\breve{\theta}_{ce}^{(j)}$ at stage j for sampling. Now for $\breve{\lambda}_{ce}^{(j+1)} = \tilde{\lambda}^{(j)}$ in Equation 4.39 and using Equation 4.14 we obtain:

$$
\breve{\lambda}_{ce}^{(j+1)} = \frac{\displaystyle\sum_{m=1}^{n_j}\left[Z_{j,m}(\omega_{j,m})\,L(\omega_{j,m};\theta;\breve{\theta}_{ce}^{(j)})\;\Sigma_{l=1}^{tr} a_l^{(j,m)}\right]}{\displaystyle\sum_{m=1}^{n_j}\left[Z_{j,m}(\omega_{j,m})\,L(\omega_{j,m};\theta;\breve{\theta}_{ce}^{(j)})\;\left[\Sigma_{i=0}^{\tau-1}\left(\Sigma_{l=1}^{tr} b_{x_i,l}\right)v_{x_i}\right]\right]}. \tag{4.40}
$$

Here $\breve{\lambda}_{ce}^{(j+1)}$ for the next stage (or $\breve{\lambda}_{ce}^{(j)}$ at the current stage), is a common value for all the transitions in subset $\tilde{\mathscr{F}}$. Here, the parameter vector $\breve{\theta}_{ce}^{(j+1)}$ for the next stage is constructed by using the above Equation 4.40 for the all the transitions in subset $\tilde{\mathscr{F}}$ grouped together in a single group.

Selection of CE update equation

In the previous sections, we proposed update equations for a multi-level CE scheme for three different cases. Equation 4.31 is for optimizing transitions of interest separately as shown by Equation 4.16. Equation 4.36 is for optimizing transitions in groups as shown by Equation 4.20. Finally, Equation 4.40 is for optimizing all transitions as a single group as shown by Equation 4.22. The main difference between these update equations is in the computation part. For the purpose of simplicity, Equation 4.36 can be considered as our main update equation for a multi-level CE scheme. The case given by Equation 4.16 can be assumed in Equation 4.20 as each transition being a group itself (i.e., $k = 1,..,tr$). The one-dimensional change of measure given by Equation 4.22 can also be assumed in Equation 4.20 as the case when there is a single group $k = 1$ having tr number of transitions within it. The likelihood ratio is computed at each state change (i.e., firing of a transition in the Markovian SPN) and Equation 4.21 is used, where the transitions are either uniquely identified or in groups or as a single group using index k in the subsets $\mathscr{F}$ and $\tilde{\mathscr{F}}$.

4.3.4 Application of CE Optimization Scheme for Markovian SPNs: Algorithm

In the current work, Markovian SPNs are used to conveniently represent the HRMS models and previously it was discussed how Markovian SPNs can be used to estimate U via IS in the current context. Also, we showed how a multi-level CE update equation can be used for optimizing the IS rates of transitions in a SPN. The CE scheme is supposed to be used in a pre-simulation to obtain CE optimized IS rates for the transitions of interest (i.e., the transitions in the subset $\mathscr{F}$ of $\digamma$). We now propose a multi-level CE algorithm where the problem defined by $f(x;\theta)$, where $\{\theta = \theta_{\mathscr{F}} \cup \theta_{\mathscr{R}}\}$, is broken down into a series of less rare problems.

Description of the Algorithm (3)

In the Algorithm 3, it is considered that the problem is defined by the original density $f(x;\theta)$. Certain assumptions of the algorithm are:

1. Subsets: It is assumed that original vector θ (with set $\digamma$ containing all transitions) contains rates of transitions of interest $\theta_{\mathscr{F}}$ (subset $\mathscr{F} \in \digamma$) and rates of transitions of non-interest $\theta_{\mathscr{R}}$ (subset $\mathscr{R} \in \digamma$). Same is true for the IS density and the vectors forming it (divided in subsets $\tilde{\mathscr{F}}$ and $\mathscr{R}$) similarly.

2. Grouping: The grouping is defined for subset $\tilde{\mathscr{F}}$ (and $\mathscr{F}$) by giving unique value of k index for each group. Transitions within a group have the same k values and consequently optimized together to obtain common value of IS rates for that group.

Algorithm 3 Cross Entropy Algorithm for Markovian SPNs

```
Inputs: Original problem $f(x;\theta)$, transitions of interest or not ($\mathscr{F}$ and $\mathscr{R}$ subsets), grouping strategy by $k$ in subsets $\mathscr{F}$ and $\tilde{\mathscr{F}}$, no. of pre-simulation stages $j = 1,...,S$ and number of cycles $(n_j)$ at stage $j$
Output: Vector of CE optimized IS rates for the problem $f(x;\theta)$
Procedure:
Redefine problem: Create an unstable system by increasing rates in subset $\mathscr{F}$ to form a new easy sub-problem $f(x;\theta^{(j=1)})$ s.t. $\{\theta^{(j=1)} = \theta_{\mathscr{F}}^{(j=1)} \cup \theta_{\mathscr{R}}\}$
Initial IS vector: Same as the new target problem $\check{\theta}_{ce}^{(j=1)} = \theta^{(j=1)}$.    ▷ standard regenerative simulation
for each pre-simulation stage from $j = 1$ to $S$
    simulate $n_j$ cycles using $\theta^{(j)}$ as new sub-problem and $\check{\theta}_{ce}^{(j)}$ as IS change of measure.
    for each cycle at stage $j$
        Initialize: sum of parameter $a_{l,k}^{(j,m)}$ to zero for all groups.
        for each state change in a cycle: $i = 0$ to $\tau - 1$
            Initialize: sum of parameter $b_{x_i,k}^{(j,m)}$ to zero for all groups at each state change.
            At each firing: Compute the sum for $a_{l,k}^{(j,m)}$ for each group $k$
            At each state: Compute the sum of $b_{x_i,k}^{(j,m)}$ and multiply with sojourn time $v_{x_i}$ in that state for each group $k$
            At each state change: Compute downtime $Z_j(\cdot)$ as a sum and the likelihood ratio $L(\cdot;\theta^{(j)};\check{\theta}_{ce}^{(j)})$
        end for

        Compute the sum of numerator and denominator for each group $k$ using Equation 4.36
    end for

    For each group $k$ of transitions: Compute next stage common IS rate $\check{\lambda}_{ce,k}^{(j+1)}$ to form $\{\check{\theta}_{ce}^{(j+1)} = \check{\theta}_{\tilde{\mathscr{F}}}^{(j+1)} \cup \theta_{\mathscr{R}}\}$ via Equation 4.36
    Progressive rarity shifting: Decrease the rates in $\theta_{\mathscr{F}}^{(j)}$ to form $\theta_{\mathscr{F}}^{(j+1)}$
    if $\theta_{\mathscr{F}}^{(j+1)} \leq \theta_{\mathscr{F}}$ then    ▷ If failures rates go below the values in original problem $\theta_{\mathscr{F}}$
            $\theta_{\mathscr{F}}^{(j+1)} \leftarrow \theta_{\mathscr{F}}$
    end if
    New rarer problem for $j+1$ stage: $f(x;\theta^{(j+1)})$ s.t. $\{\theta^{(j+1)} = \theta_{\mathscr{F}}^{(j+1)} \cup \theta_{\mathscr{R}}\}$
    $j \leftarrow j+1$
end for
return $\tilde{\theta}_{ce}^{*} = \check{\theta}_{ce}^{(j+1)}$ obtained after final pre-simulation stage $j = S$.
Use $\tilde{\theta}_{ce}^{*}$ as IS change of measure to estimate $Z$ for original problem given by $f(x;\theta)$.
```

3. Arithmetic operations: All arithmetic operations denoted for a set (or a vector) denotes the same arithmetic operation on each element (i.e., the rates in this case) of that set (or vector).

4. Other inputs: The number of pre-simulation stages j are pre-defined and the number of regenerative IS cycles n_j to be simulated at each j are also given.

With the above assumptions, the algorithm firstly redefines the original problem given by θ, s.t. $\{\theta = \theta_{\mathscr{F}} \cup \theta_{\mathscr{R}}\}$, into a new sub-problem $\{\theta^{(j=1)} = \theta_{\mathscr{F}}^{(j=1)} \cup \theta_{\mathscr{R}}\}$ for the first stage $j = 1$. Since IS is to be applied only on failure transitions in our case, $\theta_{\mathscr{F}}^{(j=1)}$ can be formed by increasing the failure rates given in the original vector $\theta_{\mathscr{F}}$. The IS vector $\breve{\theta}_{ce}^{(j=1)}$ is considered to be the same as the new sub-problem to be solved at this stage, s.t. $\{\breve{\theta}_{ce}^{(j=1)} = \breve{\theta}_{\tilde{\mathscr{F}}}^{(j=1)} \cup \theta_{\mathscr{R}}\}$ and $\breve{\theta}_{\tilde{\mathscr{F}}}^{(j=1)} = \theta_{\mathscr{F}}^{(j=1)}$. Assuming this makes the system unstable, a standard regenerative simulation is performed (likelihood ratio would be one in such case) and Equation 4.36 is used to find the IS rates (for each group in $\tilde{\mathscr{F}}$) of transitions for the next stage $j+1$.

In the subsequent stages, we create a a newer and rarer sub-problem $\{\theta^{(j+1)} = \theta_{\mathscr{F}}^{(j+1)} \cup \theta_{\mathscr{R}}\}$. This is done by decreasing the failure rates of components considered in the current stage given by $\{\theta^{(j)} = \theta_{\mathscr{F}}^{(j)} \cup \theta_{\mathscr{R}}\}$ (i.e., the rates in vector $\theta_{\mathscr{F}}^{(j)}$). The algorithm uses the IS vector $\{\breve{\theta}_{ce}^{(j+1)} = \breve{\theta}_{\tilde{\mathscr{F}}}^{(j+1)} \cup \theta_{\mathscr{R}}\}$ obtained from the current stage j (for the next stage $j+1$) as a solution of Equation 4.36 for each group of transitions. In the final stage of pre-simulation, the original problem $\{\theta = \theta_{\mathscr{F}} \cup \theta_{\mathscr{R}}\}$ is solved using the IS rates obtained from the previous stage. The solution obtained from the final stage is the CE optimized IS vector of rates $\tilde{\theta}_{ce}^{*}$ that can be used in main simulations as a IS change of measure for the original problem $f(x;\theta)$.

Remark: It is to be noted that at Step 12 of Algorithm 3, the parameter $a_{l,k}$ is computed over a cycle as a sum of number of times any of the transitions within the group k have fired in the Markovian SPN. Similarly, at Step 13, the parameter $b_{x_i,k,l}$ is computed at each state change (in the cycle) as a sum of all the transitions of a group k that are enabled at a current state x_i.

From the above discussion, some questions arise regarding how to shift the original problem to a series of less rare sub-problems, the selection of number of pre-simulation stages $(j = 1,..,S)$ and the number of cycles to be simulated at each stage n_j. For this purpose, we used certain heuristic rules that are described as follows.

- **Progressive shifting of problem**: We co-relate the original problem θ with the number of stages S used to solve it. Starting from $\{\theta = \theta_{\mathscr{F}} \cup \theta_{\mathscr{R}}\}$, for the first stage $j = 1$, we take the $1/S$th root of the failure rates of all elements in the vector $\theta_{\mathscr{F}}$ to form a new vector $\theta_{\mathscr{F}}^{(j=1)}$ containing same elements but with higher failure rates. This way we form the new sub-problem $\{\theta^{(j=1)} = \theta_{\mathscr{F}}^{(j=1)} \cup \theta_{\mathscr{R}}\}$. At each subsequent stage, we raise each element of $\theta_{\mathscr{F}}^{(j=1)}$ to the power equal to the number of current stage (i.e., $j = 2,3,4,...,S$) to form $\theta_{\mathscr{F}}^{(j+1)}$ and consequently $\{\theta^{(j+1)} = \theta_{\mathscr{F}}^{(j+1)} \cup \theta_{\mathscr{R}}\}$. With this heuristic, the failure rates are slowly decreased at each new stage forming a new increasingly rarer sub-problem to be solved using the IS rates obtained from the previous stage. We increase the choice of the total number of stages S depending on the rarity of the failure rates in original problem θ for each new main simulation. Also, for moderately sized problems, repair transitions can also be considered as transitions of interest and the same heuristic can be used, except for the condition in Step 22 of the Algorithm 3 repair transitions should be exempted.

The simplicity of this heuristic approach helps in creating an unstable system at first by increasing failure rates (and decreasing repair rates, if chosen). Then we gradually create a more stable system with increasingly rare failures by decreasing failure rates (and increasing repair rates, if chosen) from the first stage onward until the original problem is reached.

- **Number of cycles n_j at each stage j:** The number of cycles at each stage j should be enough to be able to solve the problem at that stage. This can be easily chosen according to the IS rates used at a given stage j for the failure transitions such that it is sufficient to sample the firing of the transitions of interest within the chosen value of n_j number of cycles.

In the next sections, we present the numerical results of using the Algorithm 3 to find CE optimized IS rates of Markovian SPNs and using those IS rates in main simulations to estimate the steady-state unavailability U. The main focus is on the gain obtained compared to a standard regenerative MC simulation and also the RE property as the rarity of the original problem is increased. Also, in all the examples where the proposed CE algorithm is applied, each problem is solved by breaking it down into smaller sub-problems as per the heuristic rule defined above for progressive shifting of rarity in each problem.

4.4 Example 1: A 3 State Birth-and-Death Process

Let us consider a three state simple birth-and-death process where the sample space $\{\mathscr{K} = \{0,1,2\} = \mathscr{U} \cup \mathscr{D}\}$ s.t. $\{\mathscr{D} = \{2\}\}$ is the state when the system is failed. Figure 4.1 represents a SPN model of such a small system, even though a CTMC model can be directly evaluated for such a small example. However, our goal is to show the usefulness of the Algorithm 3 to optimize Markovian SPN's transitions. In this example, the underlying CTMC model is also exactly the same as in Figure 4.1.

4.4.1 Model Description

The model in Figure 4.1 can be considered equivalent to a repairable system with 2-out-of-3 (2oo3) redundancy, where if 2 or more components are failed, the system is failed. The initial state of the system is $\{0\}$ (token in place $\{0\}$ in Figure 4.1) where no components are failed. The transition *fail1* represents failure of one component and *fail2* transition represents

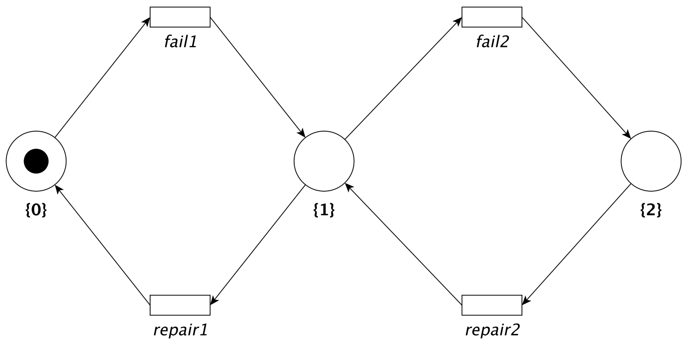

Figure 4.1 Birth & Death Process: SPN model of a 3 state simple birth-and-death process.

failure of second component. Repair actions are represented by firing of transitions *repair1* or *repair2*. We consider the original parameter vector of rates out of each state (denoting an event of failure or repair of a component) for the four possible transitions *fail1*, *fail2*, *repair2* and *repair1* as $\{\theta = (\lambda_1, \lambda_2, \mu_2, \mu_1)\}$ respectively. Firing of transitions results in movements of tokens between places $\{0\}, \{1\}$ and $\{2\}$, where each place represents a state of the underlying CTMC in this model. The objective is to estimate the steady-state unavailability U, i.e., the long run fraction of time spent in place $\{2\}$ by the above system. The probability of the respective transitions between different states and the holding times in states (places in the above model), are shown in the Table 4.1.

The value of the rates (per hour) with exponential distribution of holding times are: $\lambda_1 = \lambda_2$ and $\mu_2 = \mu_1 = 2.0$. For this model, the exact numerical value of the steady-state unavailability is also obtained numerically.

4.4.2 Empirical Results and Interpretations

Results from a standard regenerative simulation for estimation of U with $N = 10^6$ cycles simulated are shown in Table 4.2, when $\lambda_1 = \lambda_2 \to 0$. As expected from standard simulations, the RE increases rapidly with the rarity. For $\lambda_2 \leq 10^{-05}$, the point estimates $\hat{U}$ become inaccurate compared to U and the relative 95% CI width is also increased by a huge margin. When $\lambda_1 = \lambda_2 < 10^{-5}$, we start getting empirical values of 0.0 for the estimators and 95% CI bounds. This is the general problem in estimation of rare events using standard methods.

It is to be noted that we are able to obtain estimates $\hat{U}$ by the standard regenerative simulation for magnitudes $\hat{U} \leq 10^{-7}$ (for $\lambda_1 = \lambda_2 \leq 10^{-3}$) even with just $N = 10^6$ runs because $\hat{U}$ depends on μ_2 that remains a constant. For example, when $\lambda_1 = \lambda_2 = 10^{-5}$, the probability $p(1,2) = \lambda_2/(\lambda_2 + \mu_1) \approx 5.0 \times 10^{-6}$ (see Table 4.1) and hence probability of a token reaching place $\{2\}$ (equivalent to the CTMC in state 2) is possible in 10^6 cycles. The holding time in $h_{\{2\}} = 0.5$ remains a constant.

Progressive rarity shifting and choice of parameters of Algorithm 3: In this small example, IS is considered to be applied on the two transitions *fail2* and *repair1*. The probability of other transitions *fail1* and *repair2* is one (see Table 4.1) so IS is not needed for them. The subsets of the parameter vector θ are: $\{\theta_{\mathscr{F}} = (\lambda_2, \mu_1)\}$ and $\{\theta_{\mathscr{R}} = (\lambda_1, \mu_2)\}$.

Table 4.1 Birth & Death Process: Transition probabilities and holding times.

Probability	$p(0,1) = p(2,1) = 1$	$p(1,2) = \lambda_2/(\lambda_2+\mu_1)$	$p(1,0) = \mu_1/(\lambda_2+\mu_1)$
Holding Time	$h_{\{0\}} = 1/\lambda_1$	$h_{\{1\}} = 1/(\lambda_2+\mu_1)$	$h_{\{2\}} = 1/\mu_2$

Table 4.2 Birth & Death Process: Standard regenerative simulation ($N = 10^6$).

