Chance-constrained problems and rare events: an importance sampling approach

Javiera Barrera Tito Homem-de-Mello Eduardo Moreno
Bernardo K. Pagnoncelli Gianpiero Canessa

February 25, 2014

Abstract

We study chance-constrained problems in which the constraints involve the probability of a rare event. We discuss the relevance of such problems and show that the existing sampling-based algorithms cannot be applied directly in this case, since they require an impractical number of samples to yield reasonable solutions. Using a Sample Average Approximation (SAA) approach combined with importance sampling (IS) techniques, we show how variance can be reduced uniformly over a suitable approximation of the feasibility set, and as a result the problem can be solved with much fewer samples. We provide sufficient conditions to obtain such uniform variance reduction and prove asymptotic convergence of the combined SAA-IS approach. We apply our methodology to a telecommunications problem, find IS distributions that satisfy the conditions laid out for uniform variance reduction in that context and present numerical results to illustrate the ideas.

1 Introduction

Chance-constrained programming was first introduced in [12] and has been extensively studied since then.In many situations a decision maker wants a constraint to be satisfied with some pre-specified probability, that is, violation may occur for some realizations that as a whole have small probability. Applications include finance [6, 16], energy [2], water pollution management [25], mining [11] and telecommunications [24, 40]. For a theoretical background we refer to [32] and Chapter 4 of [38].

Although chance-constrained programming is a very flexible modeling tool to incorporate uncertainty into optimization problems, the resulting problems are usually hard to solve. The requirement of having to satisfy a certain constraint with high probability involves computing a multidimensional integration, which can only be performed exactly for certain distributions. In addition, the set of feasible solutions satisfying the chance constraint is usually non convex and therefore unsuitable for most optimization algorithms. Even the evaluation of feasibility for a given candidate solution cannot be done explicitly and one has to employ Monte Carlo simulation to check it.

Different approaches have been proposed in the recent years to deal with chance-constrained problems. Among these approaches, we mention the concept of efficient points [5, 15], the Bernstein approximation [29], combinatorial patterns [23] and data-driven optimization [20]. Another common technique is a sampling-based method known as Sample Average Approximation (SAA). The SAA generates a sample from the original distribution of the problem and creates an approximate problem with new sampled constraints that replace the original chance constraint. Such an approach has been well studied in the literature; a series of theoretical results ensure that the optimal value and the set of optimal solutions of those approximate problems converge to their true counterparts under mild conditions [30, 26].

The most important parameter to be chosen in SAA is the number of samples, or scenarios, that will be drawn from the original distribution. A series of papers ([7, 8, 9, 26]) study SAA (sometimes called "scenario approach") and find a lower bound on the number of samples such that the solution of the sampled problem is feasible to the original chance-constrained problem with high probability. The theoretical bound is often too conservative in practice, but constraint discarding schemes can be used to improve the quality of the obtained solutions [31]. In most applications of chance constraints the probability of violation should not exceed 10%, 5% or 1%. The formulas for the sample size in the literature are usually of the order of one over the probability of violation, so the resulting problems are of manageable size.

In this paper we consider chance-constrained optimization problems with rare events, that is, problems in which the violation probability is very small, e.g, 10^{-6}. In this case the theoretical guidelines would lead to extremely large problems that cannot be solved due to computational limitations. Working with such small values raises two important concerns. The first is how relevant those problems are. For example, in the aviation industry the maximum tolerated probability of a catastrophic event per flight hour is 10^{-9} [39]; in structural engineering the maximum tolerated probability of failures is of order 10^{-4} [17]. Many other such examples exist.

The second concern is from an algorithmic perspective: if the violation is so small, would not it be better to simply forbid violation, assigning the value zero to the probability of violation and solving a "robustified" the problem? We will show through an example that the answer is no. Significantly different solutions may arise when some violation is allowed, even if the value is as small as 10^{-6}. This is expected for distributions that have a relatively long tail.

New techniques are needed to tackle chance-constrained problems in the presence of rare events. We propose an integration of SAA with importance sampling (IS), a technique widely used in simulation to estimate probabilities of rare events that dates back to [21] and [34]. Importance sampling is still an active research topic that generates great interest among researchers; see, for instance, [22] and references therein.

In the optimization context the use of importance sampling is scarcer (e.g., [14, 19]). We are not aware of any previous work that uses importance sampling in chance-constrained programming. The difficulty is that for each decision variable there might be a different optimal IS estimator. The challenge is to choose an estimator that is *uniformly* efficient, that is, that lowers the variance for all possible solutions.

In this paper we discuss such issues in detail. We define precisely what is meant by uniform variance reduction and give a sufficient condition to accomplish that goal. We also discuss how the choice of an IS distribution that depends on the decision variables of the optimization problem is instrumental to achieve larger variance reduction, which in turn means that the problem can be solved with much fewer samples. Reducing the number of samples is crucial especially in cases in which generating samples can be expensive. In addition, we extend some of the convergence results for SAA available in the literature to the case with importance sampling.

We apply our methodology to an optical network problem in which customers want to communicate with each other with a certain rate. A central planner wants to minimize installed capacity subject to having a low probability that the links of the network do not have enough capacity to allow communication (the so-called *blocking probability*). We construct IS estimators for this problem and provide a series of results showing that variance can be uniformly controlled. We also show how the resulting problems can be modeled as mixed-integer programs, a non-trivial task especially when the IS distribution depends on the decision variables.

A key insight obtained from our construction is the need to reduce variance uniformly on a suitable *outer approximation* of the feasibility set—i.e., it is important to reduce variance at infeasible points so the infeasibility can be detected, but it is counterproductive to reduce variance at points that are "too far" from the feasibility set. Another insight is the existence of a trade-off between the amount of variance reduction and the complexity of the resulting optimization problem, which means that larger reduction is achieved at the expense of a problem with more variables and more constraints.

We illustrate our findings with extensive numerical computations for a ring network topology. The numerical results corroborate our theoretical findings, demonstrating that when the IS distribution is carefully constructed the problem can be solved with a small number of samples, even when the probability of violation is very small.

The rest of the paper is organized as follows. In Section 2 we introduce the rare-event chance-constrained problem, discuss IS techniques and introduce a formal definition for uniform variance reduction that we use throughout the paper. We also present convergence results showing the asymptotic validity of the approximating formulations. In Section 3 we describe in detail the telecommunication problem and present mixed integer programming formulations for the rare-event chance-constrained problem with importance sampling. We construct IS measures that ensure uniform variance reduction over an appropriate set in Section 4. Numerical results are presented in Section 5 and concluding remarks are discussed in Section 6.