$\lambda_1 = \lambda_2$	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	RE
10^{-1}	2.3753×10^{-03}	2.3728×10^{-03}	$[2.3429 \times 10^{-03}, 2.4027 \times 10^{-03}]$	2.3225×10^{-10}	4.504	1.0460×10^{-09}	0.00642
10^{-3}	2.4988×10^{-07}	2.5063×10^{-07}	$[2.1914 \times 10^{-07}, 2.8211 \times 10^{-07}]$	2.5807×10^{-16}	6.769	1.7469×10^{-15}	0.06410
10^{-5}	2.5000×10^{-11}	4.4000×10^{-12}	$[-8.3683 \times 10^{-13}, 9.6369 \times 10^{-12}]$	7.1389×10^{-24}	7.200	5.1400×10^{-23}	0.60724
10^{-7}	2.5000×10^{-15}	0.00	[0.00, 0.00]	0.00	–	–	–

Table 4.3 Birth & Death Process: Results of regenerative IS with CE optimization.

					$\mathbf{n_j = 10^4}$, $\mathbf{N = 10^6}$					
$\lambda_1 = \lambda_2$	S	$\check{\lambda}^{(j=S)}_{CE_2}, \check{\mu}^{(j=S)}_{CE_1}$	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-1}	2	1.1234, 1.0189	2.3753×10^{-03}	2.3768×10^{-03}	$[2.3621 \times 10^{-03}, 2.3915 \times 10^{-03}]$	5.6177×10^{-11}	7.179	4.0329×10^{-10}	**0.00315**	**2.59**
10^{-3}	3	1.0331, 1.0320	2.4988×10^{-07}	2.4929×10^{-07}	$[2.4760 \times 10^{-07}, 2.5098 \times 10^{-07}]$	7.4541×10^{-19}	7.124	5.3103×10^{-18}	**0.00346**	**328.95**
10^{-5}	4	1.0292, 1.0292	2.5000×10^{-11}	2.4943×10^{-11}	$[2.4773 \times 10^{-11}, 2.5112 \times 10^{-11}]$	7.4744×10^{-27}	8.479	6.3375×10^{-26}	**0.00347**	**811.04**
10^{-7}	5	1.0286, 1.0286	2.5000×10^{-15}	2.4943×10^{-15}	$[2.4774 \times 10^{-15}, 2.5112 \times 10^{-15}]$	7.4742×10^{-35}	6.799	5.0817×10^{-34}	**0.00347**	–
10^{-9}	6	1.0281, 1.0281	2.5000×10^{-19}	2.4943×10^{-19}	$[2.4774 \times 10^{-19}, 2.5112 \times 10^{-19}]$	7.4739×10^{-43}	7.780	5.8147×10^{-42}	**0.00347**	–

Table 4.4 Birth & Death Process: Optimizing *fail2* & *repair1* or only *fail2*.

Rarity → Transition Optimized	$\lambda_1 = \lambda_2 = 10^{-01}$ $\hat{\sigma}^2$ (RE)	$\lambda_1 = \lambda_2 = 10^{-03}$ $\hat{\sigma}^2$ (RE)	$\lambda_1 = \lambda_2 = 10^{-05}$ $\hat{\sigma}^2$ (RE)	$\lambda_1 = \lambda_2 = 10^{-07}$ $\hat{\sigma}^2$ (RE)	$\lambda_1 = \lambda_2 = 10^{-09}$ $\hat{\sigma}^2$ (RE)
fail2 & *repair1*	5.6177×10^{-11} (0.0032)	7.4541×10^{-19} (0.0035)	7.4744×10^{-27} (0.0035)	7.4742×10^{-35} (0.0035)	7.4739×10^{-43} (0.0035)
fail2	8.9633×10^{-11} (0.0040)	1.1884×10^{-18} (0.0044)	1.1767×10^{-26} (0.0044)	1.1769×10^{-34} (0.0044)	1.1769×10^{-42} (0.0044)

The original problem is defined by the choice of $\lambda_1 = \lambda_2$. For each λ_2, the values of the rates of transitions *fail2* and *repair1* are changed at each stage of the pre-simulation (progressively decreasing the rate of *fail2* transition and increasing the rate of *repair1* transition from first pre-simulation stage onward) as per the heuristic rule given in Section 4.3.4.

Each transition of interest is considered as a group itself and Equation 4.36 is used in Algorithm 3 at any stage with different k index for *fail2* and *repair1* in the IS vector $\check{\theta}_{ce}^{(j)}$. There are $j = 1,...,S$ pre-simulation stages, with $n_j = 10^4$ cycles simulated at each stage, and the main simulation uses $N = 10^6$ cycles for estimation. Also, we increase the number of stages S as $\lambda_1 = \lambda_2 \longrightarrow 0$ to break down rare problems into higher number of smaller easily solvable sub-problems in S stages. Results are presented in Table 4.3 and we compare the performance of the estimators asymptotically in terms of rarity (i.e., $\lambda_1 = \lambda_2 \longrightarrow 0$) obtained by simulation using the CE Algorithm 3.

As it is evident from the results obtained, the CE optimized values of the IS rates (for each $\lambda_1 = \lambda_2$) are obtained from the proposed CE scheme ($\tilde{\lambda}_{ce_2}^{(j=S)}$ and $\tilde{\mu}_{ce_1}^{(j=S)}$) from the final pre-simulation stage ($j = S$), as shown in Table 4.3. Using the specific IS rates obtained, we compare the results for any specific value of $\lambda_1 = \lambda_2$ with the standard regenerative simulation. The empirical values of RE and the variance of the estimator ($\hat{\sigma}^2$), obtained via CE scheme is comparatively far better (i.e., lower) than a standard regenerative simulation (for each $\lambda_1 = \lambda_2$), see Tables 4.2 & 4.3. Also, using the CE scheme, the RE (or the relative width of the 95% CI) is bounded as the rarity increases (i.e., the BRE property). In terms of gain (previously defined by Equation 2.25), for example when $\lambda_1 = \lambda_2 = 10^{-5}$, the ratio of the $\hat{\sigma}_{wn}^2$ obtained from a standard regenerative simulation and the CE optimization scheme with regenerative IS, we obtain a gain of 811 times approximately.

Application of CE scheme on failure transitions only: Previously, we also discussed application of the CE scheme only on the failure transitions. In Table 4.4, we compare the results (in terms of estimated values of RE and variance) when applying the CE scheme on both *fail2* and *repair1* or only on *fail2* transitions. The values of S, n_j and N are the same for each $\lambda_1 = \lambda_2$ as was used in the results shown in Table 4.3. It is observable from the empirical results in Table 4.4, that when the CE scheme is applied only on the *fail2* transition, we still observe a BRE property asymptotically. In terms of the values of RE and estimated variance $\hat{\sigma}^2$, optimizing both *fail2* and *repair1* results in slightly more accuracy (i.e., lower RE and $\hat{\sigma}^2$) as compared to optimizing only *fail2*. This is possible as optimizing more number of transitions via CE scheme also improves the application of IS in estimations, as we also obtain a higher dimensional optimizer. However, as the number of transitions of interest increases (in larger models), this can result in more statistical noise with same number of cycles simulated. A solution to this problem would be to either increase the number of pre-simulation cycles (n_j) or to use a grouping approach.

CE scheme leads to Variance Minimization: The entire objective of the application of IS and CE scheme in conjunction is to obtain variance reduction of the final estimator $\hat{U}$. It was previously discussed in Section 2.4.1, that the optimal VM and CE densities are asymptotically (in terms of rarity of EOI) identical or very close. The $f(\cdot;\tilde{\theta}^*_{ce})$ density which is closest to $g^*(x)$ in terms of CE distance is also the one for which the asymptotic variance of the estimator is minimum [32]. From Algorithm 3, the optimized CE density (given by $\tilde{\theta}^*_{ce}$) contains the optimized IS rates $\check{\lambda}^{(j=S)}_{ce_2}$ and $\check{\mu}^{(j=S)}_{ce_1}$ for this example. Let us consider Equation 4.10 where the goal is to maximize a CE function to reduce the CE distance with respect to IS change of measure $\tilde{\theta}$. For $\check{\theta} = \tilde{\theta}$, the Equation 4.10 can be rewritten as:

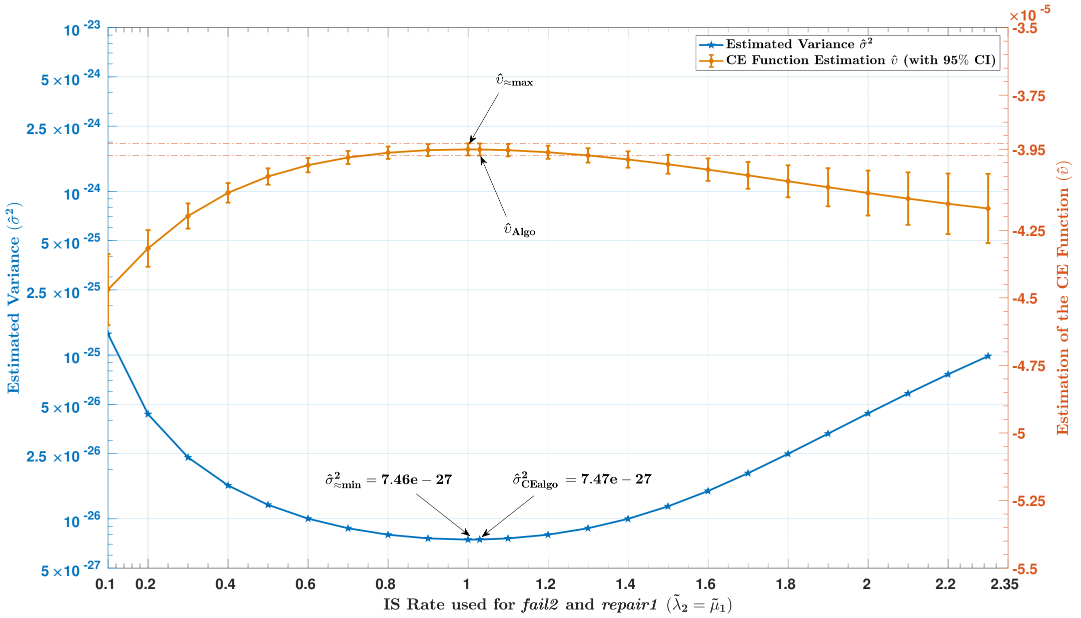

Figure 4.2 Birth & Death Process: Variance ($\hat{\sigma}^2$) & CE function $\hat{\upsilon}$ with respect to IS rates ($\tilde{\lambda}_2 = \tilde{\mu}_1$) for $\lambda_1 = \lambda_2 = 10^{-05}$.

$$\max_{\tilde{\theta}\in\Theta} \upsilon(\tilde{\theta}) = \max_{\tilde{\theta}\in\Theta} \mathbb{E}_{\tilde{\theta}}\left[|Z|\, L(X;\theta;\tilde{\theta})\; log\; f(x;\tilde{\theta})\right], \tag{4.41}$$

where $\upsilon(\tilde{\theta})$ is the CE function that can be estimated by:

$$\hat{\upsilon}(\tilde{\theta}) = \frac{1}{N}\, \Sigma_{m=1}^{N}\, Z_m(\omega_m)\, L(\omega_m;\theta;\tilde{\theta})\; log\; f(\omega_m;\tilde{\theta}). \tag{4.42}$$

In the above equation, N is the total number of cycles simulated and sampling is done from the IS density given by $\tilde{\theta}$.

We try to maximize the above Equation 4.42, by trial and error using a simple regenerative IS simulation technique. In the current example, for $\lambda_1 = \lambda_2 = 10^{-05}$, we try to estimate $\hat{\upsilon}$ for different values of $\tilde{\lambda}_2 = \tilde{\mu}_1$, where $\{\tilde{\theta} = \tilde{\theta}_{\tilde{\mathscr{F}}} \cup \theta_{\mathscr{R}}\}$ and $\{\tilde{\theta}_{\tilde{\mathscr{F}}} = (\tilde{\lambda}_2, \tilde{\mu}_1)\}$. Figure 4.2 shows the result of the empirical values of $\hat{\upsilon}(\tilde{\theta})$ and the estimated variance $\hat{\sigma}^2$ of $\hat{U}$ for different values of $\tilde{\lambda}_2 = \tilde{\mu}_1$ tried. In this case there is no pre-simulation and $N = 10^6$.

For a specific value of $\lambda_1 = \lambda_2 = 10^{-05}$, from Algorithm 3, we previously obtained the optimized IS rates $\check{\lambda}^{(j=4)}_{ce_2} = \check{\mu}^{(j=4)}_{ce_1} = 1.0292$ (in Table 4.3). In Figure 4.2 we can see the results obtained by changing the values of $\tilde{\lambda}_2 = \tilde{\mu}_1$ over a specific range of values (between 0.01 and 2.3), beyond that, the $\hat{\sigma}^2$ continues to increase. It is observed in Figure 4.2 that the approximated maxima for the Equation 4.42 is around $\tilde{\lambda}_2 = \tilde{\mu}_1 \approx 1.0$ which is very close to the values obtained from Algorithm 3 (where $\check{\lambda}_{ce_2}^{(j=4)} = \check{\mu}_{ce_1}^{(j=4)} = 1.0292$). The value of estimated variance obtained from the trial and error scheme here ($\hat{\sigma}^2 = 7.46 \times 10^{-27}$) is also approximately the one obtained using the IS rates from the CE scheme ($\hat{\sigma}^2 = 7.47 \times 10^{-27}$).

Thus, from this small example, we can conclude that the proposed CE scheme based on progressive shifting of rarity, is able to estimate U efficiently and also leads to minimized variance asymptotically. Also, the progressive shifting within each problem $f(x;\theta)$ leads to BRE property asymptotically, when the original problem rarity increases $\lambda_1 = \lambda_2 \to 0$.

4.5 Example 2: A 2oo3 System with Logistics

In this case, a more complex example of a 2oo3 (2-out-of-3) system is considered where logistics aspects are included. The Markovian SPN has 273 different markings in the reachability graph (i.e., states of the underlying CTMC). The model is briefly described in the following section.

4.5.1 Model Description of a 2oo3 System

The SPN model used in this example is a model of a 2oo3 redundant subsystem that can be considered as part of a larger system. The model includes logistics to represent some practical aspects of real passenger rail systems. Logistics aspects included here are availability of spares, a restoration team in a depot (one repair person in this case), timed inspection of components for any failures and travel time to the site of components for repair/inspection. The Markovian SPN model of this system consists of immediate transitions also (along with exponential transitions) and the reachability graph is obtained after eliminating the markings due to the immediate transitions (also called as vanishing markings). A detailed description of the Markovian SPN model of the system is given in Appendix B. The down state $\{\mathscr{D}\}$ (i.e., system failure) is reached if 2 or more components are failed. Each components (A, B or C) have two kind of failure modes: detected and undetected. There is also a common cause failure (ccf) mode that causes all the components of the 2oo3 module to fail simultaneously. There are seven failure transitions (transitions of interest), namely, detected failures of components (*detA*, *detB* and *detC*), undetected failures of components (*udetA*, *udetB* and *udetC*) and the *ccf*. Following rates are the ones used for all other transitions:

- Rate at which spares become available = 5.0
- Rate at which spares become unavailable = 0.1
- Rate at which undetected failures are detected = 10.0
- Rate at which on-site technicians start inspections = 3.0

- Rate of timed inspections = 0.1
- Travel rate for technicians = 1.0
- Repair rate = 1.0

In this example, for all the detected and undetected failures of all components (i.e., A, B or C), the failure rate used is λ. There is also a ccf transition considered that can fire from any given marking of the SPN. This ccf transition is important to be considered for a real passenger rail system. For example, a power supply failure in certain rail systems can act as a common cause that brings all the inherent components of a system down simultaneously (and consequently the system too). The *ccf* transition (event) in this case can be triggered at any instant of time from any given marking of the SPN in a cycle, including the initial state $\{\mathbf{0}\}$ for this example, and its firing brings the system to the down state $\{\mathscr{D}\}$ directly. In this example, we consider a system can be balanced or unbalanced depending on the failure rate chosen for the ccf transitions relative to the failure rate chosen for individual component failures. The general definition is given in Section 4.2.1.

In the current example, let us consider $p_1(\lambda)$ is probability of a path to failure due to component failures in a cycle and $p_2(\lambda)$ be the probability of a path to failure due to a ccf event. We exclude the possibility of a ccf event occurring after a failure of an individual component within a cycle. The rate of a ccf is usually lower than the one of individual components. If we consider for a balanced system, where ccf occurs with a rate of λ^2 and individual components fail with a rate λ, then in such a case, $p_1(\lambda) = \Theta(p_2(\lambda))$ as $\lambda \to 0$ (as presented in Section 4.2.1). In an unbalanced system, where we consider ccf occurs with a rate of 0.01λ, then $p_1(\lambda) = o(p_2(\lambda))$ asymptotically (as presented in Section 4.2.1). For an unbalanced case here, the possibility of a ccf transition firing would become increasingly higher asymptotically as compared to the rate of failures of two components in a cycle. This would result in ccf being the most dominant transition in the current example bringing the system to a failed state $\{\mathscr{D}\}$ as compared to any two components failing in a cycle. Under this assumption of ccf being a dominating transition or not, we consider the system to be balanced or unbalanced respectively.

It is to be noted that generally, in HRMS models, a direct path to failure from an initial state is not considered. In our case here, we consider the ccf as they are more representative of real passenger rail systems where ccf is considered for redundant subsystems, like the current example. In the next sections, we discuss the results obtained via application of the CE Algorithm 3 on the current 2oo3 system with logistics when it is balanced and unbalanced. The rarity of a system failure is increased by reducing the value of λ (i.e., $\lambda \to 0$). An exact numerical solution of the steady-state unavailability U is also obtainable due to the moderate size of the problem here.

4.5.2 Empirical Results and Interpretations: A Balanced 2oo3 system

In this case of a balanced 2oo3 module, the failure rates for the transitions in subset vector $\theta_{\mathscr{F}}$ are considered as:

- Failure of individual components (*detA*, *detB*, *detC*,*udetA*, *udetB*, *udetC*): $= \lambda$
- Common cause failure (*ccf*): $= \lambda^2$

In Table 4.5, results are presented from a standard regenerative simulation for $N = 10^7$ cycles. As it is expected from standard simulations, the results are increasingly inaccurate as $\lambda \to 0$. The RE and the 95% CI is increased as $\lambda \to 0$. For $\lambda < 10^{-05}$, the standard simulation gives 0.0 values for point estimates and no failures are sampled.

Table 4.5 Balanced 2oo3 system: Standard regenerative simulation ($N = 10^7$).

λ	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	RE
10^{-3}	8.6716×10^{-04}	9.0961×10^{-04}	$[8.7905 \times 10^{-04}, 9.4017 \times 10^{-04}]$	2.4312×10^{-10}	230.759	5.6102×10^{-08}	0.01714
10^{-5}	9.3715×10^{-08}	1.1976×10^{-07}	$[-1.1497 \times 10^{-07}, 3.5448 \times 10^{-07}]$	1.4342×10^{-14}	217.508	3.1195×10^{-12}	1.00000
10^{-7}	9.3790×10^{-12}	0.00	[0.00, 0.00]	0.00	–	–	–

Application of IS with regenerative simulation scheme is expected to provide variance reduction and more accurate results as compared to standard regenerative simulations. For the current example, IS is applied by changing the failure rates of all transitions in original vector $\theta_{\mathscr{F}}$ (i.e., the transitions *detA*, *detB*, *detC*, *udetA*, *udetB*, *udetC*, *ccf*).

Table 4.6 Balanced 2oo3 system: Regenerative IS simulation with $\tilde{\lambda} = 0.01$.

λ	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	RE
10^{-3}	8.6716×10^{-04}	8.5484×10^{-04}	$[8.3583 \times 10^{-04}, 8.7384 \times 10^{-04}]$	9.4004×10^{-11}	78.642	7.3926×10^{-09}	0.01134
10^{-5}	9.3715×10^{-08}	9.2455×10^{-08}	$[9.0104 \times 10^{-08}, 9.4805 \times 10^{-08}]$	1.4384×10^{-18}	76.747	1.1039×10^{-16}	0.01297
10^{-7}	9.3790×10^{-12}	9.2529×10^{-12}	$[9.0173 \times 10^{-12}, 9.4885 \times 10^{-12}]$	1.4449×10^{-26}	77.316	1.1171×10^{-24}	0.01299
10^{-9}	9.3790×10^{-16}	9.2530×10^{-16}	$[9.0174 \times 10^{-16}, 9.4886 \times 10^{-16}]$	1.4450×10^{-34}	75.209	1.0868×10^{-32}	0.01299
10^{-11}	9.3790×10^{-20}	9.2530×10^{-20}	$[9.0174 \times 10^{-20}, 9.4886 \times 10^{-20}]$	1.4450×10^{-42}	79.711	1.1518×10^{-40}	0.01299

At first, by trial-and-error, a one-dimensional change of measure $\tilde{\lambda} = 0.01$ is obtained, meaning all the failure rates of transitions in the original vector $\theta_{\mathscr{F}}$ are replaced by a common value ($\tilde{\lambda}$) in the IS vector $\tilde{\theta}_{\mathscr{F}}$, as previously explained in Section 4.3.2. This one-dimensional change of measure is found by attempting to reduce the variance of the final estimator $\hat{U}$ by manually tuning the value of $\tilde{\lambda}$. The empirical results shown in Table 4.6 show that the variance ($\hat{\sigma}^2$) is reduced by a large factor in comparison to the standard regenerative simulation. It is worth noticing that as $\lambda \to 0$, with $\tilde{\lambda} = 0.01$, the empirical value of the RE is bounded (≈ 0.01). However, it is not feasible to find the change of measure by trial and error always. For this purpose, we show the usefulness of the proposed CE Algorithm 3 to approximate optimal IS rates for transitions in subset vector $\theta_{\mathscr{F}}$.

Progressive shifting of rarity and choice of parameters in Algorithm 3: The proposed algorithm is applied to each original problem defined by λ and the growth of the RE and the estimation of the gain are considered as measures of accuracy/efficiency. For each original problem given by λ, the system is first made unstable by increasing the rates of the failure transitions of subset $\mathscr{F}$ for the first stage and then gradually decreasing these failure rates until the problem reaches the original problem defined by λ (as per the heuristic rule given in Section 4.3.4).

Table 4.7 Balanced 2oo3 system: Regenerative IS simulation with CE One-D optimization on failure subset only.

					$\mathbf{n_j = 5 \times 10^4 (n_{(j=S)} = 2 \times 10^5), N = 10^6}$					
λ	S	$\check{\lambda}_{ce}^{(j=S)}$	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-3}	3	1.8936×10^{-02}	8.6716×10^{-04}	8.6691×10^{-04}	$[8.4168 \times 10^{-04}, 8.9213 \times 10^{-04}]$	1.6563×10^{-10}	130.327	2.1586×10^{-08}	**0.01485**	**2.60**
10^{-5}	4	1.7669×10^{-02}	9.3715×10^{-08}	9.5550×10^{-08}	$[9.2571 \times 10^{-08}, 9.8529 \times 10^{-08}]$	2.3101×10^{-18}	121.702	2.8115×10^{-16}	**0.01591**	**11095.62**
10^{-7}	5	1.7959×10^{-02}	9.3790×10^{-12}	9.5828×10^{-12}	$[9.2929 \times 10^{-12}, 9.8727 \times 10^{-12}]$	2.1881×10^{-26}	119.317	2.6108×10^{-24}	**0.01544**	–
10^{-9}	6	1.8042×10^{-02}	9.3790×10^{-16}	9.5904×10^{-16}	$[9.3009 \times 10^{-16}, 9.8799 \times 10^{-16}]$	2.1816×10^{-34}	122.883	2.6809×10^{-32}	**0.01540**	–
10^{-11}	7	1.7376×10^{-02}	9.3790×10^{-20}	9.4789×10^{-20}	$[9.1958 \times 10^{-20}, 9.7620 \times 10^{-20}]$	2.0863×10^{-42}	121.306	2.5308×10^{-40}	**0.01524**	–

Table 4.8 Balanced 2oo3 system: Regenerative IS simulation with CE Multi-D optimization on failure subset only.

				$\mathbf{n_j = 5 \times 10^4 (n_{(j=S)} = 2 \times 10^5), N = 10^6}$					
λ	S	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-3}	3	8.6716×10^{-04}	8.8023×10^{-04}	$[8.6455 \times 10^{-04}, 8.9592 \times 10^{-04}]$	6.4067×10^{-11}	133.301	8.5401×10^{-09}	**0.00909**	**6.57**
10^{-5}	4	9.3715×10^{-08}	9.3335×10^{-08}	$[9.1514 \times 10^{-08}, 9.5157 \times 10^{-08}]$	8.6381×10^{-19}	128.157	1.1070×10^{-16}	**0.00996**	**28178.60**
10^{-7}	5	9.3790×10^{-12}	9.3544×10^{-12}	$[9.1776 \times 10^{-12}, 9.5313 \times 10^{-12}]$	8.1408×10^{-27}	131.964	1.0743×10^{-24}	**0.00965**	–
10^{-9}	6	9.3790×10^{-16}	9.3967×10^{-16}	$[9.2203 \times 10^{-16}, 9.5731 \times 10^{-16}]$	8.0993×10^{-35}	135.739	1.0994×10^{-32}	**0.00958**	–
10^{-11}	7	9.3790×10^{-20}	9.4398×10^{-20}	$[9.2661 \times 10^{-20}, 9.6135 \times 10^{-20}]$	7.8516×10^{-43}	139.269	1.0935×10^{-40}	**0.00939**	–

Table 4.9 Balanced 2oo3 system: Regenerative IS simulation with CE Multi-D optimization on all transitions.

				$\mathbf{n_j = 5 \times 10^4 (n_{(j=S)} = 2 \times 10^5), N = 10^6}$					
λ	S	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-3}	3	8.6716×10^{-04}	8.6745×10^{-04}	$[8.4979 \times 10^{-04}, 8.8510 \times 10^{-04}]$	8.1147×10^{-11}	248.509	2.0166×10^{-08}	**0.01038**	**2.78**
10^{-5}	4	9.3715×10^{-08}	9.3894×10^{-08}	$[9.3063 \times 10^{-08}, 9.4724 \times 10^{-08}]$	1.7967×10^{-19}	239.737	4.3073×10^{-17}	**0.00451**	**72422.94**
10^{-7}	5	9.3790×10^{-12}	9.3282×10^{-12}	$[9.2464 \times 10^{-12}, 9.4100 \times 10^{-12}]$	1.7415×10^{-27}	233.491	4.0663×10^{-25}	**0.00447**	–
10^{-9}	6	9.3790×10^{-16}	9.4160×10^{-16}	$[9.3352 \times 10^{-16}, 9.4969 \times 10^{-16}]$	1.7009×10^{-35}	242.714	4.1283×10^{-33}	**0.00438**	–
10^{-11}	7	9.3790×10^{-20}	9.4484×10^{-20}	$[9.3669 \times 10^{-20}, 9.5299 \times 10^{-20}]$	1.7287×10^{-43}	252.321	4.3619×10^{-41}	**0.00440**	–

In this case, we use $n_j = 5 \times 10^4$ for each pre-simulation stage and for the final pre-simulation stage where the original problem (given by λ) is considered, we simulate higher number of cycles (2×10^5) so that the IS rates obtained from final stage have less statistical noise. The main simulation uses $N = 10^6$ cycles.

It is to be noted that if repair or inspection transitions (that have values of rates relatively higher than failure transitions) are considered to be part of subset $\mathscr{F}$, then these rates are actually first decreased in the first pre-simulation stage and then gradually increased to reach the original rates. This way, we can also create an unstable system by increasing failure rates and decreasing repair rates and then gradually increasing the stability of the system (i.e., increasing the rarity of system failure) by decreasing failure rates and increasing repair rates, until all of them reach the original values for that problem.

We use the above heuristics for the progressive shifting of the rarity (within each problem) and with the choice of the given parameters, we show the usefulness of the proposed CE Algorithm 3 with three different grouping strategies. The strategies and their results are discussed as follows.

Strategy 1 for balanced case: One-dimensional optimization of failure transitions

In this strategy, we try applying the CE scheme to find a one-dimensional change of measure on only failure transitions and we compare the results with the standard regenerative simulation (as given in Table 4.5). We also compare the results to the results obtained from the trial and error method as previously discussed (see Table 4.6). The results obtained from this strategy are shown in Table 4.7.

With the specific assumptions of Strategy 1, all the failure transitions are considered to be part of the subset $\mathscr{F}$ and $\tilde{\mathscr{F}}$. The grouping is done as shown in Equation 4.22 with a common k index for all transitions of interest. In such a case, the IS rate is computed via CE scheme taking into account the combined contribution of all the transitions in the group towards the failure state. The results of this strategy using the proposed CE scheme show a BRE property as $\lambda \to 0$. Comparing the results with the standard regenerative simulation, for example for $\lambda = 10^{-5}$, through the CE scheme we obtain a RE$= 0.01591$ (see Table 4.7) while standard simulation gave a large RE$= 1.0$. In terms of gain, the CE scheme proves to be $\approx 11{,}095$ times better than a standard simulation. If we compare the results of the CE scheme with the trial and error method (in Tables 4.6 & 4.7), for example when $\lambda = 10^{-9}$, we obtain from the CE scheme the optimized one-dimensional IS rate $\check{\lambda}_{ce}^{(j=S)} = 0.018$ (with RE$= 0.015$), while the trial and error method using $\tilde{\lambda} = 0.01$ provided a RE estimate of 0.013. We can say from these empirical values that the CE scheme approximates the CE optimized IS rates and we obtain large variance reductions.

Strategy 2 for balanced case: Multi-dimensional optimization of failure transitions

In this strategy, we try applying the CE scheme to find the CE optimized IS rates of each failure transition separately. Therefore, each transition is a group within itself as shown by Equation 4.16. The results are presented in Table 4.8.

In this case also we observe the BRE property as $\lambda \to 0$. Also, the gain in terms of work-normalized variance, for example when $\lambda = 10^{-5}$ is $\approx 28,178$ times as compared to a standard regenerative simulation (see Table 4.5). The values of the estimated RE in this case is lower than the one-dimensional strategy as the IS rates of the failure transitions are optimized separately and hence their contributions are captured more accurately, in sufficient n_j number of cycles.

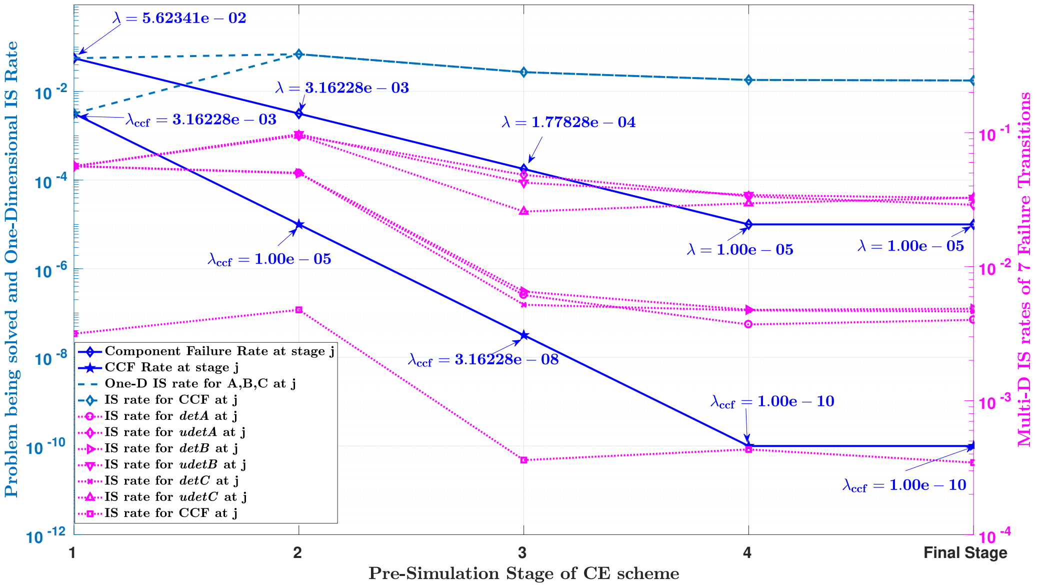

Figure 4.3 Progressive shifting of the problem and IS rates with One-D & Multi-D optimization.

Representation of Progressive Shifting of the Problem for One-D and Multi-D case: We present the generic idea of the Algorithm 3 in Figure 4.3 in context of this example for the original problem where $\lambda = 10^{-5}$. The figure shows how a one-dimensional IS change of measure for all failure transitions (as given in Table 4.7) and multi-dimensional IS change of measure for each failure transition (as given in Table 4.8) are obtained in 4 pre-simulation stages. At $j = 1$ stage, the failure rates of components as well as the ccf transition are increased and an unstable system is obtained. The IS rates for each transition are chosen to be the same as their original rates, in both one-dimensional and multi-dimensional strategies. In the subsequent stages ($j = 2$ and 3), the rates of both component failures and ccf transition are decreased gradually as per the heuristic rules discussed in Section 4.3.4. The last stage of pre-simulation is at $j = 4$ in the Figure 4.3, where the original problem (where components failure rates are $\lambda = 10^{-5}$ and *ccf* transition is $\lambda^2 = 10^{-10}$) is solved using the IS rates from the previous stage. Finally, the final stage is the main simulation which uses the IS rates obtained from the stage $j = 4$.

Strategy 3 for balanced case: Multi-dimensional optimization of all transitions

In this final strategy, all the transitions, including the ones of failures, repairs and inspections of the model are considered to be transitions of interest. Thus, all these transitions are considered to be in subsets $\mathscr{F}$ and $\tilde{\mathscr{F}}$ and the CE optimization scheme is applied on each of these transitions uniquely, meaning each transition being a group within itself. The results obtained via this strategy are shown in Table 4.9 and the RE is bounded in this case also (≈ 0.004) as $\lambda \to 0$. In this case, for $\lambda = 10^{-5}$, the gain is the maximum $\approx 72,422$ times and also the specific values of RE (for each λ) are much lower as compared to the results obtained from the first two strategies.

Again, the model being of moderate size, more number of transitions can be included to be optimized and results are better. However, as previously discussed, application of IS on large dimensional problems could lead to likelihood degeneracy issues and also in larger models it could result in more statistical noise in the solution of the equation used for the Algorithm 3. Nevertheless, the CE scheme works efficiently for this example, and we obtain the desired BRE property as well as huge variance reduction (and gains) using the proposed CE Algorithm 3 here.

4.5.3 Empirical Results and Interpretations: An Unbalanced 2oo3 System

The second case considered here is that of an unbalanced 2oo3 system. In this case, the failure rates for the transitions in subset vector $\theta_{\mathscr{F}}$ are considered as:

- Failure of individual components (*detA*, *detB*, *detC*, *udetA*, *udetB*, *udetC*): $= \lambda$
- Common cause failure (*ccf*): $= 0.01\lambda$

Here, when $\lambda \to 0$, $p_1(\lambda) = o(p_2(\lambda))$, where $p_1(\lambda)$ is sample path probability due to component failures, and $p_2(\lambda)$ is due to a possible ccf event. Due to this, $p_2(\lambda)$ relatively increases as compared to $p_1(\lambda)$ asymptotically. The effect of this in the current example is that asymptotically ($\lambda \to 0$), ccf becomes the most dominating transition towards failure set $\{\mathscr{D}\}$; while individual component failures contribution towards $\{\mathscr{D}\}$ is relatively very low (or even negligible).

Table 4.10 Unbalanced 2oo3 system: Standard regenerative simulation ($N = 10^7$).

λ	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	RE
10^{-3}	8.8784×10^{-04}	9.2871×10^{-04}	$[8.9807 \times 10^{-04}, 9.5934 \times 10^{-04}]$	2.4428×10^{-10}	209.205	5.1105×10^{-08}	0.01683
10^{-5}	3.3239×10^{-07}	2.4840×10^{-07}	$[-1.5285 \times 10^{-08}, 5.1209 \times 10^{-07}]$	1.8100×10^{-14}	205.643	3.7220×10^{-12}	0.54160
10^{-7}	2.3994×10^{-09}	0.00	[0.00, 0.00]	0.00	–	–	–

For this unbalanced case also we first performed standard regenerative simulations, results of which are shown in Table 4.10. As a commonly known observation, the RE increases as $\lambda \to 0$ and for $\lambda < 10^{-05}$, the standard simulation gives useless empirical values of 0.0 for point estimations as no system failure event occurs even for $N = 10^7$ cycles simulated.

Table 4.11 Unbalanced 2oo3 system: Regenerative IS simulation with CE One-D optimization on failure subset only.