2 Chance-constrained programming and SAA

In this paper we consider problems of the form

$$\min_{x\in X} h(x) \text{ s.t. } \mathbb{P}\big\{G_i(x,\xi)\le 0\big\} \ge 1-\alpha_i\ , \quad i=1,\dots,M. \tag{1}$$

Here $X\subset\mathbb{R}^n$, ξ is a random vector defined on an underlying probability space $(\Omega,\mathcal{F},\mathbb{P})$ with probability distribution supported on a set $\Xi\subset\mathbb{R}^d$, $\alpha_i\in(0,1)$, $h:\mathbb{R}^n\mapsto\mathbb{R}$ is a real valued function and $G_i:\mathbb{R}^n\times\mathbb{R}^d\mapsto\mathbb{R}$ is a real-valued function. By choosing $M>1$ we allow multiple chance constraints, each one with its own reliability level α_i.

Problem (1), which we will refer to as the original problem, can be written in the following equivalent form:

$$\min_{x\in X} h(x) \text{ s.t. } p_i(x)\le\alpha_i\ , \quad i=1,\dots,M, \tag{2}$$

where

$$p_i(x) := \mathbb{P}\{G_i(x,\xi)>0\} = \mathbb{E}_\xi\left[\mathbb{1}_{\{G_i(x,\xi)>0\}}\right]\ , \tag{3}$$

where $\mathbb{1}_A$ is the indicator function of the event A, i.e., $\mathbb{1}_A=1$ if A occurs and $\mathbb{1}_A=0$ otherwise. In the discussion that follows we will assume that $M=1$, and drop the subscript. Later we will consider the case $M>1$.

For a given sample $(\xi^1,\dots,\xi^N)$ of size N from the distribution of ξ, a natural approximation of function $p(x)$ in (3) is

$$\hat{p}(x) \ := \ \frac{1}{N}\sum_{j=1}^{N}\mathbb{1}_{\{G(x,\xi^j)>0\}}\ , \tag{4}$$

that is, $\hat{p}(x)$ is equal to the proportion of indices j such that $G(x,\xi^j)>0$. The Sample Average Approximation (SAA) problem associated with the generated sample is defined as

$$\min_{x\in X}\ h(x) \text{ s.t. } \hat{p}(x)\le\gamma\ . \tag{5}$$

Following [30], we allow the tolerance levels $\gamma\ge 0$ of the SAA problem and α of the original problem to be different. A very important result regarding feasibility of chance-constrained programs was derived in [8]. The authors find a bound on the probability that a solution of the SAA problem with $\gamma=0$ (the so-called scenario problem) violates the original problem. Using the simpler expression shown in [10], they prove that if the sample size N satisfies, for a d-dimensional problem,

$$N\ge\frac{2}{\alpha}\left(\ln\frac{1}{\beta}+d\right)\ , \tag{6}$$

then the optimal solution of the SAA problem violates the chance constraint in the original problem with probability at most β. The result is very powerful because it does not depend on the distribution of ξ, requiring only convexity of the function G with respect to x. A similar result is derived in [26] for the case of a general function G but when the feasibility set X is finite.

As mentioned earlier, one of the goals of this paper is to solve chance constrained problems in which the value of α is very small, say, of order 10^{-6}. Using formula (6) we see that the resulting sample size would be impractical, of order 10^6. For this reason the applicability of (6) is limited in our context. The same comment applies to the sample size estimates derived in [26].

Motivated by those difficulties, we propose the use of IS techniques within the SAA. In the remainder of this section we formulate the SAA problem with importance sampling and show that the modified problem is still consistent.

2.1 Importance Sampling Techniques

Importance sampling (IS) is a well known simulation technique used to reduce variance (see, e.g., [4] for a comprehensive discussion). For completeness, we review next the basic ideas of IS. Let μ denote the measure in $\mathbb{R}^d$ induced by the random vector ξ, i.e. $\mu(A) = \mathbb{P}(\xi \in A)$. We want to estimate $p(x)$ for all $x \in X$. Now let us consider $\hat{\xi}$ a new random vector with induced measure ν, which we will call the IS measure, such that μ is absolute continuous with respect to ν, i.e., $\mu(A) = 0$ if $\nu(A) = 0$. Let $\hat{\xi}^1, \ldots \hat{\xi}^N$ be i.i.d. copies of $\hat{\xi}$. Define

$$\hat{p}^{\mathrm{IS}}(x) \ := \ \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{\{G(x,\hat{\xi}^j)>0\}} L(\hat{\xi}^j) \ , \tag{7}$$

where $L(\cdot)$ is the likelihood ratio $L(\cdot) \ := \ \frac{d\mu}{d\nu}(\cdot)$. The function L is the *Radon-Nikodym derivative* of μ with respect to ν, i.e., L is a function $\mathbb{R}^d \mapsto \mathbb{R}$ such that $\mu(A) = \int_A L(s) d\nu(s)$ for any Borel set $A \subset \mathbb{R}^d$. Note that for any measurable function $f : \mathbb{R}^d \mapsto \mathbb{R}$ we have that $\int_{\mathbb{R}^d} f(s)\, d\mu(s) \ = \ \int_{\mathbb{R}^d} f(s) L(s)\, d\nu(s)$, that is,

$$\mathbb{E}_\xi \left[f(\xi) \right] = \mathbb{E}_{\hat{\xi}} \left[f(\hat{\xi}) L(\hat{\xi}) \right]. \tag{8}$$

In particular, by taking $f \equiv 1$ we have $\mathbb{E}_{\hat{\xi}}[L(\hat{\xi})] = 1$. In the case where both ξ and $\hat{\xi}$ have discrete support L is the ratio between the respective probabilities mass functions, whereas in the case where both ξ and $\hat{\xi}$ have probability densities L is the ratio between the respective probability densities. For each $x \in X$, we have that $\hat{p}^{\mathrm{IS}}(x)$ is an unbiased estimator of $p(x)$, since

$$\mathbb{E}_{\hat{\xi}} \left[\hat{p}^{\mathrm{IS}}(x) \right] \ = \ \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_{\hat{\xi}} \left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}} L(\hat{\xi}) \right] \ = \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_\xi \left[\mathbb{1}_{\{G(x,\xi)>0\}} \right] = p(x).$$