					$\mathbf{n_j = 5 \times 10^4 (n_{(j=S)} = 2 \times 10^5), N = 10^6}$					
λ	S	$\check{\lambda}_{ce}^{(j=S)}$	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-3}	3	1.9093×10^{-02}	8.8784×10^{-04}	8.8905×10^{-04}	$[8.6421 \times 10^{-04}, 9.1388 \times 10^{-04}]$	1.6056×10^{-10}	122.684	1.9699×10^{-08}	**0.01425**	**2.59**
10^{-5}	4	2.0741×10^{-02}	3.3239×10^{-07}	3.3210×10^{-07}	$[3.2662 \times 10^{-07}, 3.3758 \times 10^{-07}]$	7.8180×10^{-18}	121.198	9.4753×10^{-16}	**0.00842**	**3928.16**
10^{-7}	5	2.1925×10^{-02}	2.3994×10^{-09}	2.4091×10^{-09}	$[2.3671 \times 10^{-09}, 2.4511 \times 10^{-09}]$	4.5903×10^{-22}	129.443	5.9418×10^{-20}	**0.00889**	–
10^{-9}	6	2.2100×10^{-02}	2.3902×10^{-11}	2.4012×10^{-11}	$[2.3592 \times 10^{-11}, 2.4431 \times 10^{-11}]$	4.5835×10^{-26}	127.792	5.8574×10^{-24}	**0.00892**	–
10^{-11}	7	2.2969×10^{-02}	2.3901×10^{-13}	2.4124×10^{-13}	$[2.3679 \times 10^{-13}, 2.4568 \times 10^{-13}]$	5.1428×10^{-30}	133.611	6.8713×10^{-28}	**0.00940**	–

Table 4.12 Unbalanced 2oo3 system: Regenerative IS simulation with CE Multi-D optimization on failure subset only.

				$\mathbf{n_j = 5 \times 10^4 (n_{(j=S)} = 2 \times 10^5), N = 10^6}$					
λ	S	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-3}	3	8.8784×10^{-04}	8.7901×10^{-04}	$[8.6437 \times 10^{-04}, 8.9364 \times 10^{-04}]$	5.5758×10^{-11}	131.572	7.3362×10^{-09}	**0.00850**	**6.97**
10^{-5}	4	3.3239×10^{-07}	3.3454×10^{-07}	$[3.2854 \times 10^{-07}, 3.4053 \times 10^{-07}]$	9.3563×10^{-18}	126.259	1.1813×10^{-15}	**0.00914**	**3150.76**
10^{-7}	5	2.3994×10^{-09}	2.2975×10^{-09}	$[2.0711 \times 10^{-09}, 2.5239 \times 10^{-09}]$	1.3346×10^{-20}	112.435	1.5005×10^{-18}	**0.05028**	–
10^{-9}	6	2.3902×10^{-11}	2.2795×10^{-11}	$[2.0359 \times 10^{-11}, 2.5230 \times 10^{-11}]$	1.5438×10^{-24}	120.189	1.8555×10^{-22}	**0.05451**	–
10^{-11}	7	2.3901×10^{-13}	2.2751×10^{-13}	$[2.0331 \times 10^{-13}, 2.5171 \times 10^{-13}]$	1.5245×10^{-28}	124.568	1.8991×10^{-26}	**0.05427**	–

Table 4.13 Unbalanced 2oo3 system: Regenerative IS simulation with CE Multi-D optimization on all transitions.

				$\mathbf{n_j = 5 \times 10^4 (n_{(j=S)} = 2 \times 10^5), N = 10^6}$					
λ	S	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-3}	3	8.8784×10^{-04}	8.9074×10^{-04}	$[8.7864 \times 10^{-04}, 9.0283 \times 10^{-04}]$	3.8078×10^{-11}	247.607	9.4285×10^{-09}	**0.00693**	**5.42**
10^{-5}	4	3.3239×10^{-07}	3.3261×10^{-07}	$[3.2928 \times 10^{-07}, 3.3595 \times 10^{-07}]$	2.9002×10^{-18}	206.048	5.9758×10^{-16}	**0.00512**	**6228.53**
10^{-7}	5	2.3994×10^{-09}	2.3999×10^{-09}	$[2.3810 \times 10^{-09}, 2.4187 \times 10^{-09}]$	9.2524×10^{-23}	187.155	1.7316×10^{-20}	**0.00401**	–
10^{-9}	6	2.3902×10^{-11}	2.3894×10^{-11}	$[2.3717 \times 10^{-11}, 2.4071 \times 10^{-11}]$	8.1531×10^{-27}	188.266	1.5350×10^{-24}	**0.00378**	–
10^{-11}	7	2.3901×10^{-13}	2.3931×10^{-13}	$[2.3747 \times 10^{-13}, 2.4115 \times 10^{-13}]$	8.8598×10^{-31}	193.323	1.7128×10^{-28}	**0.00393**	–

As previously done for the balanced case, here also a one-dimensional change of measure ($\tilde{\lambda} = 0.01$) was used (by trial and error). Experimental results showed a huge variance reduction (and BRE property), similar to the balanced case. However, our objective is to find the optimal IS rates using CE Algorithm 3.

The progressive shifting of rarity and the choice of the parameters for Algorithm 3 for this case have the same reasoning as explained for the balanced case previously. In this unbalanced case also, we apply the three different strategies as done previously for the balanced module of this 2oo3 system.

Strategy 1 for unbalanced case: One-dimensional optimization of failure transitions

In this case, we apply the proposed CE scheme on all failure transitions (divided in subsets $\mathscr{F}$ and $\tilde{\mathscr{F}}$) and a one-dimensional change of measure is obtained where all failure transitions are considered to be in a single group (as shown in Equation 4.22). The results of using Strategy 1 are shown in Table 4.11, where it is observable that even in this unbalanced system, the RE is approximately 0.009 value as $\lambda \to 0$. In fact, the slight increase in RE as $\lambda \to 0$ can be resolved by using more number of pre-simulation stages to break down the original problems further or by using more number of cycles at each stage. In terms of gain with respect to standard simulation (see Table 4.10), for example when $\lambda = 10^{-05}$, we see that the CE scheme applied using the one-dimensional change of measure provides approximately $3,928$ times better results. Also, the exact U is bounded within the 95% CI.

Strategy 2 for unbalanced case: Multi-dimensional optimization of failure transitions

In the second strategy, the CE scheme is applied on all failure transitions and they are optimized separately as shown in Equation 4.16. Results are shown in Table 4.12. The results from Strategy 2 are less accurate (in terms of higher RE and $\hat{\sigma}^2$) as compared to the results obtained through one-dimensional strategy. The RE is still bounded (≈ 0.05). Compared to a standard simulation, we do obtain better results due to the BRE property. For example, when $\lambda = 10^{-05}$, the gain obtained by using this strategy is ≈ 3150 times.

Remark: In this strategy, we observe that the RE values for higher values of λ (less rarity) are better compared to the one-dimensional Strategy 1 used above. However, as $\lambda \to 0$, the RE values obtained are worse than the ones of the one-dimensional strategy. Figure 4.4 shows this trend between Strategy 1 and Strategy 2. As previously presented for the balanced case, the Strategy 2 (results from Table 4.12) must provide better accuracy and lower RE as optimizing each failure transition is better in terms of optimization, as compared to optimizing them in a group to find an averaged value of IS rates common for all of them. However, in Figure 4.4, as $\lambda \to 0$, the system becomes increasingly imbalanced, i.e., $p_1(\lambda) = o(p_2(\lambda))$. The effect of this imbalance in Algorithm 3 is that the ccf becomes the most dominating transition and the contribution of individual failures of components becomes relatively negligible as $\lambda \to 0$. This would result in the component failure transitions providing zero values for the update Equation 4.36 and only the ccf transition would be

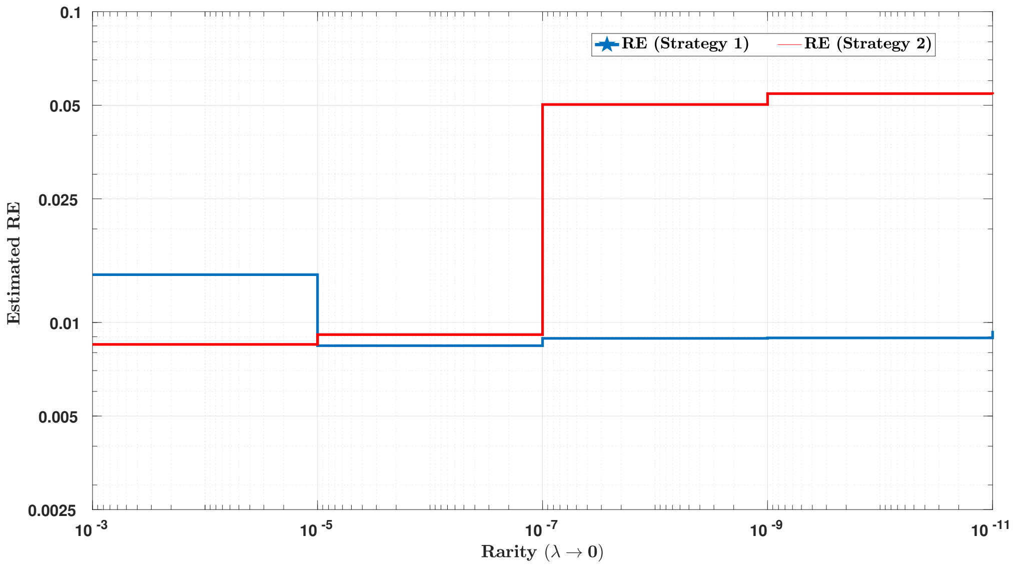

Figure 4.4 2oo3 Unbalanced System: Estimated RE with increasing imbalance.

accounted for. Thus, here we can conclude that when applying the proposed CE scheme on failure transitions only, it should be kept in mind if there exist a single SPN transition in the failure subset that dominates all other failure transitions or not. One-Dimensional change of measure (i.e., grouping) is a good strategy in such a case to find averaged values of CE optimized IS rates. Grouping helps in this case to take into account a combined contribution of all failure transitions, even if this optimization approach is lower dimensional (i.e., more restrictive) than a multi-dimensional one. Nevertheless, we observe a BRE property in this case too and the exact values of U are also bounded within the 95% CI.

Strategy 3 for unbalanced case: Multi-dimensional optimization of all transitions

In this strategy, the CE Algorithm 3 is used for optimizing the IS rates of all the transitions of the SPN model (except immediate transitions). Results of using this strategy are shown in Table 4.13. It is observable that optimizing all the transitions results in the lowest estimator variance $\hat{\sigma}^2$ and RE obtained when compared to any other IS strategy used in results in Tables 4.10-4.12. The RE is bounded (≈ 0.004) while the $\hat{\sigma}^2$ is also lowest. In this case also, as $\lambda \to 0$, we observed failure transitions of individual components not being accounted for in the CE scheme due to the increasing imbalance of the system. It is also likely that since individual component failures are not contributing to reach the failure set as $\lambda \to 0$, IS is not supposed to be applied on them. However, the optimization of other non-failure transitions along with the ccf failure transition, provides better results. Here, for $\lambda = 10^{-05}$, we obtain a gain of approximately $6,228$ times as compared to the standard regenerative simulation.

To summarize, there are several possible conclusions that can be drawn from this example where specifically we considered the performance of the CE scheme for a balanced and unbalanced system. For any system (balanced or unbalanced), optimizing all transitions with the CE scheme in the Markovian SPN provides best results due to a higher dimensional optimization (less restrictive optimizer $\tilde{\theta}^{*}_{ce}$). However, as previously discussed, it is only possible in case of moderately sized systems. In large systems it could result in likelihood ratio degeneracy and more statistical noise. Application of the CE scheme on only failure transitions, both one-dimensional and multi-dimensional strategies provide very good results for a balanced system. However, in case of unbalanced systems (especially large scale) where all transitions can not be included for the optimization scheme, and there is a single failure transition that dominates all other failure transitions, the best strategy is to use grouping. Grouping helps to reduce the statistical noise, and consider the combined contributions of all failure transitions, even though the optimizer $\tilde{\theta}^{*}_{ce}$ would be lower dimensional (more restrictive). In case of large systems, this grouping could be useful for each redundant subsystem module, like the 2oo3 considered here.

Another important conclusion that can be drawn from this example is that for a sufficient number of cycles n_j in each pre-simulation stage, the CE Algorithm 3 here is able to capture the contributions of each transition towards the failure set $\{\mathscr{D}\}$. For the balanced system, since both component failure transitions and ccf contribute towards reaching $\{\mathscr{D}\}$, the proposed algorithm optimizes all the failure transitions. However, in an unbalanced system where individual components do not contribute towards failure set (as $\lambda \to 0$), the algorithm is able to capture that IS is not supposed to be applied to the failure transitions of individual components, as *ccf* is the most contributing transition.

4.6 Example 3: Multiple 2oo3 System Modules in Series Configuration

In this example, we consider 3 modules of the 2oo3 system (as discussed in Example 2) connected in a series configuration. It is equivalent to having three SPN models of Appendix B in series. The reachability graph of this Markovian SPN has $20,346,417$ markings (i.e., states of the underlying CTMC). There are 51 discrete places, 51 timed transitions, 9 immediate transitions and 21 failure transitions (18 components and 3 ccf) in the SPN model.

In terms of logistics, each module has its own team of repair personnel considered, with a single repair person for each module. Also, each 2oo3 module in this example has its own spares (to represent different kind of spares needed for different modules) and the spare availability is modeled the same way as in Example 2, for each module here. The system is considered down if any of the 2oo3 module is failed, where any individual module fails if 2 or more components within it are failed. Here again, we analyze the performance of the simulation methods discussed in previous sections with respect to rarity, that is increased as $\lambda \to 0$. The rates of different transitions (exponential distribution of holding times) in the SPN are considered as:

- Module 1 failure rates: Detected (λ), undetected (λ) and common cause (λ^2)
- Module 2 failure rates: Detected (1.3λ), undetected (1.3λ) and common cause ($1.3\lambda^2$)
- Module 3 failure rates: Detected (1.5λ), undetected (1.5λ) and common cause ($1.5\lambda^2$)

- Repair rates : Module 1 (4.0), Module 2 (3.0) and Module 3 (3.5) respective components within each module
- Rate at which spares become available = 5.0
- Rate at which spares become unavailable = 0.1
- Rate at which undetected failures are detected = 10.0
- Rate at which on-site technicians start inspections = 3.0
- Rate of timed inspections = 0.1
- Travel rate for technicians = 1.0

4.6.1 Empirical Results and Interpretations: Three 2oo3 Modules in Series Configuration

The system is considered balanced in this case because within each module, the contribution of the different failure transitions (detected, undetected and ccf) are similar towards the target set $\{\mathscr{D}\}$. Each module has different failure rates and repair rates for the components within it. A standard regenerative simulation (with $N = 10^6$ cycles) yields increasingly inaccurate results as $\lambda \to 0$ (see Table 4.14) and the RE increases rapidly. For $\lambda < 10^{-05}$, we get the undesired empirical value of 0.0 for the point estimates as no failure of the system is recorded.

Table 4.14 Three 2oo3 system: Standard regenerative simulation ($N = 10^6$).

λ	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	RE
10^{-3}	3.6167×10^{-03}	3.6379×10^{-03}	$[3.3861 \times 10^{-03}, 3.8896 \times 10^{-03}]$	1.6501×10^{-08}	181.071	2.9878×10^{-06}	0.03531
10^{-5}	4.0139×10^{-07}	1.1009×10^{-06}	$[-7.5173 \times 10^{-07}, 2.9535 \times 10^{-06}]$	8.9345×10^{-13}	162.744	1.4540×10^{-10}	0.85858
10^{-7}	4.0182×10^{-11}	0.00	[0.00, 0.00]	0.00	–	–	–

Progressive shifting of rarity and choice of parameters in Algorithm 3: For this example, the CE scheme is applied only on the failure transitions, thus forming the subsets $\mathscr{F}$ and $\tilde{\mathscr{F}}$. For each value of λ that forms the original problem to be solved, it is broken down in smaller and easily solvable sub-problems according to the heuristic rule previously defined for progressive shifting of rarity within each problem (in Section 4.3.4). This is simply done by first increasing the failure rates of all the failure transitions (in subset $\mathscr{F}$) to create an unstable system and then gradually decreasing their failure rates until the original problem is reached, which is solved in the final pre-simulation stage. The choice of n_j is the same as that for the Example 2 previously. When the system does not have rare event problem, for example when $\lambda = 10^{-3}$, we use only one pre-simulation stage where only a standard regenerative simulation is performed. There are three different grouping strategies used and we observe the results obtained from these strategies. The results are as discussed in the following text.

Table 4.15 Three 2oo3 systems: Regenerative IS simulation with CE One-D optimization (for each module grouped together) on failure subset only.

				$\mathbf{n_j = 5 \times 10^4 (n_{(j=S)} = 2 \times 10^5), N = 10^6}$					
λ	S	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-3}	1	3.6167×10^{-03}	3.5815×10^{-03}	$[3.4018 \times 10^{-03}, 3.7613 \times 10^{-03}]$	8.4120×10^{-09}	978.926	8.2348×10^{-06}	**0.02561**	**0.36**
10^{-5}	3	4.0139×10^{-07}	4.2456×10^{-07}	$[3.4820 \times 10^{-07}, 5.0093 \times 10^{-07}]$	1.5180×10^{-15}	1226.814	1.8624×10^{-12}	**0.09177**	**78.07**
10^{-7}	4	4.0182×10^{-11}	3.6191×10^{-11}	$[3.2566 \times 10^{-11}, 3.9816 \times 10^{-11}]$*	3.4205×10^{-24}	1133.500	3.8771×10^{-21}	**0.05110**	–
10^{-9}	5	4.0182×10^{-15}	3.8848×10^{-15}	$[3.4749 \times 10^{-15}, 4.2947 \times 10^{-15}]$	4.3742×10^{-32}	1121.263	4.9046×10^{-29}	**0.05384**	–
10^{-11}	6	4.0182×10^{-19}	3.0923×10^{-19}	$[2.7836 \times 10^{-19}, 3.4010 \times 10^{-19}]$*	2.4805×10^{-40}	1205.414	2.9901×10^{-37}	**0.05093**	–

Table 4.16 Three 2oo3 systems: Regenerative IS simulation with CE One-D optimization on failure subset only.

				$\mathbf{n_j = 5 \times 10^4 (n_{(j=S)} = 2 \times 10^5), N = 10^6}$					
λ	S	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-3}	1	3.6167×10^{-03}	3.5677×10^{-03}	$[3.3917 \times 10^{-03}, 3.7436 \times 10^{-03}]$	8.0561×10^{-09}	955.625	7.6986×10^{-06}	**0.02516**	**0.388**
10^{-5}	3	4.0139×10^{-07}	3.6774×10^{-07}	$[3.2528 \times 10^{-07}, 4.1021 \times 10^{-07}]$	4.6946×10^{-16}	1311.586	6.1573×10^{-13}	**0.05892**	**236.15**
10^{-7}	4	4.0182×10^{-11}	3.7572×10^{-11}	$[3.3926 \times 10^{-11}, 4.1217 \times 10^{-11}]$	3.4593×10^{-24}	1178.209	4.0758×10^{-21}	**0.04950**	–
10^{-9}	5	4.0182×10^{-15}	3.9585×10^{-15}	$[3.5798 \times 10^{-15}, 4.3371 \times 10^{-15}]$	3.7315×10^{-32}	1132.600	4.2263×10^{-29}	**0.04880**	–
10^{-11}	6	4.0182×10^{-19}	3.7937×10^{-19}	$[3.4408 \times 10^{-19}, 4.1466 \times 10^{-19}]$	3.2421×10^{-40}	1236.488	4.0088×10^{-37}	**0.04746**	–

Table 4.17 Three 2oo3 systems: Regenerative IS simulation with CE Multi-D optimization on failure subset only.

				$\mathbf{n_j = 5 \times 10^4 (n_{(j=S)} = 2 \times 10^5), N = 10^6}$					
λ	S	Exact Soln. (U)	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-3}	1	3.6167×10^{-03}	3.5296×10^{-03}	$[3.4232 \times 10^{-03}, 3.6360 \times 10^{-03}]$	2.9475×10^{-09}	1113.700	3.2826×10^{-06}	**0.01538**	**0.91**
10^{-5}	3	4.0139×10^{-07}	3.9528×10^{-07}	$[3.5775 \times 10^{-07}, 4.3280 \times 10^{-07}]$	3.6663×10^{-16}	1549.406	5.6806×10^{-13}	**0.04844**	**255.964**
10^{-7}	4	4.0182×10^{-11}	3.7991×10^{-11}	$[3.5298 \times 10^{-11}, 4.0684 \times 10^{-11}]$	1.8878×10^{-24}	1397.318	2.6379×10^{-21}	**0.03617**	–
10^{-9}	5	4.0182×10^{-15}	3.6629×10^{-15}	$[3.4685 \times 10^{-15}, 3.8573 \times 10^{-15}]$*	9.8357×10^{-33}	1399.208	1.3762×10^{-29}	**0.02708**	–
10^{-11}	6	4.0182×10^{-19}	3.9090×10^{-19}	$[3.6810 \times 10^{-19}, 4.1370 \times 10^{-19}]$	1.3529×10^{-40}	1409.951	1.9076×10^{-37}	**0.02976**	–

*Exact value slightly out of 95% CI bounds: Since we are estimating a 95% CI, there is always a 5% probability of the exact solution U being out of the estimated bounds. Another possible reason could be that the CE scheme is a stochastic approximation method, so statistical noise can result in such observations. Also, using a different seed, we found the 95% CI bounds covering the exact solution. Additionally, using more number of pre-simulation stages instead of changing the seed, we found the RE values to be approximately similar while the 95% CI bounds covering the exact solution.

Strategy 1 (Three 2oo3 in series): Grouped one-dimensional CE optimization for each module

In this strategy, failure transitions of each module are grouped together and we use the proposed CE Algorithm 3 to find a common value of CE optimized IS rates for all failure transitions within each module (as represented by Equation 4.20). The results are presented in Table 4.15. With this strategy, the empirical results show that we obtain a BRE property (RE ≈ 0.05) as $\lambda \to 0$. Also, in terms of gain, we obtain a gain of ≈ 78 times as compared to the standard simulation (see Table 4.14), for $\lambda = 10^{-5}$. The progressive shifting of rarity for each λ in Algorithm 3, thus helps in obtaining large variance reduction.

Strategy 2 (Three 2oo3 in series): One-dimensional CE optimization for all failure transitions grouped together

The second strategy groups all failure transitions of all three modules in a single group. It thus attempts to use the CE Algorithm 3 to find a common CE optimized IS rate for all of the failure transitions (i.e., a single IS rate value). The results are shown in Table 4.16. Here also the RE is bounded (≈ 0.05) as $\lambda \to 0$ and we obtain an even larger variance reduction compared to standard simulation. In this case, the gain is ≈ 236 times for $\lambda = 10^{-5}$, when compared to the standard simulation.

Strategy 3 (Three 2oo3 in series): Multi-dimensional CE optimization for each failure transition

Finally, we apply the CE Algorithm 3 by optimizing each failure transition separately (as previously shown by Equation 4.16). In this case, optimizing each failure transition separately gives better results and the RE is bounded (at ≈ 0.03), as shown by the results in Table 4.17. The gain is also maximum (≈ 256 times) in this case, for $\lambda = 10^{-5}$ and obviously due to the BRE property, it will be increasingly higher as $\lambda \to 0$.

Thus, from this example of a large system, we observe that the CE scheme provides CE optimized IS rates. When these IS rates are used in the main regenerative IS simulations, we finally obtain estimates of U with the BRE property (as $\lambda \to 0$), in terms of empirical values. It is to be noted that in terms of gain using the CE schemes, for very high values of $\lambda = 10^{-03}$, we do not see enough gain (see Table 4.15-4.17) as compared to standard simulation. The gain is lesser than one because even though the proposed CE Algorithm 3 reduces the $\hat{\sigma}^2$ and the RE, the algorithm is based on pre-simulations that uses more computation time as compared to standard simulation. However, since the RE is bounded in all the three strategies used for the CE scheme (in Tables 4.15-4.17), the gain will obviously be increasingly higher as $\lambda \to 0$.

Figure 4.5 also shows the evolution of the empirical values of the RE obtained when rarity is increased (i.e., $\lambda \to 0$). In Figure 4.5, it is observable that as $\lambda \to 0$, the RE from the three strategies used for CE optimization (grouped one-dimensional, one-dimensional and multi-dimensional optimization of failure transitions) are bounded. The slight oscillations

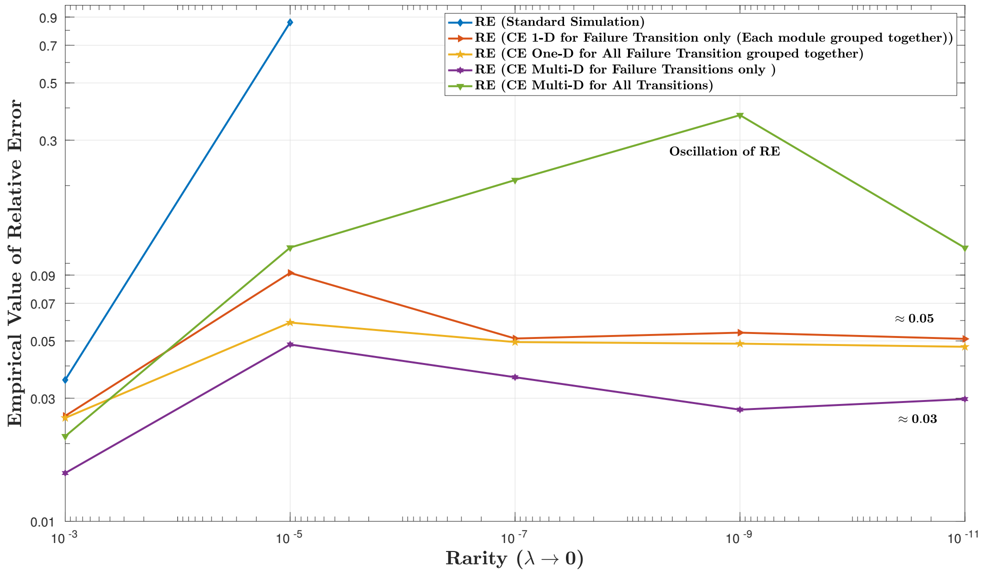

Figure 4.5 Three 2oo3 System: Evolution of estimated RE as $\lambda \to 0$.

are due to stochasticity. Application of the CE scheme on all transitions (except immediate transitions) resulted in more statistical noise in the given number of n_j used at the specific stages. Since the CE scheme is based on sample average approximation, increasing the number of cycles in the pre-simulation should be able to overcome the issue of statistical noise. However, the possibility of likelihood degeneracy [31, 97] needs to be also checked and avoided in such cases, as previously discussed.

4.6.2 A Large Example of 4 Modules of 2oo3 Subsystems with Logistics

In the previous examples, we showed the efficiency of the CE Algorithm 3 for various cases. However, in all the previously presented examples, an exact solution of the underlying CTMC was obtainable (via SPNP). Our goal is to also show the effectiveness of the proposed algorithm on very large systems, where the exact solution is not obtainable.

For the above purpose, we now consider 4 modules of the 2oo3 subsystem connected in a series configuration. The first three modules have the same specifics (in terms of rates and logistical aspects) as presented for the previous example of three 2oo3 modules in series. We add a fourth module with the same SPN, but with different failure and repair rates. These rates are as given by:

- Module 4 failure rates: Detected (0.5λ), undetected (0.5λ) and common cause ($0.5\lambda^2$)
- Module 4 repair rate : 5.0

The SPN of this model has $5,554,571,841$ markings in the reachability graph. For such a large system, we were unable to obtain the exact numerical solution (in SPNP). The model has 28 failure transitions, 7 for each of the 4 modules. We use the CE Algorithm 3, to find the optimal IS rates for these failure transitions of the Markovian SPN model here.

First, we performed a standard regenerative simulation (with $N = 10^6$ cycles) and as expected, as the rarity increases (i.e., $\lambda \to 0$), we start obtaining increasingly erroneous estimations (see Table 4.18). For $\lambda < 10^{-5}$, we obtain the useless 0.0 empirical values for the point estimates as no failure event is recorded. [1]

Table 4.18 Four 2oo3 systems: Standard regenerative simulation ($N = 10^6$).

λ	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	RE
10^{-3}	3.6100×10^{-03}	$[3.3571 \times 10^{-03}, 3.8630 \times 10^{-03}]$	1.6657×10^{-08}	365.76	6.0926×10^{-06}	0.03575
10^{-5}	$6.8052 \times 10^{-08}*$	$[-6.5330 \times 10^{-08}, 2.0143 \times 10^{-07}]*$	$4.6310 \times 10^{-15}*$	306.85	$1.4210 \times 10^{-12}*$	1.00
10^{-7}	0.00	$[0.00, 0.00]$	0.00	–	–	–

To show the effectiveness of our proposed Algorithm 3, we consider to apply IS on the 28 failure transitions using three different strategies: a grouped change of measure for each module, a one-dimensional change of measure with a common value of IS rate, and a multi-dimensional optimization for each transition individually. The progressive shifting of rarity is done in the same way as explained for the previous examples. Since here we consider a very large system, we use a higher number of pre-simulation cycles in the Algorithm 3 to reduce the statistical noise. Here, we use $n_j = 10^5$ cycles for all pre-simulation stages, while using $n_{j=S} = 4 \times 10^5$ cycles for the final pre-simulation stage, where the original problem is solved. The final simulation uses $N = 10^6$ cycles. Also, the number of stages S are increased as the rarity of the original problem increases, to break down rarer problems into higher number of sub-problems. .

Table 4.19 Four 2oo3 systems: Regenerative IS simulation with CE One-D optimization (for each module grouped together) on failure subset only.

			$\mathbf{n_j = 1 \times 10^5 (n_{(j=S)} = 4 \times 10^5), N = 10^6}$					
λ	S	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-3}	1	3.7473×10^{-03}	$[3.5432 \times 10^{-03}, 3.9515 \times 10^{-03}]$	1.0845×10^{-08}	2303.27	2.4979×10^{-05}	**0.02779**	**0.24**
10^{-5}	3	3.7299×10^{-07}	$[3.3612 \times 10^{-07}, 4.0985 \times 10^{-07}]$	3.5385×10^{-16}	4158.76	1.4716×10^{-12}	**0.05043**	**10.51**
10^{-7}	4	3.9732×10^{-11}	$[3.3754 \times 10^{-11}, 4.5710 \times 10^{-11}]$	9.3014×10^{-24}	3700.87	3.4423×10^{-20}	**0.07676**	–
10^{-9}	5	4.2733×10^{-15}	$[3.4875 \times 10^{-15}, 5.0592 \times 10^{-15}]$	1.6076×10^{-31}	4134.88	6.6473×10^{-28}	**0.09383**	–
10^{-11}	6	3.7774×10^{-19}	$[3.4349 \times 10^{-19}, 4.1200 \times 10^{-19}]$	3.0543×10^{-40}	3183.56	9.7235×10^{-37}	**0.04627**	–

In the first case, grouping failure transitions of each module in a single group, we obtain a large variance reduction as $\lambda \to 0$. From the empirical results presented in Table 4.19, we observe that the RE is bounded, approximately between 0.05 to 0.09 estimated values (oscillation due to stochasticity). In terms of gain compared to the standard regenerative

[1] *The empirical values obtained in Table 4.18 for $\lambda = 10^{-5}$ are erroneous due to stochasticity. Using a different seed, we obtained empirical values of 0.0 with $N = 10^6$ cycles. For this purpose, we increased that number of cycles $N = 10^7$ to obtain a slightly better estimator. The empirical values obtained using $N = 10^7$, larger sample size were: $\hat{U} = 1.0968 \times 10^{-7}$, 95%CI as $[-2.7916 \times 10^{-08}, 2.4728 \times 10^{-07}]$, $\hat{\sigma}^2 = 4.9286 \times 10^{-15}$, computation time (s) = 3138.030, $\hat{\sigma}^2_{wn} = 1.5466 \times 10^{-11}$, and RE= 0.64006. We use these values for any comparisons (of gain or work-normalized variance) with the CE method later on for this example, when $\lambda = 10^{-5}$.

simulation, we observe for $\lambda = 10^{-5}$, a gain of approximately 10 times using this CE algorithm with this strategy.

In the second case, we group all the 28 failure transitions in a single group to obtain a common value of IS rates for all of them using the proposed CE algorithm. Results presented in Table 4.20, show the RE is bounded ($\approx$ 0.07 to 0.09) as $\lambda \to 0$. In this case also, we observe that a gain of ≈ 3 times for $\lambda = 10^{-5}$, when we compared it to the result from the standard method. Here, the gain is lesser than the first strategy as a one-dimensional change of measure is very low dimensional optimization (highly restrictive) and thus variance is reduced to a lesser extent.

Table 4.20 Four 2oo3 systems: Regenerative IS simulation with CE One-D optimization on failure subset only.

			$\mathbf{n_j = 5 \times 10^4 (n_{(j=S)} = 2 \times 10^5),\ N = 10^6}$					
λ	S	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-3}	1	3.6219×10^{-03}	$[3.3929 \times 10^{-03}, 3.8510 \times 10^{-03}]$	1.3658×10^{-08}	2187.18	2.9872×10^{-05}	**0.03227**	**0.20**
10^{-5}	3	3.6723×10^{-07}	$[2.9903 \times 10^{-07}, 4.3542 \times 10^{-07}]$	1.2106×10^{-15}	4034.22	4.8836×10^{-12}	**0.09475**	**3.17**
10^{-7}	4	4.0185×10^{-11}	$[3.3816 \times 10^{-11}, 4.6555 \times 10^{-11}]$	1.0561×10^{-23}	3247.76	3.4301×10^{-20}	**0.08087**	–
10^{-9}	5	3.8717×10^{-15}	$[3.3286 \times 10^{-15}, 4.4149 \times 10^{-15}]$	7.6790×10^{-32}	3277.80	2.5170×10^{-28}	**0.07157**	–
10^{-11}	6	4.6983×10^{-19}	$[3.8664 \times 10^{-19}, 5.5303 \times 10^{-19}]$	1.8018×10^{-39}	3320.71	5.9832×10^{-36}	**0.09035**	–

In the final strategy, we use the proposed CE algorithm to obtain the CE optimized IS rates for each of the 28 failure transitions individually. In this case (see results in Table 4.21), we observe the lowest values of the RE, also bounded (between $\approx$ 0.04 to 0.06) when $\lambda \to 0$. However, this multi-dimensional optimization uses a higher computation time when comparing to the previous two grouping strategies, but with a higher dimensional (less restrictive) optimizer $\tilde{\theta}^*_{ce}$. In terms of gain, for $\lambda = 10^{-5}$, this multi-dimensional strategy using the proposed CE Algorithm 3 is ≈ 4 times better than the standard method. Again, for higher values of $\lambda = 10^{-3}$, the system does not suffer from rare event issues. Due to this we obtain a gain less than one using the proposed CE algorithm, due to a higher computation time even if it reduces the variance slightly.

Table 4.21 Four 2oo3 systems: Regenerative IS simulation with CE Multi-D optimization on failure subset only.

			$\mathbf{n_j = 1 \times 10^5 (n_{(j=S)} = 5 \times 10^5),\ N = 10^6}$					
λ	S	Point Est. ($\hat{U}$)	95% CI	Variance Est. ($\hat{\sigma}^2$)	Time(s)	$\hat{\sigma}^2_{wn}$	**RE**	**Gain**
10^{-3}	1	3.7679×10^{-03}	$[3.6366 \times 10^{-03}, 3.8992 \times 10^{-03}]$	4.4883×10^{-09}	2569.06	1.1531×10^{-05}	**0.01778**	**0.53**
10^{-5}	3	4.3421×10^{-07}	$[3.8254 \times 10^{-07}, 4.8589 \times 10^{-07}]$	6.9515×10^{-16}	5308.18	3.6900×10^{-12}	**0.06072**	**4.19**
10^{-7}	4	3.9577×10^{-11}	$[3.4587 \times 10^{-11}, 4.4566 \times 10^{-11}]$	6.4798×10^{-24}	4176.15	2.7061×10^{-20}	**0.06432**	–
10^{-9}	5	3.7599×10^{-15}	$[3.4590 \times 10^{-15}, 4.0608 \times 10^{-15}]$	2.3566×10^{-32}	4169.59	9.8261×10^{-29}	**0.04083**	–
10^{-11}	6	4.3979×10^{-19}	$[3.9374 \times 10^{-19}, 4.8584 \times 10^{-19}]$	5.5201×10^{-40}	3922.33	2.1652×10^{-36}	**0.05342**	–

4.7 Conclusions from the Chapter

The motivation of the work (as discussed in Section 4.1) was to estimate the steady-state unavailability U of HRMS that also include complex logistics. We used SPNs to conveniently represent these HRMS in the form of Markovian SPNs (from which the underlying CTMCs can be extracted). The main objective was to develop a multi-level pre-simulation scheme based on minimizing the CE distance between the zero-variance IS density and the IS change of measure used. The approach provided CE optimized IS rates (within the same parametric family) that could we used in main simulations.

We proposed an efficient Algorithm 3, that exploited the regenerative structure of the underlying CTMCs of the Markovian SPNs. Also, we proposed the novel idea of progressively shifting the rarity within each problem itself, by changing failure rates (in some cases repair rates also) to create a sequence of easily solvable sub-problems with increasing rarity until reaching the original problem.

Another idea applied in the proposed algorithm was to use grouping for the IS change of measure. Grouping is helpful in the practical application of the algorithm to reduce the statistical noise and consider combined effect of transitions in a group towards the system failure. However, grouping also provides a lower dimensional (i.e., more restrictive) CE optimizer $\tilde{\theta}^*_{ce}$. A multi-dimensional strategy is better in terms of a higher dimensional (less restrictive) optimizer $\tilde{\theta}^*_{ce}$, but it also takes a higher computation time in the algorithm. We do not yet have a robust heuristic to select the best possible grouping strategy in general problems.