Notice that the variance of the standard estimator $\hat{p}(x)$ in (4) is given by

$$N\mathrm{Var}(\hat{p}(x)) \ = \ p(x) - p(x)^2 \ = \ \mathbb{E}_\xi \left[\mathbb{1}_{\{G(x,\xi)>0\}} \right] - p(x)^2$$

whereas the variance of the new estimator $\hat{p}^{\mathrm{IS}}(x)$ is given by

$$N\mathrm{Var}(\hat{p}^{\mathrm{IS}}(x)) = \mathbb{E}_{\hat{\xi}} \left[\mathbb{1}^2_{\{G(x,\hat{\xi})>0\}} L(\hat{\xi})^2 \right] - p(x)^2 = \mathbb{E}_\xi \left[\mathbb{1}_{\{G(x,\xi)>0\}} L(\xi) \right] - p(x)^2,$$

where the second equality follows from (8). Thus, we see that by choosing the IS measure ν in such a way that the event $\mathbb{1}_{\{G(x,\xi)>0\}}$ becomes more likely under that distribution, the variance of the new estimator will be smaller. In fact, an optimal choice for ν is to put all of its weight on the set $\{\xi : G(x,\xi) > 0\}$; however, such a choice is impractical as it requires knowledge of the original probability $p(x)$.

There is an extensive literature in simulation on how to chose a “good” IS measure ν, especially in the context of estimation of rare-event probabilities; see, for instance, [22] for a recent account. Another approach is to restrict the choice of IS distributions to some parametric family, say, indexed by θ, and then to find the “best” (in some sense) parameter θ^*. For example, θ^* can be the parameter that minimizes the variance of the IS estimator subject to some restrictions, or one that minimizes some kind of distance to the best (but idealized) distribution; see [36] and [35] for discussions. Many other approaches, which typically work by exploiting somehow the structure of the problem, exist as well. In the telecommunications problem described in Section 3 we use a parametric distribution but also exploit the structure of the problem in order to obtain further variance reduction.

2.2 Enhanced IS estimators

We move one step further and allow the IS estimator to depend on the decision variable x. As we shall see later, this generalization will allow us to obtain better estimators. Before we proceed, we introduce an assumption and some notation. Let $I := \{1, \ldots, d\}$ (recall that d is the dimension of ξ).

Assumption 1. The original probability measure μ and the IS probability measure ν are product measures in $\mathbb{R}^d$, i.e., that all components of ξ and $\hat{\xi}$ are independent.

Definition 1. Given a subset $J \subseteq I$ and $x \in X$, we say that the function $G(x, \cdot)$ is *J-determined* if there is a function $G_J : \mathbb{R}^n \times \mathbb{R}^{|J|} \mapsto \mathbb{R}$ such that

$$G\left(x, (z_i)_{i \in I}\right) \;=\; G_J\left(x, (z_j)_{j \in J}\right)$$

for any vector $z \in \mathbb{R}^d$. In words, only the coordinates z_j for $j \in J$ matter for calculation of $G(x, z)$.

Next, let I_x be a set-function that chooses a subset of I for each $x \in X$. Given $x \in X$, there exists a Borel measurable function ϕ_x on $\mathbb{R}^{|I_x|}$ such that

$$\phi_x\left((z_i)_{i \in I_x}\right) \;=\; \mathbb{E}_{\hat{\xi}}\left[L(\hat{\xi}) \,|\, (\hat{\xi}_i)_{i \in I_x} = (z_i)_{i \in I_x}\right];$$

see, for instance, [13]. Define now a function $L_x : \mathbb{R}^d \mapsto \mathbb{R}$ in such a way that

$$L_x(\hat{\xi}) \;=\; \phi_x\left((\hat{\xi}_i)_{i \in I_x}\right) \;=\; \mathbb{E}_{\hat{\xi}}\left[L(\hat{\xi}) \,|\, (\hat{\xi}_i)_{i \in I_x}\right]. \tag{9}$$

By construction, L_x is I_x-determined. The following lemma shows that when the function $G(x, \cdot)$ is I_x-determined, the IS estimator of $p(x)$ constructed with the likelihood function L_x is unbiased and its variance is at most the same as the variance of the estimator $\hat{p}^{\mathrm{IS}}(x)$ defined in (7).

Lemma 2. *Suppose that the set-function I_x defined above is such that $G(x, \cdot)$ is I_x-determined for each $x \in X$. Given an i.i.d. sample $(\hat{\xi}^1, \ldots, \hat{\xi}^N)$ from the distribution of $\hat{\xi}$, let*

$$\hat{p}^{\mathrm{IS}_0}(x) \;:=\; \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}_{\{G(x, \hat{\xi}^j) > 0\}} L_x(\hat{\xi}^j). \tag{10}$$

Then $\hat{p}^{\mathrm{IS}_0}(x)$ is also an unbiased estimator of $p(x)$. Moreover,

$$\mathrm{Var}(\hat{p}^{\mathrm{IS}_0}(x)) \;=\; \mathrm{Var}(\hat{p}^{\mathrm{IS}}(x)) - \frac{1}{N}\mathbb{E}_{\hat{\xi}}\left[\mathrm{Var}\left(\mathbb{1}_{\{G(x,\hat{\xi})>0\}} \,|\, (\hat{\xi}_i)_{i \in I_x}\right)\right] \tag{11}$$

Proof. First let us prove that the estimator $\hat{p}^{\mathrm{IS}_0}(x)$ is unbiased, for which it suffices to show that $\mathbb{E}_{\hat{\xi}}\left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}} L_x(\hat{\xi})\right] = p(x)$. Indeed, we have

$$\begin{aligned}
\mathbb{E}_{\hat{\xi}}\left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}} L_x(\hat{\xi})\right] &= \mathbb{E}_{\hat{\xi}}\left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}} \mathbb{E}_{\hat{\xi}}\left[L(\hat{\xi}) \,|(\hat{\xi}_i)_{i \in I_x}\right]\right] \\
&= \mathbb{E}_{\hat{\xi}}\left[\mathbb{E}_{\hat{\xi}}\left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}} L(\hat{\xi}) \,|\, (\hat{\xi}_i)_{i \in I_x}\right]\right] \qquad (12) \\
&= \mathbb{E}_{\hat{\xi}}\left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}} L(\hat{\xi})\right] = p(x), \qquad (13)
\end{aligned}$$

where the second equality follows from the assumption that $G(x, \cdot)$ is I_x-determined, which implies that $G(x, \hat{\xi})$ is measurable with respect to the sigma-algebra generated by $(\hat{\xi}_i)_{i \in I_x}$.

For the second assertion of the lemma, note that

$$\mathbb{E}_{\hat{\xi}}\left[\mathbb{1}^2_{\{G(x,\hat{\xi})>0\}} L_x(\hat{\xi})^2\right] \;=\; \mathbb{E}_{\hat{\xi}}\left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}} \left(\mathbb{E}_{\hat{\xi}}\left[L(\hat{\xi}) \,|\, (\hat{\xi}_i)_{i \in I_x}\right]\right)^2\right]$$

$$
\begin{aligned}
&= \mathbb{E}_{\hat{\xi}}\left[\left(\mathbb{E}_{\hat{\xi}}\left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}}L(\hat{\xi}) \mid (\hat{\xi}_i)_{i\in I_x}\right]\right)^2\right] \\
&= \mathbb{E}_{\hat{\xi}}\left[\mathbb{E}_{\hat{\xi}}\left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}}L(\hat{\xi})^2 \mid (\hat{\xi}_i)_{i\in I_x}\right]\right. \\
&\qquad\qquad \left. - \operatorname{Var}\left(\mathbb{1}_{\{G(x,\hat{\xi})>0\}}L(\hat{\xi}) \mid (\hat{\xi}_i)_{i\in I_x}\right)\right] \\
&= \mathbb{E}_{\hat{\xi}}\left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}}L(\hat{\xi})^2\right] - \mathbb{E}_{\hat{\xi}}\left[\operatorname{Var}\left(\mathbb{1}_{\{G(x,\hat{\xi})>0\}}L(\hat{\xi}) \mid (\hat{\xi}_i)_{i\in I_x}\right)\right]
\end{aligned}
$$

and therefore

$$
\begin{aligned}
N\operatorname{Var}(\hat{p}^{\mathrm{IS}_0}(x)) &= \mathbb{E}_{\hat{\xi}}\left[\mathbb{1}^2_{\{G(x,\hat{\xi})>0\}}L_x(\hat{\xi})^2\right] - p(x)^2 \\
&= \mathbb{E}_{\hat{\xi}}\left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}}L(\hat{\xi})^2\right] - \mathbb{E}_{\hat{\xi}}\left[\operatorname{Var}\left(\mathbb{1}_{\{G(x,\hat{\xi})>0\}}L(\hat{\xi}) \mid (\hat{\xi}_i)_{i\in I_x}\right)\right] - p(x)^2 \\
&= N\operatorname{Var}(\hat{p}^{\mathrm{IS}}(x)) - \mathbb{E}_{\hat{\xi}}\left[\operatorname{Var}\left(\mathbb{1}_{\{G(x,\hat{\xi})>0\}} \mid (\hat{\xi}_i)_{i\in I_x}\right)\right].
\end{aligned}
$$

□

It is important to observe that $\hat{p}^{\mathrm{IS}}(x)$ corresponds to a particular case of $\hat{p}^{\mathrm{IS}_0}(x)$ defined in Lemma 2, obtained by choosing $I_x = I$ for all x.

2.3 Uniform variance reduction

As mentioned earlier, a major challenge that arises when using IS in the context of solving chance-constrained problems with SAA is that we need to use the same IS measure to perform the importance sampling for all $x \in X$. For certain IS measures, variance can be reduced for some point in X but increase for others. We propose the definition of a ε-**Uniform variance reduction** to capture the fact that we need to introduce a new estimator that reduces the variance uniformly with respect to x.

Definition 3. Let $\hat{p}(x)$ be the standard Monte Carlo estimator of $p(x)$ defined in (4), and let $\hat{p}'(x)$ be another estimator of the same quantity. Given $\varepsilon \in [0,1]$, we say that $\hat{p}'(\cdot)$ has *ε-Uniform Variance Reduction* with respect to $\hat{p}(\cdot)$ if for each $x \in X$ we have

$$
\operatorname{Var}\left(\hat{p}'(x)\right) \;\le\; \varepsilon \operatorname{Var}(\hat{p}(x)).
$$

The following proposition gives a sufficient condition to obtain estimators with ε-Uniform Variance Reduction. Before that, we provide a definition.

Definition 4. Let ε be such that $1 \ge \varepsilon \ge p(x)$ for all $x \in X$, and consider the IS estimator $\hat{p}^{\mathrm{IS}_0}(x)$ defined in (10). We say that $\hat{p}^{\mathrm{IS}_0}(\cdot)$ is an *ε-Uniformly Bounded IS Estimator* of $p(\cdot)$ if for all $x \in X$ we have

$$
\mathbb{1}_{\{G(x,\hat{\xi})>0\}}L_x(\hat{\xi}) \;\le\; \varepsilon \qquad \text{w.p.1.} \tag{14}
$$

Proposition 5. *If $\hat{p}^{\mathrm{IS}_0}(\cdot)$ is an ε-Uniformly Bounded IS Estimator of $p(\cdot)$ then it has ε-Uniform Variance Reduction with respect to $\hat{p}(\cdot)$.*

Proof. We have

$$
\begin{aligned}
N\operatorname{Var}(\hat{p}^{\mathrm{IS}_0}(x)) &= \mathbb{E}_{\hat{\xi}}\left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}}L_x(\hat{\xi})^2\right] - \left(\mathbb{E}_{\hat{\xi}}\left[\mathbb{1}_{\{G(x,\hat{\xi})>0\}}L_x(\hat{\xi})\right]\right)^2 \\
&= \mathbb{E}_{\xi}\left[\mathbb{1}_{\{G(x,\xi)>0\}}L_x(\xi)\right] - p(x)^2,
\end{aligned} \tag{15}
$$

where the second equality follows from (8) and (13). Moreover, since $\hat{p}^{\mathrm{IS}_0}(\cdot)$ is an ε-Uniformly Bounded IS Estimator of $p(\cdot)$, we have that (14) holds and hence absolute continuity of μ with respect to ν implies that $\mathbb{1}_{\{G(x,\xi)>0\}}L_x(\xi) \le \varepsilon$ w.p.1, which in turn implies that

$$
\mathbb{1}_{\{G(x,\xi)>0\}}L_x(\xi) \;\le\; \varepsilon \mathbb{1}_{\{G(x,\xi)>0\}} \qquad \text{w.p.1.}
$$

It follows that

$$\begin{aligned} N\mathrm{Var}(\hat{p}^{\mathrm{IS}_0}(x)) &= \mathbb{E}_\xi\left[\mathbb{1}_{\{G(x,\xi)>0\}}L_x(\xi)\right] - p(x)^2 \\ &\le \mathbb{E}_\xi\left[\mathbb{1}_{\{G(x,\xi)>0\}}\varepsilon\right] - p(x)^2 = \varepsilon p(x) - p(x)^2 \\ &\le \varepsilon(p(x) - p(x)^2) = N\varepsilon\mathrm{Var}(\hat{p}(x)), \end{aligned}$$

which finishes the proof. □

2.4 Consistency

When using sampling-based methods for stochastic optimization such as SAA, it is important to use consistent estimators because they provide theoretical guarantees that the optimal value and solution set of the approximate problem converge to their deterministic counterparts (see, e.g., [37] and [18] for a general discussion of that issue). Here, we extend the consistency results of [30] derived for chance-constrained problems to the setting of SAA with importance sampling and show that consistency is present if the significance levels of the SAA and true problems are the same.