Another applicability issue that the proposed CE scheme solves in the examples presented is the choice of the IS vector for the first pre-simulation stage. In the proposed algorithm, this is done based on the number of pre-simulations stages (S), where in the first stage we perform only a standard regenerative simulation for a system which is unstable (i.e., non-rare system failures). This approach only attempts to capture the contribution of the respective transitions of interest as per the update equation used in the Algorithm 3 (the likelihood ratio being one).

Finally, we tried four different examples (with increasing size of the models) where complex logistics was also considered. The proposed CE scheme provided accurate results while also adhering to the desired BRE property (in terms of empirical values) as the rarity of a problem increased. The gain (in terms of work-normalized variance) when compared to standard regenerative simulations was increasingly higher, due to the BRE property, as rarity of system failure increased. In practical applications, for analysis of large scale and highly reliable Markovian SPNs, the proposed CE scheme can be very helpful in performing automated IS simulations with only few inputs required.

Chapter 5

Conclusions

The objective of the current work and the entire dissertation was to propose efficient algorithms for use of IS methods that can accurately estimate RAM metrics of interest, especially in case of highly reliable systems where system failures are rare events. The work has focused on two RAM metrics: first, unreliability (or reliability) of static networks; second, steady-state unavailability (or contrarily the availability) of dynamic systems (under Markovian assumptions). These RAM metrics are helpful in determining the LCC of systems and thus can help rail system suppliers, such as Alstom, to make well-informed decisions and policies (e.g., maintenance, spares availability and inspection policies, to name a few). Usually, when standard MC simulations become impractical due to rare event issues, IS techniques are helpful. IS techniques help to accelerate these simulations (the notion of obtaining less variance in same computational budget) while providing accurate and efficient estimation of the metrics. However, since the use of IS techniques require to know specifically a good change of measure that reduces the variance of the final estimator, the objective of the current work tapered towards algorithms that can approximate or find these good (or even optimal) change of measures to apply IS in real problems.

5.1 Conclusions of the thesis

The first work in this dissertation focused on approximating the zero-variance IS change of measure scheme, where the networks (or systems) are static in nature and nodes in a graph model are the failing components. We proposed an efficient algorithm [11] in this context (in Chapter 3) that adapted the work on link failure case of static networks [29]. The usefulness of the proposed algorithm was shown for different networks and a real case of a DCS subsystem of Alstom was also studied. The second work as presented in Chapter 4, focused on using a multi-level CE scheme to find optimal IS rates of failure transitions in Markovian SPN models. We developed this idea for the underlying CTMCs of Markovian SPNs and showed application on various examples. In both the works, in terms of measures of accuracy, the results obtained from the respective algorithms showed BRE properties (for some static networks even VRE property) asymptotically (i.e., when the event of interest became rarer). In the following paragraphs, we describe the main conclusions of the current work and the algorithms therein.

An approximate zero-variance IS algorithm for static networks reliability estimation with node failures: This Algorithm 2, namely the *Approximate Zero-Variance IS using Ford-Fulkerson adapted Algorithm 1*, considered node failures in static networks and sequentially sampled the nodes. The main idea behind the algorithm was to find mincuts with maximal probabilities between the source-and-terminal nodes and sequentially sample the state of each node in those paths. The RAM metric of interest here was the unreliability between the source-and-terminal nodes of a static network. The unreliability metric could also be considered as the steady-state unavailability of such systems. From the results shown in Chapter 3 for various examples, we obtained BRE as well as VRE properties in an asymptotic regime (in terms of rarity). Taking into account a BRE (or VRE) property, the gain with respect to a standard simulation increased as the event of interest became rarer. The practical aspect of this work was to compute the reliability of systems like the DCS, where communication between a set of nodes is critically important for real passenger rail systems. The proposed algorithm was able to estimate the reliability of a DCS network, with a BRE property (as probability of failure became rarer) in both homogeneous and heterogeneous cases.

Availability estimation of Markovian reliability systems with logistics using CE: The Algorithm 3 proposed here, namely *Cross Entropy Algorithm for Markovian SPNs*, utilized the idea of minimizing the CE distance between the zero-variance IS density and the IS density used as a change of measure. The novelty of the proposed Algorithm 3 lied in the breakdown of difficult (i.e., rarer) problems into a set of easily solvable sub-problems in a multi-level CE scheme, where rarity was slowly increased in each sub-problem that is being solved. The modeling of HRMS in the form of Markovian SPNs conveniently represented the underlying CTMCs and made it relatively easier to model systems, as is the general reasoning behind the use of SPNs. Also, the proposed CE algorithm was able to find the optimal IS rates, specifically for failure transitions, in Markovian SPN models. Results from various examples that tried to mimic certain subsystems of a real passenger rail system (while also including complex logistics) showed BRE property being observed. Consequently, a considerable variance reduction and gain is obtained in the results.

The aforementioned work attempted to address the problems of application of IS methods for static and dynamic systems (under Markovian assumptions). However, as any scientific research is an evolutionary stride, there is always a room for improvement. Keeping in mind the future possibilities of the current work, we propose certain ideas that are worth considering.

5.2 Perspectives: Directions for Future work

Previously in Chapter 3 where static network reliability was estimated using approximate zero-variance IS scheme, we discussed that in our examples, certain enumerations of nodes provided slightly lower values of RE. As the goal is to estimate RAM metrics with lowest possible values of variance and RE as the event of interest becomes rarer, it could be a valuable contribution to co-relate the RE with the ordering of nodes in the graph models. Another aspect is the trade-off between accuracy and computation time. The proposed Algorithm 2 provided highly accurate results but at each step it computed two mincuts with maximal probability, resulting in a significant more computation time (bounded when $\varepsilon \to 0$) than a standard MC simulation. Even though the proposed method, due to BRE and VRE properties, is always better (in terms of gain) than a standard MC simulation, it is also a trade-off between choosing a more accurate estimation or a faster estimation with huge variance. Best case obviously being the accurate estimation with lowest possible computation time. In order to reduce the issue of computational effort, graph reduction techniques could prove to be very useful.

The second method for dynamic systems discussed in Chapter 4 also has certain room for improvements. In the Algorithm 3, we increased the number of pre-simulation stages (S) as the original problem became rarer. This was done to breakdown rarer problems into higher number of easier sub-problems, where less rarer problems required lower S and more rarer problems required a higher S. Finding a robust heuristic to be able to automatically find the required minimum number of pre-simulation stages S for a particular original problem could be a useful idea for the future work. This would also make it possible to optimize the computational effort during the pre-simulation stages.

The above ideas for future possibilities of research provide a very general direction for improvement of the proposed methods here. However, there are two following main ideas that we consider are specifically useful for future work, and can improve the methods for real applications to analyze passenger rail systems.

5.2.1 Application of CE to Non-Markovian SPN transitions

In the Algorithm 3 proposed in Chapter 4 for CE optimization of transitions of Markovian SPNs, we exploited the regenerative structure of the underlying CTMCs. Due to the regenerative property, each cycle provided iid samples. However, in Non-Markovian SPNs, the transitions are not limited to exponential distributions of holding times only. There are several distributions like Weibull, log-normal, triangular etc., that can be used in SPNs. This, would result in Non-Markovian SPNs. In Non-Markovian systems, the regenerative structure is lost and the samples from simulation are not iid. In this context, the proposed CE method could be useful in finding the optimal IS change of measures for transitions with general distributions by exploring further.

Steady-state measures, as discussed in Chapter 4, are usually represented by a ratio (expected downtime and expected length of a cycle in Markovian systems) [36]. In non-Markovian systems, this ratio representation for steady-state quantities can be in terms of "A-cycles", where "A" is a set of some states (e.g., all components operational) [17, 36]. However, these A-cycles are not iid and hence application of IS and estimation of variance becomes complicated [36]. In literature, a "splitting" technique (not to be confused with splitting technique for rare events simulation discussed previously) is suggested where IS is applied for estimating the numerator (expected downtime in a A-cycle). The denominator (expected length of a A-cycle) is estimated under the original measure [17, 36]. This method is similar to Measure Specific Dynamic Importance Sampling (MSDIS). For application of IS to estimate the denominator, the system is allowed to reach the steady-state, discarding the initial samples (e.g., D_i being downtime in a A-cycle) [36]. The variance in such cases can be estimated by the method of Batch Means [36]. Also, CE optimization has been used for non-exponential distributions like the Weibull distribution for parameter updating in IS context [31]. Integrating and obtaining a CE update equation for optimizing parameters of general distributions in Non-Markovian SPNs, while using A-cycles could be an important direction for future research. This could also make it very useful for RAM practitioners to model and analyze real passenger rail systems with less assumptions.

5.2.2 Description of Failure Modes

When modeling systems with inherent components in SPNs, it is easier to model when specific components can be considered as working, failed or working in degraded mode. However, the failure mode for the entire system needs to be specified in SPNs (in general for any Markov modeling approach too). The approximate zero-variance IS method in Chapter 3 used mincuts with maximal probabilities to estimate reliability efficiently. Extracting those mincuts in a dynamic system and defining the failure modes of SPN models based on the components included in those mincuts having maximal probability of failure, could prove useful in easier modeling of SPNs for the practitioners.

Publications from the thesis

In the current work, we proposed mainly two algorithms [11, 30] for estimation of reliability metrics of highly reliable (static and dynamic) systems in context of rare events simulations. For static network reliability estimation, the metric of interest was the probability of the source-and-terminal nodes being disconnected. For dynamic systems, we included complex logistics, while modeling the systems as Markovian SPNs, and using a Cross-Entropy optimization scheme to find optimal IS rates of the failure transitions of the SPN. The empirical results from the two algorithms in their respective context (static and dynamic) showed a large variance reduction and also Bounded Relative Error (in some cases Vanishing Relative Error) property too. Results obtained from the current work are also published as shown below.

Peer-reviewed International Conferences

- [11] Ajit Rai, Rene C. Valenzuela, Bruno Tuffin, Gerardo Rubino, and Pierre Dersin. "Approximate zero-variance importance sampling for static network reliability estimation with node failures and application to rail systems." In Proceedings of the 2016 Winter Simulation Conference (WSC), Washington D.C., USA., pp. 3201-3212., 2016.

- [30] Ajit Rai, Bruno Tuffin, Rene C. Valenzuela, Gerardo Rubino, and Pierre Dersin. "Availability estimation of Markovian reliability systems with logistics via cross-entropy." In Proceedings of the 13th International Conference in Monte Carlo & Quasi-Monte Carlo Methods in Scientific Computing (MCQMC), Rennes, France, July 2018 (Abstract presented).

Bibliography

[1] P. Dersin and R.C. Valenzuela. Application of non-Markovian stochastic Petri Nets to the modeling of rail system maintenance and availability. In *Proceedings of the 2012 Winter Simulation Conference (WSC)*, pages 1–12. IEEE, 2012.

[2] IEC. International Standard: IEC 60050-192, CEI 60050-192, Edition 1.0, 2015. Personal Webport for Alstom Transport.

[3] G.M.P. Simões. RAMS analysis of railway track infrastructure. Master's thesis, Instituto Superior Técnico-Universidade Técnica de Lisboa, Portugal, 2008.

[4] M. G. Park. *RAMS Management of Railway Systems*. PhD thesis, College of Engineering & Physical Sciences, University of Birmingham, Birmingham, UK, 2014.

[5] F. Corvaro, G. Giacchetta, B. Marchetti, and M. Recanati. Reliability, Availability, Maintainability (RAM) study, on reciprocating compressors API 618. *Petroleum*, 3(2):266–272, 2017.

[6] J.P. Williams. Predicting process systems. *Hydrocarbon Engineering*, 6(7):29–33, 2001.

[7] J.A. Buzacott. Markov approach to finding failure times of repairable systems. *IEEE Transactions on Reliability*, 19(4):128–134, 1970.

[8] V.G. Kulkarni. *Introduction to Modeling and Analysis of Stochastic Systems*. Springer Texts in Statistics. Springer, New York, 2012.

[9] R.K. Sharma and S. Kumar. Performance modeling in critical engineering systems using RAM Analysis. *Reliability Engineering & System Safety*, 93(6):913–919, 2008.

[10] A. Goyal, P. Shahabuddin, P. Heidelberger, V.F. Nicola, and P.W. Glynn. A unified framework for simulating Markovian models of highly dependable systems. *IEEE Transactions on Computers*, 41(1):36–51, 1992.

[11] A. Rai, R.C. Valenzuela, B. Tuffin, G. Rubino, and P. Dersin. Approximate zero-variance importance sampling for static network reliability estimation with node failures and application to rail systems. In *Proceedings of the 2016 Winter Simulation Conference (WSC)*, pages 3201–3212. IEEE, 2016.

[12] R. David and H. Alla. Petri Nets for modeling of dynamic systems: A survey. *Automatica*, 30(2):175–202, 1994.

[13] B. Tuffin. *La simulation de Monte Carlo*. Hermès, 2010.

[14] G. Rubino and B. Tuffin. Introduction to rare event simulation. In G. Rubino and B. Tuffin, editors, *Rare Event Simulation Using Monte Carlo Methods*, chapter 1, pages 1–13. John Wiley & Sons, 2009.

[15] P.H. Borcherds. Importance sampling: an illustrative introduction. *European Journal of Physics*, 21(5):405–411, 2000.

[16] M. Denny. Introduction to importance sampling in rare-event simulations. *European Journal of Physics*, 22(4):403–411, 2001.

[17] P. Heidelberger. Fast simulation of rare events in queueing and reliability models. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 5(1):43–85, 1995.

[18] P.-T. de Boer and D.P. Kroese. Estimating buffer overflows in three stages using cross-entropy. In *Proceedings of the 2002 Winter Simulation Conference (WSC)*, volume 1, pages 301–309. IEEE, 2002.

[19] Areva and EDF. EDF pre-construction safety report: Issue 06. UK EPR™ GDA submission. Technical report, Areva and EDF, 2012. Available at (Accessed: December 2017) http://www.epr-reactor.co.uk/ssmod/liblocal/docs/PCSR/Chapter%2015%20-%20Probabilistic%20Safety%20Analysis/Sub-Chapter%2015.7%20-%20PSA%20Discussion%20and%20Conclusions.pdf.

[20] P. L'Ecuyer, V. Demers, and B. Tuffin. Splitting for rare-event simulation. In *Proceedings of the 2006 Winter Simulation Conference (WSC)*, pages 137–148. IEEE, 2006.

[21] P. L'Ecuyer, F. Le Gland, P. Lezaud, and B. Tuffin. Splitting techniques. In G. Rubino and B. Tuffin, editors, *Rare Event Simulation Using Monte Carlo Methods*, chapter 3, pages 39–62. John Wiley & Sons, 2009.

[22] P. L'Ecuyer, M. Mandjes, and B. Tuffin. Importance sampling in rare event simulation. In G. Rubino and B. Tuffin, editors, *Rare Event Simulation Using Monte Carlo Methods*, chapter 2, pages 17–38. John Wiley & Sons, 2009.

[23] H. Kahn. Stochastic (Monte Carlo) attenuation analysis, RAND Corporation Report R-163. Technical report, The RAND Corporation, 1949.

[24] N. Metropolis and S. Ulam. The Monte Carlo method. *Journal of the American Statistical Association*, 44(247):335–341, 1949.

[25] P.W. Glynn and D.L. Iglehart. Importance sampling for stochastic simulations. *Management Science*, 35(11):1367–1392, 1989.

[26] A. Goyal, P. Heidelberger, and P. Shahabuddin. Measure specific dynamic importance sampling for availability simulations. In *Proceedings of the 1987 Winter Simulation Conference (WSC)*, pages 351–357. ACM, 1987.

[27] H. Cancela, M. El Khadiri, and G. Rubino. Rare Event Analysis by Monte Carlo Techniques in Static Models. In G. Rubino and B. Tuffin, editors, *Rare Event Simulation Using Monte Carlo Methods*, chapter 7, pages 145–170. John Wiley & Sons, 2009.

[28] I. Gertsbakh, Y. Shpungin, and R. Vaisman. Network Reliability Monte Carlo with Nodes Subject to Failure. *International Journal of Performability Engineering*, 10(2):163–172, 2014.

[29] P. L'Ecuyer, G. Rubino, S. Saggadi, and B. Tuffin. Approximate zero-variance importance sampling for static network reliability estimation. *IEEE Transactions on Reliability*, 60(3):590–604, 2011.

[30] A. Rai, B. Tuffin, R.C. Valenzuela, G. Rubino, and P. Dersin. Availability estimation of Markovian reliability systems with logistics via cross-entropy. In *Proceedings of the 13th International Conference in Monte Carlo & Quasi-Monte Carlo Methods in Scientific Computing (MCQMC)*, Rennes, France, July 2018. (Abstract accepted).

[31] R.Y. Rubinstein and D.P. Kroese. *The Cross-Entropy Method: A Unified Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*. Springer Science & Business Media, New York, USA, 2004.

[32] J.C.C. Chan, P.W. Glynn, and D.P. Kroese. A comparison of cross-entropy and variance minimization strategies. *Journal of Applied Probability*, 48(A):183–194, 2011.

[33] W. L. Smith. Regenerative stochastic processes. In *Proceedings of the Royal Society of London (Series A): Mathematical and Physical Sciences*, volume 232, pages 6–31. The Royal Society, 1955.

[34] K.S. Trivedi. *SPNP User's Manual Version 6.0.* Center for Advanced Computing and Communication (CACC), Department of Electrical and Computer Engineering, Duke University, 1999.

[35] P.W. Glynn, G. Rubino, and B. Tuffin. Robustness properties and confidence interval reliability issues. In G. Rubino and B. Tuffin, editors, *Rare Event Simulation Using Monte Carlo Methods*, chapter 4, pages 63–84. John Wiley & Sons, 2009.

[36] V.F. Nicola, P. Shahabuddin, P. Heidelberger, and P.W. Glynn. Fast simulation of steady-state availability in non-Markovian highly dependable systems. In *The Twenty-Third International Symposium on Fault-Tolerant Computing. FTCS-23. Digest of Papers.*, pages 38–47. IEEE, 1993.

[37] R.Y. Rubinstein and D.P. Kroese. *Simulation and the Monte Carlo Method, Third Edition*, volume 10 of *Wiley Series in Probability and Statistics*. John Wiley & Sons, 2017.

[38] M.J.J. Garvels. *The Splitting Method in Rare Event Simulation*. PhD thesis, Faculty of Mathematical Science, University of Twente, Enschede, The Netherlands, 2000.

[39] F. Cérou and A. Guyader. Adaptive multilevel splitting for rare event analysis. *Stochastic Analysis and Applications*, 25(2):417–443, 2007.