The first step is to formulate an importance sampling version of the SAA problem (5), which we will refer to as SAA-IS. Following the notation established in the previous section, we define SAA-IS problem as

$$\min_{x\in X} h(x) \text{ s.t. } \hat{p}^{\mathrm{IS}}(x) \le \gamma, \tag{16}$$

where $\hat{p}^{\mathrm{IS}}(x)$ is the IS estimator defined in (7). Note that problem (16) is not stated for the more general estimator $\hat{p}^{\mathrm{IS}_0}(x)$ defined in (10), i.e., we use $I_x = I$ for all x; we will comment on that shortly.

Before showing the consistency results for problem (16) we need some definitions. Recall that a sequence $f_k(\cdot)$ of extended real valued functions is said to *epiconverge* to a function $f(\cdot)$, written $f_k \overset{\mathrm{e}}{\to} f$, if for any point x the following two conditions hold: (i) for any sequence x_k converging to x one has

$$\liminf_{k\to\infty} f_k(x_k) \ge f(x), \tag{17}$$

(ii) there exists a sequence x_k converging to x such that

$$\limsup_{k\to\infty} f_k(x_k) \le f(x). \tag{18}$$

We discuss now the consistency results. Note initially that by the strong law of large numbers (SLLN) (together with Lemma 2) we have that for any x, $\hat{p}^{\mathrm{IS}}(x)$ converges w.p.1 to $p(x)$. Proposition 6 below shows that, under mild assumptions, we actually have epiconvergence.

Proposition 6. *Let $G(x,\hat{\xi})$ be a Carathéodory function (i.e., continuous in x and measurable in $\hat{\xi}$). Then the functions $p(\cdot)$ and $\hat{p}^{\mathrm{IS}}(\cdot)$ are lower semicontinuous, and $\hat{p}^{\mathrm{IS}} \overset{\mathrm{e}}{\to} p$ w.p.1.*

Proof. To simplify the notation, define the functions $\phi(x,\hat{\xi}) := \mathbb{1}_{\{G(x,\hat{\xi})>0\}}$ and $\psi(x,\hat{\xi}) := \mathbb{1}_{\{G(x,\hat{\xi})>0\}}L(\hat{\xi})$. In [30, Proposition 2.1] it was shown that function ϕ is random lower semicontinuous. Combining this result with Corollary 14.46 in [33] we see that ψ is also random lower semicontinuous[1]. The lower semicontinuity of p was already established in [30]. The lower semicontinuity of $\hat{p}^{\mathrm{IS}}$ follows from Fatou's lemma, since for every $\bar{x} \in \mathbb{R}^n$ we have

$$\begin{aligned} \liminf_{x\to\bar{x}} \hat{p}^{\mathrm{IS}}(x) &= \liminf_{x\to\bar{x}} \frac{1}{N}\sum_{j=1}^{N} \psi(x,\hat{\xi}^j) \\ &\ge \frac{1}{N}\sum_{j=1}^{N} \liminf_{x\to\bar{x}} \psi(x,\hat{\xi}^j) \ge \frac{1}{N}\sum_{j=1}^{N} \psi(\bar{x},\hat{\xi}^j) = \hat{p}^{\mathrm{IS}}(\bar{x}). \end{aligned}$$

The epiconvergence $\hat{p}^{\mathrm{IS}} \overset{\mathrm{e}}{\to} p$ w.p.1 is then direct from [3, Theorem 2.3]. □

[1] Random lower semicontinuous functions are called normal integrands in [33].

It is important to observe the role of the likelihood function $L(\xi)$ in the above result. As mentioned earlier, the result does not apply to the more general likelihood function $L_x(\hat{\xi})$, the reason being that the dependence of L on x may destroy the lower semicontinuity required to show the epiconvergence in Proposition 6. By Proposition 2.2 in [30], we have that for $\gamma = \alpha$, under mild regularity conditions, $\hat{\vartheta}$ and $\hat{\Gamma}$ converge w.p.1 to their counterparts of the true problem.

The above discussion shows that consistency of optimal values and optimal solutions is preserved when using appropriate IS estimators. Note however that these results are stated with a continuous feasibility set X in mind. On the other hand, when X is finite there is no need to require the epiconvergence result in Proposition 6, since finiteness of X automatically implies uniform convergence of $\hat{p}^{\mathrm{IS}}$ to p. In fact, in that case convergence holds for the more general estimator $\hat{p}^{\mathrm{IS}_0}$, since the dependence of L_x on x does not preclude uniform convergence when X is finite. It is also worth noticing that the above convergence results extend readily to the case of multiple (but finitely many) chance constraints.

In the next section we describe a rare-event chance-constrained optimization model for a problem in telecommunications, which we solve using an SAA approach combined with importance sampling as laid out in this section. As we shall see, the use of IS techniques can be very effective for that problem, provided the IS distribution is properly chosen. After presenting the problem and the model formulation, we will discuss a particular choice of an IS distribution that yields promising results.

3 Joint routing and dimensioning problem for optical networks

We present now a problem arising in optical networks that illustrates the IS techniques discussed in Section 2.1. We start by describing the problem and its relevance and then we write explicit mixed integer programming formulations that approximate the problem using different IS estimators.

3.1 Problem description

Nowadays, the only technology that provides the high transmission speeds required in telecommunication are optical networks with Wavelength Division Multiplexing (WDM) technology. WDM allows transmission of multiple information channels (wavelengths) using a unique optical fiber. As a result, optical WDM networks are widely deployed as transport networks around the globe.

In an end-to-end dynamic optical WDM network, every time a connection request (i.e. a request to establish an optical channel from the source to the destination node) is generated, the resource allocation algorithm must find a route and an available optical channel in each link of that route. We assume a network equipped with wavelength conversion capability, that is, the resource allocation algorithm must only find a route with at least one optical channel available in each link of this route, regardless the wavelength of the channel. The algorithm in charge of finding a route is known as a routing algorithm.

When, in an end-to-end network, a connection is requested but there is not enough capacity in some link of the route assigned to the origin-destination pair, a *blockage* occurs. The performance of a routing algorithm in dynamic networks is typically measured in terms of the blocking probability. Routing algorithm A is better than routing algorithm B if it obtains a lower value of blocking probability. The blocking probability of a routing algorithm is in turn very much affected by the dimensioning of the network, that is, the number of wavelengths or capacity allocated to each link. If all network links are equipped with as many wavelengths as required in the worst case, then all routing algorithms would obtain zero blocking probability. Since wavelengths are costly resources, network operators aim at equipping the network with the minimum number of wavelengths per link such that the blocking obtained by the routing algorithms is below a given threshold. To do so, the typical approach is to first select a routing algorithm and then dimension the network in order to guarantee a given blocking probability according to the routes determined by the routing algorithm. Since routing affects dimensioning, such sequential approach tends to be suboptimal. To circumvent that problem, we will work with a model that solve both problems in a joint fashion.

We consider a network topology represented by a directed graph $G = (V, A)$ where V is the set of network vertices and A is the set of unidirectional arcs or links. The link capacity is measured in terms of number of wavelengths for link $a \in A$ and is denoted by w_a. Let $\mathcal{C} \subset V \times V$ be the set of connections that should be routed through the network. To facilitate the notation, we shall identify each connection with a number $c = 1, \ldots, C$, where $C = |\mathcal{C}|$. Each connection is associated to a source s_c and destination t_c nodes. We assume the traffic generated by source node s_c to destination node t_c is governed by an ON-OFF model [1]. For this traffic model, the source is assumed to transmit at the maximum bitrate and, therefore, in the long run the traffic load ρ_c corresponds to the fraction of time that connection c was transmitting data through the network. If the traffic load ρ_c is the same for every connection c we say that the traffic load is *homogeneous*, otherwise it is *heterogeneous*.

For source node s_c, we model the ON-OFF process as independent Bernoulli random variables $\xi = (\xi_c)_{c=1,\ldots,C}$, where $\xi_c = 1$ means connection c is in the ON state and thus the probability that such connection is requested is equal to ρ_c. Hence, the number of requested connections (referred to as active connections from now on) using a given link a is also a random variable. In this context, the blocking probability of link a is the probability that the number of active connections exceeds its capacity w_a.

Let α be the maximum blocking probability acceptable at every link. The *Joint Routing and Dimensioning* **(JRD)** problem of an optical network consists in finding routes r_c for each connection $c = 1, \ldots, C$ and capacities w_a for each arc $a \in A$ such that the minimum number of wavelengths is used. A common framework to model these kind of problems in telecommunications is by using multicommodity network flow models [28]. The chance-constrained JRD optimization problem can be stated as

$$\textbf{(CC-JRD)} \quad \min \sum_{a \in A} w_a$$

$$\mathcal{N} y_{c,\cdot} = d_c \quad \forall c = 1, \ldots, C \tag{19}$$

$$\mathbb{P}\left(\sum_{c=1}^{C} \xi_c y_{c,a} \le w_a \right) \ge 1 - \alpha \quad \forall a \in A \tag{20}$$

$$w_a \in \mathbb{N},\ y_{c,a} \in \{0, 1\} \quad \forall a \in A,\ \forall c = 1, \ldots, C \tag{21}$$

The integer variable w_a represents the capacity of arc a, while the binary variable $y_{c,a}$ takes value 1 if connection c is routed through arc a, and 0 otherwise. Equation (19) is an abbreviated form of the flow constraints to route each connection c from s_c to t_c. More specifically, we have

$$\sum_{a \in \delta^+(v)} y_{c,a} - \sum_{a \in \delta^-(v)} y_{c,a} = \begin{cases} 1 & v = s_c \\ -1 & v = t_c \\ 0 & \text{otherwise} \end{cases} \quad \forall v \in V$$

where $\delta^+(v)$ and $\delta^-(v)$ are respectively the set of arcs that start in node v and the set of arcs that end in node v. Finally, chance constraints (20) indicate that capacity constraints should be satisfied with probability at least $1 - \alpha$.

For the homogeneous case with traffic load ρ and for a fixed link a, the random variable $\sum_{c=1}^{C} \xi_c y_{c,a}$ follows a binomial distribution with parameters $\sum_{c=1}^{C} y_{c,a}$ and ρ. For a given value of $\sum_{c=1}^{C} y_{c,a}$, the minimum value of w_a such that constraint (20) is satisfied can be easily computed. Since the optimal value of w_a is a discrete step-increasing function on $\sum_{c=1}^{C} y_{c,a}$, problem **(CC-JRD)** can be reformulated as an integer program and solved to optimality (see [27, 41]). For the heterogeneous case, however, this argument is no longer valid, since the dimensioning of w_a depends not only on the number of connections routed through link a, but also on which connections are being routed — and enumerating all possibilities is clearly not practical. Nevertheless, the homogeneous case will serve as a benchmark for our IS approximations since the optimal solution is known. In the heterogeneous case we rely on bounds for the optimal value in order to evaluate the quality of the approximations obtained.

We close this section by emphasizing that, although the probability α is very small, we cannot simply approximate the chance-constrained problem by setting $\alpha = 0$ — a strategy that might seem appealing given that it is much easier to solve the problem when $\alpha = 0$. As it turns out, an optimal solution in the case of small α can differ considerably from the optimal solution with $\alpha = 0$. In fact, note that if $\alpha = 0$ then the optimal solution satisfies $w_a = \sum_{c=1}^{C} y_{c,a}$, so the objective function of **(CC-JRD)** is equal to minimize $\sum_{a \in A} \sum_{c=1}^{C} y_{c,a}$. Hence, the optimal solutions correspond to route each connection using its corresponding shortest path. Consider for example the case of ring topologies with 7 and 9 nodes. In that case, each link has either 6 (in case of 7 nodes) or 10 (in case of 9 nodes) connections routed through, so the optimal objective values are 84 and 180, respectively. But, if $\alpha = 10^{-6}$ then a capacity at each link of 5 (in case of 7 nodes) and 7 (in case of 9 nodes) is enough, resulting in objective values of $5 \cdot 14 = 70$ and $7 \cdot 18 = 126$, respectively, which are significant smaller than the optimal values in the $\alpha = 0$ case. This example shows that allowing for a small amount of violation, e.g 10^{-6}, yields substantially different solutions from the 100% reliable case. More interestingly, none of these solutions are optimal for $\alpha = 10^{-6}$. The optimal values for these instances are 68 and 117, respectively, and the optimal routing differs considerably from the shortest path routing (see Figure 1).