[40] P. L'Ecuyer, V. Demers, and B. Tuffin. Rare events, splitting, and quasi-Monte Carlo. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 17(2):9, 2007.

[41] F. LeGland and N. Oudjane. A sequential particle algorithm that keeps the particle system alive. In *Proceedings of 13th European Signal Processing Conference, 2005*, pages 1–4. IEEE, 2005.

[42] M. Villén-Altamirano and J. Villén-Altamirano. RESTART: A method for accelerating rare event simulations. In J.W. Cohen and C.D. Pack, editors, *The 13th International Teletraffic Congress, Queuing Performance and Control in ATM*, pages 71–76, 1991.

[43] J.M. Hammersley and D.C. Handscomb. *Monte Carlo methods*. Methuen, London, 1964.

[44] J.H. Halton. A retrospective and prospective survey of the Monte Carlo method. *SIAM Review*, 12(1):1–63, 1970.

[45] M.H. Kalos and P.A. Whitlock. *Monte Carlo methods: Basics*, volume 1. John Wiley & Sons, 1986.

[46] M. Mahmoodian and A. Alani. Time-dependent reliability analysis of corrosion affected structures. In S. Kadry and A. El Hami, editors, *Numerical Methods for Reliability and Safety Assessment (Multiscale and Multiphysics Systems)*, chapter Part III (17), pages 459–458. Springer International Publishing, 2015.

[47] M.K. Nakayama. A characterization of the simple failure-biasing method for simulations of highly reliable Markovian systems. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 4(1):52–88, 1994.

[48] S. Parekh and J. Walrand. A quick simulation method for excessive backlogs in networks of queues. *IEEE Transactions on Automatic Control*, 34(1):54–66, 1989.

[49] P.-T. de Boer. Analysis of state-independent importance-sampling measures for the two-node tandem queue. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 16(3):225–250, 2006.

[50] P.-T. de Boer and V.F. Nicola. Adaptive state-dependent importance sampling simulation of Markovian queueing networks. *Transactions on Emerging Telecommunications Technologies*, 13(4):303–315, 2002.

[51] P.-T. de Boer. *Analysis and Efficient Simulation of Queueing Models of Telecommunication Systems*. PhD thesis, University of Twente, The Netherlands, 2000.

[52] P.-T. de Boer. Rare-event simulation of non-Markovian queueing networks using a state-dependent change of measure determined using cross-entropy. *Annals of Operations Research*, 134(1):69–100, 2005.

[53] T.S. Zaburnenko. *Efficient Heuristics for Simulating Rare Events in Queuing Networks*. PhD thesis, University of Twente, Enschede, The Netherlands, 2008.

[54] F. den Hollander. *Large Deviations*. Fields Institute Monographs (Book 14). American Mathematical Society, 2008.

[55] S. Juneja and P. Shahabuddin. Rare-Event Simulation Techniques: An Introduction and Recent Advances. In S.G. Henderson and B.L. Nelson, editors, *Simulation*, volume 13 of *Handbooks in Operations Research and Management Science*, chapter 11, pages 291–350. Elsevier, 2006.

[56] J.A. Bucklew. *Introduction to Rare Event Simulation*. Springer Series in Statistics. Springer Science & Business Media, 2004.

[57] S. Asmussen. Large deviations in rare events simulation: Examples, counterexamples and alternatives. In K.-T. Fang, H. Niederreiter, and F.J. Hickernell, editors, *Monte Carlo and Quasi-Monte Carlo Methods 2000*, pages 1–9. Springer, Berlin, 2002.

[58] J.T. Lewis and R. Russell. An introduction to large deviations for teletraffic engineers. *Dublin Institute for Advanced Studies*, pages 1–45, 1997.

[59] S.R.S. Varadhan. *Large Deviations and Applications (CBMS-NSF Regional Conference Series in Applied Mathematics, No. 46)*, volume 75. SIAM, 1984.

[60] M. Cottrell, J.-C. Fort, and G. Malgouyres. Large deviations and rare events in the study of stochastic algorithms. *IEEE Transactions on Automatic Control*, 28(9):907–920, 1983.

[61] R.Y. Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89–112, 1997.

[62] R.Y. Rubinstein. Combinatorial optimization, cross-entropy, ants and rare events. In S. Uryasev and P.M. Pardalos, editors, *Stochastic Optimization: Algorithms and Applications*, pages 303–363. Springer, Boston, MA, USA, 2001.

[63] R.Y. Rubinstein. The cross-entropy method for combinatorial and continuous optimization. *Methodology and Computing in Applied Probability*, 1(2):127–190, 1999.

[64] P. L'Ecuyer and B. Tuffin. Effective approximation of zero-variance simulation in a reliability setting. In *Proceedings of the 2007 European Simulation and Modeling Conference*, pages 48–54, Ghent, Belgium, 2007. EUROSIS.

[65] T.P.I. Ahamed, V.S. Borkar, and S. Juneja. Adaptive importance sampling technique for Markov chains using stochastic approximation. *Operations Research*, 54(3):489–504, 2006.

[66] P.Y. Desai and P.W. Glynn. A Markov chain perspective on adaptive Monte Carlo algorithms. In *Proceedings of the 2001 Winter Simulation Conference (WSC)*, pages 379–384. IEEE, 2001.

[67] C. Kollman, K. Baggerly, D. Cox, and R. Picard. Adaptive importance sampling on discrete Markov chains. *Annals of Applied Probability*, 9(2):391–412, 1999.

[68] P. L'Ecuyer, J.H. Blanchet, B. Tuffin, and P.W. Glynn. Asymptotic robustness of estimators in rare-event simulation. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 20(1):6:1–6:41, 2010.

[69] P. L'Ecuyer and B. Tuffin. Approximate zero-variance simulation. In *Proceedings of the 2008 Winter Simulation Conference (WSC)*, pages 170–181. IEEE, 2008.

[70] P. Shahabuddin. *Simulation and Analysis of Highly Reliable Systems*. PhD thesis, Stanford University, Stanford, CA, USA, 1990.

[71] M.K. Nakayama. *Simulation of Highly Reliable Markovian and Non-Markovian Systems*. PhD thesis, Stanford University, Stanford, CA, USA, 1991.

[72] P. Heidelberger, P. Shahabuddin, and V.F. Nicola. Bounded relative error in estimating transient measures of highly dependable non-Markovian systems. *ACM Transactions on Modeling and Computer Simulation (TOMACS)*, 4(2):137–164, 1994.

[73] S. Asmussen and P.W. Glynn. *Stochastic Simulation: Algorithms and Analysis*, volume 57. Springer-Verlag, New York, 2007.

[74] B. Tuffin, W. Sandmann, and P. L'Ecuyer. Robustness properties in simulations of highly reliable systems. In *Proceedings of RESIM 2006*, University of Bamberg, Germany, October 2006.

[75] I. Eusgeld, B. Fechner, F. Salfner, M. Walter, P. Limbourg, and L. Zhang. Hardware reliability. In I. Eusgeld, F. Freiling, and R.H. Reussner, editors, *Dependability Metrics: Advanced Lectures*, chapter 9, pages 59–103. Springer, Berlin, 2008.

[76] P.D.F. Conradie, C.J. Fourie, P.J. Vlok, and N.F. Treurnicht. Quantifying System Reliability in Rail Transportation in an Ageing Fleet Environment. *South African Journal of Industrial Engineering*, 26(2):128–142, 2015.

[77] K. Bourouni. Availability Assessment of a Reverse Osmosis Plant: Comparison between Reliability Block Diagram and Fault Tree Analysis Methods. *Desalination*, 313:66–76, 2013.

[78] R. Sedgewick and M. Schidlowsky. *Algorithms in Java, Part 5*. Addison-Wesley, Boston, 2003.

[79] G. Rubino. Network reliability evaluation. In J. Walrand, K. Bagchi, and G.W. Zobrist, editors, *Network Performance Modeling and Simulation*, pages 275–302. Gordon and Breach Science Publishers, Inc., Newark, NJ, USA, 1998.

[80] M.O. Ball. Computational complexity of network reliability analysis: An overview. *IEEE Transactions on Reliability*, 35(3):230–239, 1986.

[81] B.H. Lindqvist. On the Statistical Modeling and Analysis of Repairable Systems. *Statistical Science*, pages 532–551, 2006.

[82] H. Ascher and H. Feingold. *Repairable Systems Reliability: Modeling, Inference, Misconceptions and their Causes*. Marcel Dekker, New York, 1984.

[83] G. Ciardo, J. Muppala, and K.S. Trivedi. SPNP: Stochastic Petri Net package. In *Proceedings of the Third International Workshop on Petri Nets and Performance Models, PNPM89*, pages 142–151. IEEE, 1989.

[84] G. Ciardo and K.S. Trivedi. A decomposition approach for stochastic reward net models. *Performance Evaluation*, 18(1):37–59, 1993.

[85] B. Tuffin. Highly reliable Markovian systems interval availability estimation by importance sampling. In *Proceedings of the 2014 Winter Simulation Conference (WSC)*, pages 553–563. IEEE, 2014.

[86] A. Ridder. Importance sampling simulations of Markovian reliability systems using cross-entropy. *Annals of Operations Research*, 134(1):119–136, 2005.

[87] C. Alexopoulos and B.C. Shultes. Estimating reliability measures for highly-dependable Markov systems, using balanced likelihood ratios. *IEEE Transactions on reliability*, 50(3):265–280, 2001.

[88] P.W. Glynn. Simulation algorithms for regenerative processes. In S.G. Henderson and B.L. Nelson, editors, *Simulation*, volume 13 of *Handbooks in Operations Research and Management Science*, chapter 16, pages 477–500. Elsevier, 2006.

[89] P. Shahabuddin. Importance sampling for the simulation of highly reliable Markovian systems. *Management Science*, 40(3):333–352, 1994.

[90] Z. Zheng and P.W. Glynn. Extensions of the regenerative method to new functionals. In *Proceedings of the 2016 Winter Simulation Conference (WSC)*, pages 289–301. IEEE, 2016.

[91] V. Volovoi. Modeling of system reliability petri nets with aging tokens. *Reliability Engineering & System Safety*, 84(2):149 – 161, 2004.

[92] R.A. Sahner and K.S. Trivedi. Reliability modeling using sharpe. *IEEE Transactions on Reliability*, R-36(2):186–193, June 1987.

[93] J.B. Dugan, K.S. Trivedi, M.K. Smotherman, and R.M. Geist. The hybrid automated reliability predictor. *AIAA Journal of Guidance, Control and Dynamics*, 9(3):319–331, 1986.

[94] A. Goyal, W.C. Carter, E. de Souza e Silva, S.S. Lavenberg, and K.S. Trivedi. The system availability estimator. In *Proceedings of the 16th International Symposium on Fault-Tolerant Computing*, pages 84–89, 1986.

[95] H. Choi and K.S. Trivedi. Approximate performance models of polling systems using stochastic Petri Nets. In *INFOCOM'92. Eleventh Annual Joint Conference of the IEEE Computer and Communications Societies*, volume 3, pages 2306–2314. IEEE, 1992.

[96] C. Hirel, B. Tuffin, and K.S. Trivedi. SPNP: Stochastic Petri Nets. version 6.0. In *International Conference on Modelling Techniques and Tools for Computer Performance Evaluation*, pages 354–357. Springer, 2000.

[97] R.Y. Rubinstein and P.W. Glynn. How to deal with the curse of dimensionality of likelihood ratios in Monte Carlo simulation. *Stochastic Models*, 25(4):547–568, 2009.

[98] S. Hunter, T. Philip, and K.S. Trivedi. Combined performance and availability analysis of a switched network application. In *1997 IEEE International Conference on Communications*, volume 1, pages 241–245, Montréal, Québec, Canada, 1997. IEEE.

[99] O.C. Ibe and K.S. Trivedi. Stochastic Petri Net models of polling systems. *IEEE Journal on Selected Areas in Communications*, 8(9):1649–1657, 1990.

Acronyms

Important Acronyms used

BRE	Bounded Relative Error
CBTC	Communication Based Train Control
ccf	Common Cause Failure
CE	Cross Entropy
CI	Confidence Interval (by default 95%)
CLT	Central Limit Theorem
CMC	Crude Monte-Carlo
CSPL	C-based Stochastic Petri Nets Language
CTMC	Continuous Time Markov Chain
DCS	Data Communication System
DES	Discrete Event Simulations
DTMC	Discrete Time Markov Chain
EOI	Event Of Interest
ERG	Extended Reachability Graph
FTA	Fault Tree Analysis
GSPN	Generalized Stochastic Petri Nets
GUI	Graphical User Interface
HRMS	Highly Reliable Markovian Systems
IS	Importance Sampling
LCC	Life Cycle Cost
LDT	Large Deviations Theory
LE	Logarithmic Efficiency
MC	Monte Carlo simulation
MDT	Mean Down Time
MRM	Markov Reward Models
MSDIS	Measure Specific Dynamic Importance Sampling
MTBF	Mean Time Between Failures
MTTF	Mean Time To Failure
Multi-D	Multi-Dimensional
One-D	One-Dimensional
pdf	Probability Density Function
PN	Petri Nets
RAM	Reliability, Availability and Maintainability
RAMS	Reliability, Availability, Maintainability and Safety
RBD	Reliability Block Diagram
RE	Relative Error
r.v.	Random Variable

SOR	Successive Over-Relaxation
SPNP	Stochastic Petri Nets Package
SPN	Stochastic Petri Nets
SRN	Stochastic Reward Nets
VM	Variance Minimization
VRE	Vanishing Relative Error

Appendix A

Stochastic Petri Nets Package

Stochastic Petri Nets Package (SPNP) [34] is a powerful and versatile tool developed at the Duke University. The input language for SPNP is C-based Stochastic Petri Nets Language (CSPL), an extension of the C programming language with additional constructs to facilitate easy modeling of Stochastic Petri Net (SPN) models [96]. The SPN models' definition is based on *SPN Reward Models* or *Stochastic Reward Nets (SRNs)* which itself is based on the *Markov Reward Models (MRM)* paradigm [34, 83, 84, 95, 98, 99]. The solution methods for various SPN models can be classified into the following two broad categories:

A.1 Analytic Numeric Methods

The steady-state measures for CTMC or DTMC can be solved using numerical techniques like Steady-State SOR (Successive Overrelaxation), Steady-State Gauss-Seidel and Steady-State Power methods [96]. For evaluation of transient measures of a CTMC, standard uniformization and uniformization using the Fox and Glynn method for computing the Poisson probabilities is available [96]. The numeric solutions of the CTMC or DTMC is possible for not very big models, when the state space is too large.

A.2 Simulation

Simulation methods are commonly used when the state spaces are too large to be solved analytically. In SPNP, there are many options to solve Markovian and Non-Markovian SRNs. Some of them are:

- Independent Replications: to compute cumulative or average instantaneous measures up to a fixed simulation time.
- Batch Means: to compute steady-state measures and building of CI.
- Regenerative simulations to estimate steady-state measures for Markov models.

- Importance Sampling: Restart and Splitting, standard IS by modifying the distributions (or probabilities) of transitions and IS simulations using Regenerative simulations in Markovian context.

A complete discussion of the SPNP tool is beyond the scope of this study, however, Figure A.1 gives a brief idea of the capabilities available in the SPNP tool (version 6.0 and after). In the version 6.1 used in the current study, there is no Graphical User Interface (GUI) available as mentioned for version 6.0 in [34]. Readers are advised to refer to [34] for a complete explaination of this tool.

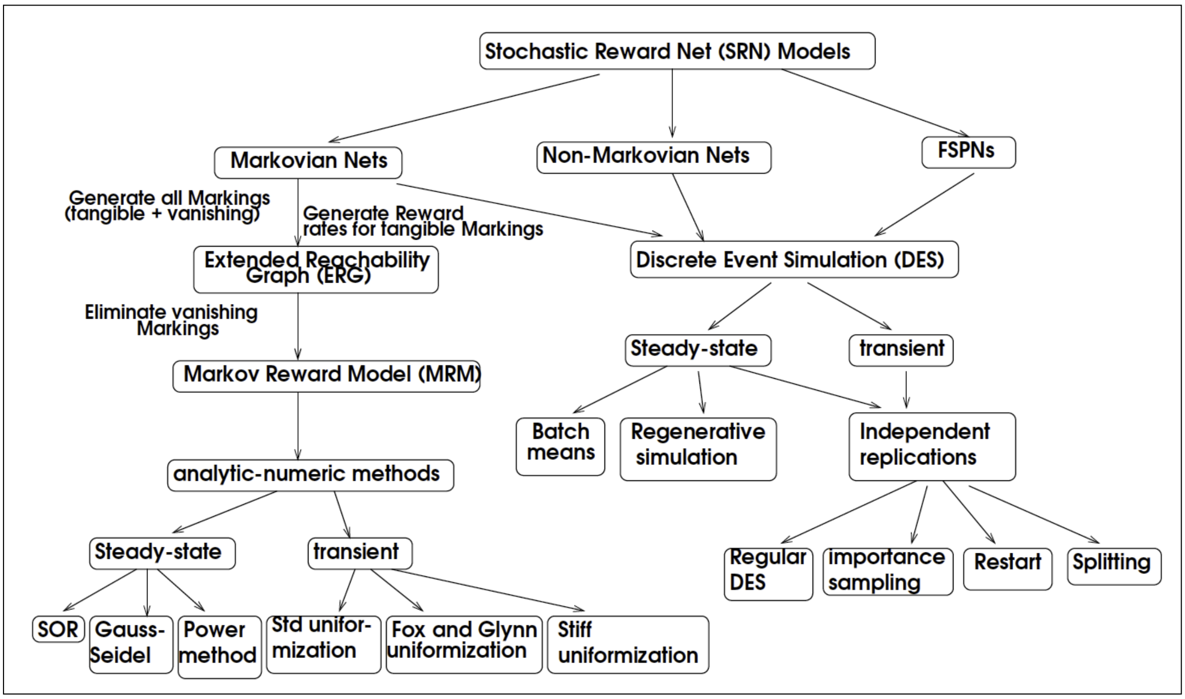

Figure A.1 Classification and analysis methods in SPNP.

A.3 Methodologies Added

We added a new methodology in the SPNP package based on multi-level CE Algorithm 3 for performing regenerative IS simulations in Markovian SPNs, while using the IS rates obtained for transitions of interest (to be defined by the user) from the CE algorithm. The CE scheme is used as a pre-simulation method.

Appendix B

SPN Models

The following SPN models have been considered in the current work.

B.1 A single 2-out-of-3 Model with Logistics

We are going to present the case with one technician and its associated spares, but this block can be repeated as much as desired without a significant effort to represent several depots. Several technicians in a depot would mean several tokens in the place "TechHome". Similarly, a single 2-out-of-3 (2oo3) system will be analyzed, but the block can be repeated at will.