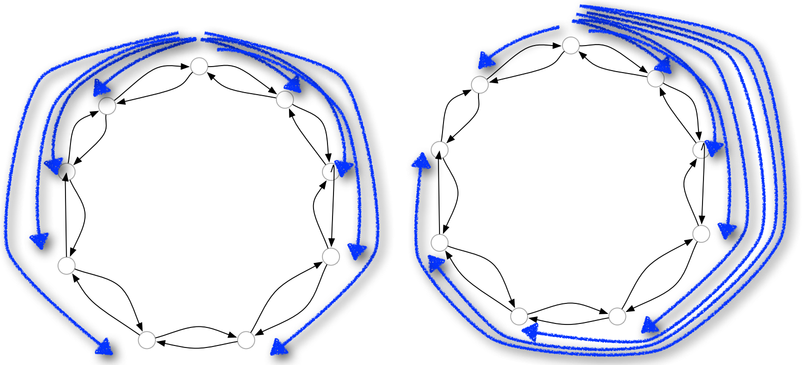

Figure 1: Example of two routings for the ring of 9 nodes for $\alpha = 10^{-6}$. In the left case, each arc satisfies $\sum_{c=1}^{C} y_{c,a} = 10$, requiring a minimum capacity of 7, and resulting in a objective value of 126. In the right case, each clockwise (counterclockwise) arc satisfies $\sum_{c=1}^{C} y_{c,a} = 28$ ($\sum_{c=1}^{C} y_{c,a} = 1$) with capacity 12 (1) resulting in a objective value of 117, which is the optimal solution.

3.2 Mixed integer programming formulations

Let y denote the vector $(y_{c,a})_{a \in A,\, c=1,\ldots,C}$ and w denote the vector $(w_a)_{a \in A}$. We shall use x to denote the joint vector (y, w), and ξ to denote the random vector $(\xi_c)_{c=1,\ldots,C}$. Let $\xi^1, \ldots, \xi^N$ be an i.i.d sample from the distribution of the random vector ξ. Using the notation of Section 2, we have that the chance constraints of problem **(CC-JRD)** can be written as

$$\mathbb{P}\big\{G_a(x,\xi) \leq 0\big\} \geq 1 - \alpha \ , \quad \text{with} \ \ G_a(x,\xi) = \sum_{c=1}^{C} \xi_c y_{c,a} - w_a \ . \tag{22}$$

Following equation (4), one estimator is

$$\hat{p}_a(x) = \frac{1}{N} \sum_{s=1}^{N} \mathbb{1}_{\{G_a(x,\xi^s)>0\}} \ . \tag{23}$$

We are interested in blocking probabilities that are very small, say of order 10^{-6} , which would lead to intractable sample sizes if $\hat{p}_a(x)$ were used in an SAA formulation. In order to construct IS estimators, we choose the IS distribution in a parametric way as follows: consider independent random variables $(\hat{\xi}_c)_{c=1,\ldots,C}$

with Bernoulli distribution with parameters $(\hat{\rho}_c)_{c=1,\dots,C}$. The likelihood ratio is

$$L(\hat{\xi}) = \prod_{c=1}^{C} \left(\frac{\rho_c}{\hat{\rho}_c}\right)^{\hat{\xi}_c} \left(\frac{1-\rho_c}{1-\hat{\rho}_c}\right)^{1-\hat{\xi}_c} , \tag{24}$$

and the IS estimator can be written as follows:

$$\hat{p}_a^{\mathrm{IS}}(x) = \frac{1}{N}\sum_{s=1}^{N} \mathbb{1}_{\{\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a} > w_a\}} L(\hat{\xi}) . \tag{25}$$

A third unbiased estimator can be constructed by noting that function G_a defined in (22) is $I_{a,y}$-determined, for $I_{a,y} = \{c = 1,\dots,C \ : \ y_{c,a} = 1\}$. Then, the modified likelihood ratio defined in (9) can be written as

$$L_{y,a}(\hat{\xi}) \ := \ \prod_{c=1}^{C} \left(\frac{\rho_c}{\hat{\rho}_c}\right)^{y_{c,a}\hat{\xi}_c} \left(\frac{1-\rho_c}{1-\hat{\rho}_c}\right)^{y_{c,a}(1-\hat{\xi}_c)} , \tag{26}$$

and consequently the estimator in (10) becomes

$$\hat{p}_a^{\mathrm{IS}_0}(x) \ := \ \frac{1}{N}\sum_{s=1}^{N} \mathbb{1}_{\{\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a} > w_a\}} L_{y,a}(\hat{\xi}) . \tag{27}$$

From (15), the variance of this estimator is

$$\begin{aligned} \mathrm{Var}(\hat{p}_a^{\mathrm{IS}_0}(x)) \ &= \ \frac{1}{N}\mathbb{E}_\xi \left[\mathbb{1}_{\{\sum_{c=1}^{C} \xi_c^s y_{c,a} > w_a\}} L_{y,a}(\xi)\right] - \frac{p(x)^2}{N} \\ &= \ \frac{1}{N}\mathbb{E}_\xi \left[\mathbb{1}_{\{\sum_{c=1}^{C} \xi_c y_c > w\}} \prod_{c=1}^{C} \left(\frac{\rho_c}{\hat{\rho}_c}\right)^{\xi_c y_c} \left(\frac{1-\rho_c}{1-\hat{\rho}_c}\right)^{(1-\xi_c)y_c}\right] - \frac{p(x)^2}{N} . \end{aligned} \tag{28}$$

From Lemma 2 we have that $\hat{p}_a^{\mathrm{IS}_0}(x)$ is an unbiased estimator of $p_a(x)$ and its variance is less than or equal to the variance of $\hat{p}_a^{\mathrm{IS}}(x)$ in (25).

In order to use these estimators in the formulation of **(CC-JRD)**, we only need to replace the chance constraint (20) by

$$\begin{aligned} \sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a} &\le w_a + z_{a,s}\mathcal{M} \quad \forall a \in A, \forall s = 1,\dots,N, & (29) \\ \sum_{s=1}^{N} L(\hat{\xi}^s) z_{a,s} &\le \alpha \quad \forall a \in A, & (30) \\ z_{a,s} &\in \{0,1\} \quad \forall a \in A, \forall s = 1,\dots,N. \end{aligned}$$

The variable $z_{a,s}$ indicates whether the capacity constraint on arc a is violated in sample s, which is captured by the big-$\mathcal{M}$ constraint (29). Finally, equation (30) approximates the chance constraint (20) as follows: for $\hat{p}_a$ we use $L(\hat{\xi}^s) \equiv 1$, and for $\hat{p}_a^{\mathrm{IS}}$ we use $L(\hat{\xi})$ as defined in equation (24).