We decompose the Petri net presentation in several sub-figures for an easier visualization.

B.1.1 The Technician(s) and the Spares Modeling

The logistics aspects of the model is represented as in Figure B.1. In Figure B.1, we can see that the blocks when the technician leaves can be copied for each site (that is, for each 2oo3 system). There is a guard function on the immediate transitions *Detected failure "unit"* (*unit* being the number of 2oo3 systems) and an inhibitor arc meaning that the technician cannot leave if there is no spare available (should wait then). The guard function for *Detected failure "unit"* is requiring that one of the components of the 2oo3 system unit has a detected failure, i.e., one of the places "Pdetected_a", "Pdetected_b" or "Pdetected_c" has a token (those places are defined in next section). The availability of spares is described by the independent Petri net at the top of the Figure B.1. A possible extension could be that if all fails before restoration in an undetected mode, then we probably need to move to the detected mode and not wait for the next inspection (would require a new immediate transition to be added). Next, we describe the description of what is between the places "OnSiteTech" and "RestOver".

B.1.2 The 2oo3 System

We present here the Petri net for a single 2oo3 systems (the junction with the previous Petri net is via places "OnSiteTech" and "RestOver"). To simplifiy the graph, we avoid inhibitor

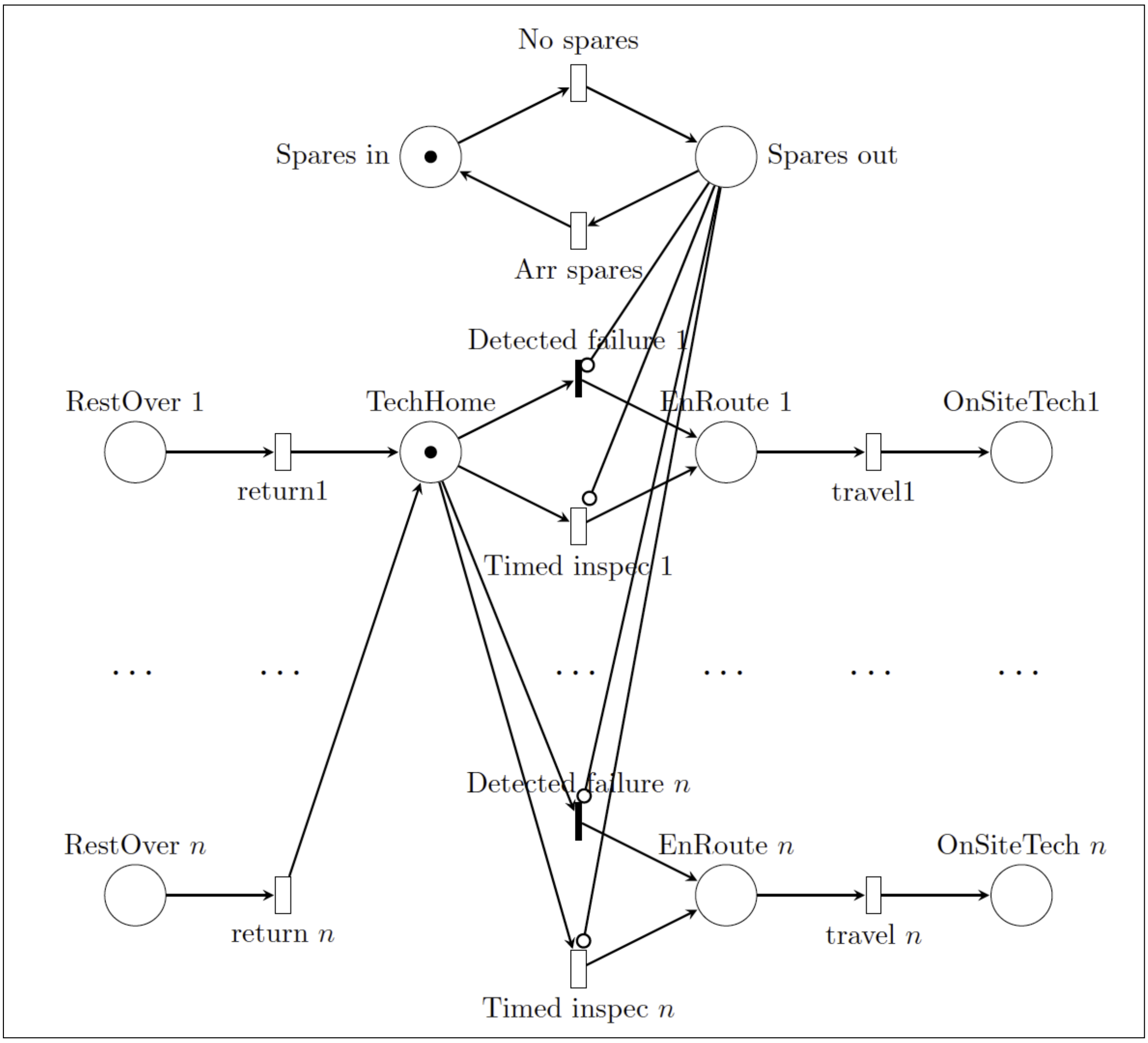

Figure B.1 Logistics part of the 2oo3 model.

arcs and rather define guard functions described after the graph. We first present some subparts, and then the whole graph of the 2oo3. Each of the three components (A, B and C) can be up, failed and detected, or failed and undetected. There is a common cause failure (CCF) that makes the 3 components fail at the same time. To be restored, the technician has to be there. When the technician inspects (place "Inspect"), the undetected failure becomes detected, and when the inspection is over, restoration starts (a token at place "Restore" meaning that the technician is working on the restoration). This is summarized as follows for component A (the same thing is extended for the other two components B and C, with corresponding arcs).

Next, we describe the Petri subnet for the 2oo3. On this graph we have put three components, plus the time to start the inspection, and an immediate transition *readyRestore*, which fires when the inspections of all components is over and all undetected failures are detected.

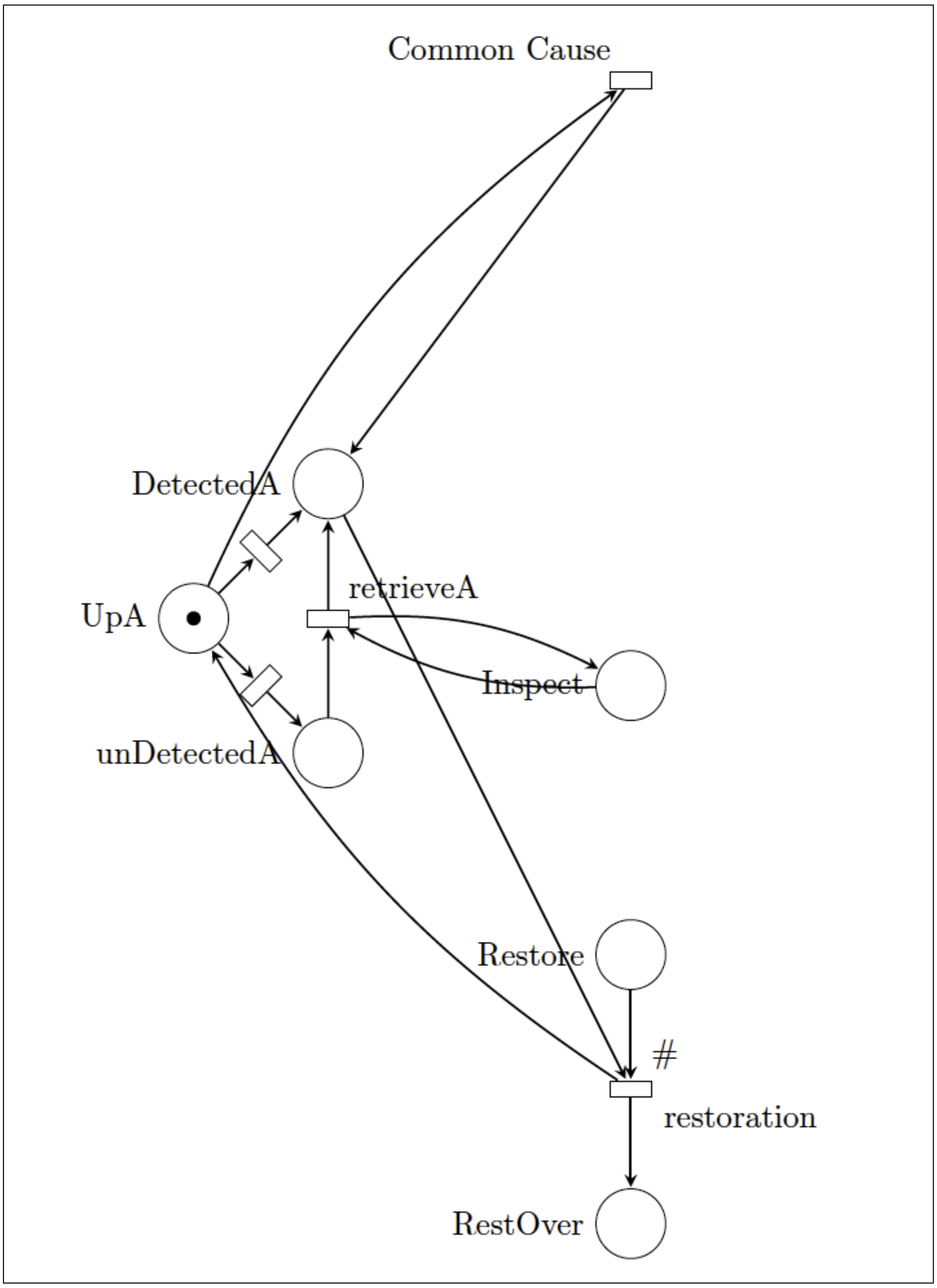

Figure B.2 Component level model.

We have introduced the immediate transition *noRestorationNeeded* to account for timed inspection for which there was actually no need to repair. then, after inspection, the technician can actually return immediately to the depot.

General Comments:

1. Guard function "gReady": On immediate transition *readyRestore*: transition can not fire if there are still some undetected failures, i.e., if one of the three places "UndetectedA", "UndetectedB" or "UndetectedC" has a token.

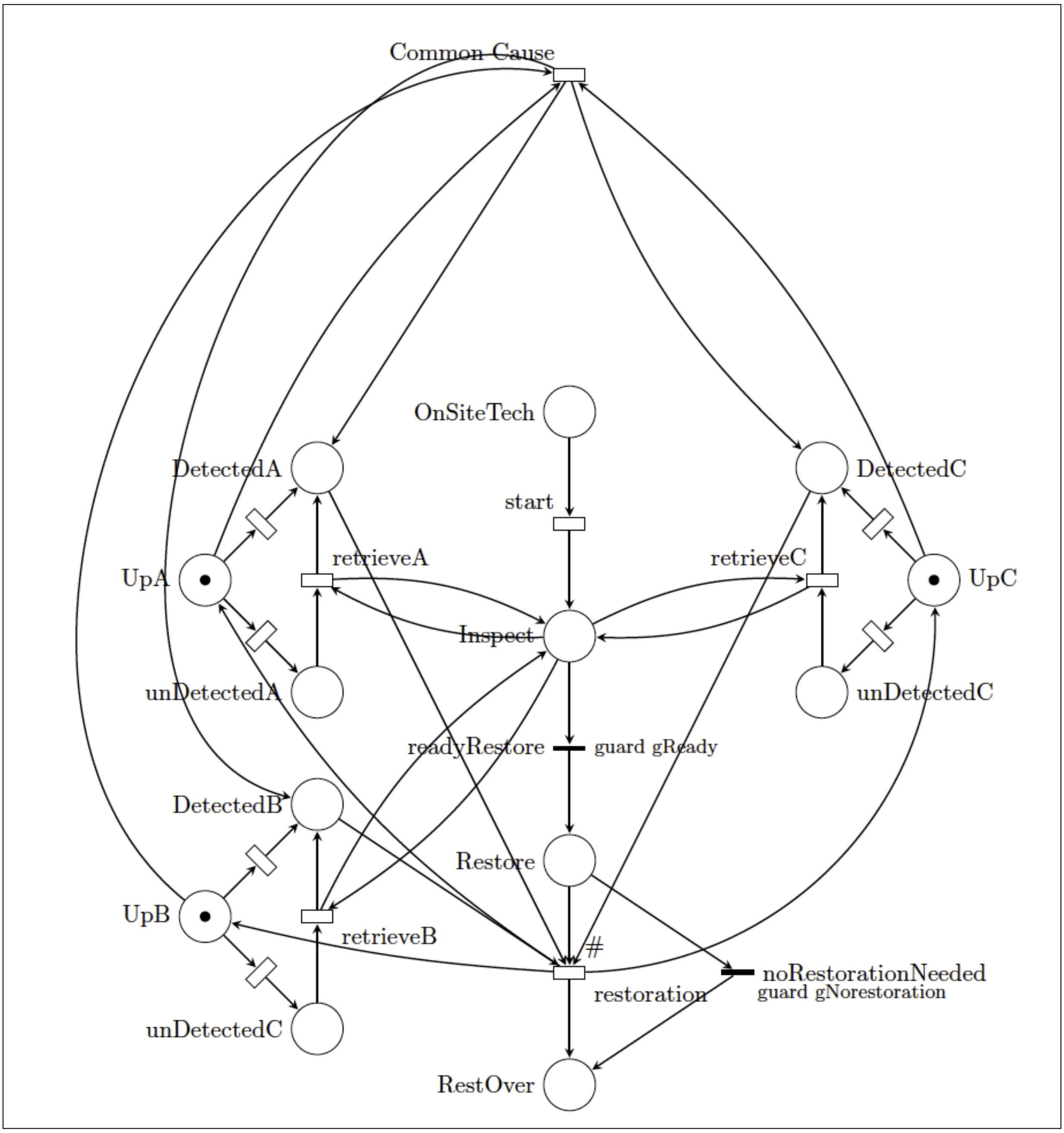

Figure B.3 Entire 2oo3 model with logistics in Figure B.1

2. For the transition *restoration*, the cardinality of the input and output arcs from "Detected$c$" and to "Upc" ($c \in \{A,B,C\}$) is the number of tokens in "Detected$c$" (1 or 0). In other words, when restoration, the detected failed components are repaired (we do not need all the components to be failed).

3. Guard function "gNorestoration": the immediate transition *noRestorationNeeded* is enabled if all three components are up and the technician is ready to restore (that is, place "Restore" contains a token). No need for restoration in this case.

In the present model, as soon as the technician starts inspection, the potential component hidden failures are tried to be identified in parallel (the three transitions *retrieveA*, *retrieveB* and *retrieveC* are in competition). This maybe correct if the technician runs some software working simultaneously at all components, but probably not if the technician does it manually. In that case, a sequential treatment would be more relevant.

This model of a single 2oo3 system unit has 7 failure transitions of three different types: Detected Failures of components (*detA*, *detB* and *detC*), undetected failures of components (*udetA*, *udetB* and *udetC*) and a common cause failure (*ccf*). Other transitions for which distributions are needed to be defined in the model are:

- Rate at which spares become available
- Rate at which spares become unavailable
- Rate at which undetected failures are detected
- Rate at which on-site technicians start inspections
- Rate of timed inspections
- Travel rate for technicians
- Repair rate

Titre : Estimation de la Disponibilité par Simulation, pour des Systèmes incluant des Contraintes Logistiques.

Mots clés : paramètres de sûreté de fonctionnement, la simulation d'événements rares, l'optimisation de CE.

Résumé : L'analyse des FDM fait partie intégrante de l'estimation du coût du cycle de vie des systèmes ferroviaires. Ces systèmes sont hautement fiables et présentent une logistique complexe. Les simulations Monte Carlo dans leur forme standard sont inutiles dans l'estimation efficace des paramètres des FDM à cause de la problématique des événements rares. C'est ici que l'échantillonnage préférentiel joue son rôle. C'est une technique de réduction de la variance et d'accélération de simulations. Cependant, l'échantillonnage préférentiel inclut un changement de lois de probabilité (changement de mesure) du modèle mathématique. Le changement de mesure optimal est inconnu même si théoriquement il existe et fournit un estimateur avec une variance zéro.

Dans cette thèse, l'objectif principal est d'estimer deux paramètres pour l'analyse des FDM: la fiabilité des réseaux statiques et l'indisponibilité asymptotique pour les systèmes dynamiques. Pour ce faire, la thèse propose des méthodes pour l'estimation et l'approximation du changement de mesure optimal et l'estimateur final. Les contributions se présentent en deux parties: la première partie étend la méthode de l'approximation du changement de mesure de l'estimateur à variance zéro pour l'échantillonnage préférentiel. La méthode estime la fiabilité des réseaux statiques et montre l'application à de réels systèmes ferroviaires. La seconde partie propose un algorithme en plusieurs étapes pour l'estimation de la distance de l'entropie croisée. Cela permet d'estimer l'indisponibilité asymptotique pour les systèmes markoviens hautement fiables avec des contraintes logistiques.

Les résultats montrent une importante réduction de la variance et un gain par rapport aux simulations Monte Carlo.

Title : Availability Estimation by Simulation for Systems including Logistics.

Keywords : reliability metrics, rare events simulations, cross-entropy optimization.

Abstract: RAM analysis forms an integral part in estimation of Life Cycle Costs (LCC) of passenger rail systems. These systems are highly reliable and include complex logistics. Standard Monte-Carlo simulations are rendered useless in efficient estimation of RAM metrics due to the issue of rare events. Systems failures of these complex passenger rail systems can include rare events and thus need efficient simulation techniques.

Importance Sampling (IS) are an advanced class of variance reduction techniques that can overcome the limitations of standard simulations. IS techniques can provide acceleration of simulations, meaning, less variance in estimation of RAM metrics in same computational budget as a standard simulation. However, IS includes changing the probability laws (change of measure) that drive the mathematical models of the systems during simulations and the optimal IS change of measure is usually unknown, even though theoretically there exist a perfect one (zero-variance IS change of measure). In this thesis, we focus on the use of IS techniques and its application to estimate two RAM metrics : reliability (for static networks) and steady state availability (for dynamic systems). The thesis focuses on finding and/or approximating the optimal IS change of measure to efficiently estimate RAM metrics in rare events context. The contribution of the thesis is broadly divided into two main axis : first, we propose an adaptation of the approximate zero-variance IS method to estimate reliability of static networks and show the application on real passenger rail systems ; second, we propose a multi-level Cross-Entropy optimization scheme that can be used during pre-simulation to obtain CE optimized IS rates of Markovian Stochastic Petri Nets (SPNs) transitions and use them in main simulations to estimate steady state unavailability of highly reliable Markovian systems with complex logistics involved. Results from the methods show huge variance reduction and gain compared to MC simulations.