The formulation of the problem for the estimator $\hat{p}_a^{\mathrm{IS}_0}$ is more delicate since the likelihood ratio $L_{y,a}(\hat{\xi})$ defined in (26) depends nonlinearly on the decision variables. Nevertheless, we can formulate an equivalent MIP problem for the homogeneous case (all parameters ρ_c equal) as follows:

$$\min \sum_{a\in A} w_a \tag{31}$$

$$\mathcal{N}y_{c,\cdot} = d_c \quad \forall c = 1,\dots,C \tag{32}$$

$$\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a} \le w_a + z_{a,s}\mathcal{M} \quad \forall a \in A, \forall s = 1,\dots,N \tag{33}$$

$$\sum_{c=1}^{C} y_{c,a} = \sum_{k=0}^{C} k v_{a,k} \quad \forall a \in A \tag{34}$$

$$\sum_{k=0}^{C} v_{a,k} = 1 \quad \forall a \in A \tag{35}$$

$$L_{a,s} \ge \sum_{k=1}^{C} F_a(s,k) v_{a,k} - (1 - z_{a,s})\mathcal{M} \quad \forall a \in A, \forall s = 1,\dots,N \tag{36}$$

$$\sum_{s=1}^{N} L_{a,s} \le \alpha N \quad \forall a \in A \tag{37}$$

$$w_a \in \mathbb{N},\ y_{c,a} \in \{0,1\}, \quad \forall a \in A, \forall c = 1,\dots,C \tag{38}$$

$$L_{a,s} \ge 0,\ z_{a,s} \in \{0,1\}, v_{a,k} \in \{0,1\} \quad \forall a \in A, \forall s = 1,\dots,N \tag{39}$$

Decision variables y and w, constraint (32) and the objective function (31) are the same as the ones defined in the original formulation **(CC-JRD)**. Decision variables z and constraint (33) are as in (29). The binary decision variable $v_{a,k}$ is equal to one if and only if $\sum_{c=1}^{C} y_{c,a} = k$, which is obtained by constraints (34) and (35). The formulation is such that the continuous decision variable $L_{a,s}$ represents $\mathbb{1}_{\{\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a} > w_a\}} L_{y,a}(\hat{\xi}^s)$ for the arc a and sample s. Indeed, the likelihood ratio for sample s when k connections are routed through a link can be pre-computed as

$$F_a(s,k) \ := \ \prod_{i=1}^{k} \left(\frac{\rho}{\hat{\rho}}\right)^{\hat{\xi}_i^s} \left(\frac{1-\rho}{1-\hat{\rho}}\right)^{(1-\hat{\xi}_i^s)} . \tag{40}$$

It is important to observe that, since all connections have the same probability and scenarios are sampled independently for each arc, this value does not depend on *which* connections are routed through a link, only on the *number* of such connections. We can compute (40) using the first k components of the sample s. Constraint (36) ensures that if k connections are routed through link a (i.e. $v_{a,k} = 1$) and if the corresponding capacity constraint is violated under sample s (i.e. $z_{a,s} = 1$) then $L_{a,s} \ge F_a(s,k)$. Together with constraint (37) we have the desired representation.

In the heterogeneous case the likelihood ratio depends on which — and not just how many — connections are routed through a link a, so the above formulation cannot be used in that setting. However, it is possible to construct a MIP formulation that approximates the nonlinearity of the likelihood ratio function. We describe that formulation next.

Recall that the term $L_{y,a}(\hat{\xi})$ from equation (26) can be written as

$$\begin{aligned} L_{y,a}(\hat{\xi}) &= \prod_{c=1}^{C} \left(\frac{\rho_c}{\hat{\rho}_c}\right)^{\hat{\xi}_c^s y_{c,a}} \left(\frac{1-\rho_c}{1-\hat{\rho}_c}\right)^{(1-\hat{\xi}_c^s) y_{c,a}} \\ &= \prod_{c=1}^{C} \left(\frac{\rho_c}{\hat{\rho}_c}\frac{1-\hat{\rho}_c}{1-\rho_c}\right)^{\hat{\xi}_c^s y_{c,a}} \prod_{c=1}^{C} \left(\frac{1-\rho_c}{1-\hat{\rho}_c}\right)^{y_{c,a}} . \end{aligned}$$

Hence, constraint $\hat{p}_a^{\mathrm{IS}_0}(x) \le \alpha$ is equivalent to

$$\frac{1}{N} \sum_{s=1}^{N} \left(\prod_{c=1}^{C} \left(\frac{\rho_c}{\hat{\rho}_c}\frac{1-\hat{\rho}_c}{1-\rho_c}\right)^{\hat{\xi}_c^s y_{c,a}} \right) \mathbb{1}_{\{\sum\limits_{c=1}^{C} \hat{\xi}_c^s y_{c,a} > w_a\}} \le \alpha \prod_{c=1}^{C} \left(\frac{1-\hat{\rho}_c}{1-\rho_c}\right)^{y_{c,a}} . \tag{41}$$

We will show in Section 4 that we can build an appropriate important sampling distribution satisfying $\frac{\rho_c}{\hat{\rho}_c}\frac{1-\hat{\rho}_c}{1-\rho_c} = e^{\lambda_x^*}$ for all $c = 1, \ldots, C$, where $\lambda_x^* > 0$ is the root of a certain equation (see Theorem 9). Hence, for such IS distribution the left-hand size of the equation only depends on the number of active connections in sample s (i.e., $\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a}$), but not on which of them are active. However, this is not possible for the right-hand size of (41), so we need to rely on an approximation of the non-linear term, replacing $\prod_{c=1}^{C}\left(\frac{1-\hat{\rho}_c}{1-\rho_c}\right)^{y_{c,a}}$ with $G_a\left(\sum_{c=1}^{C} y_{c,a}\right)$, where

$$G_a(k) := \max_{C^0 \subseteq \{1,\ldots,C\}:|C^0|=k} \prod_{c \in C^0} \frac{1-\hat{\rho}_c}{1-\rho_c} \; . \tag{42}$$

We argue that the above expression can be computed easily. In fact, by writing $\hat{\rho}_c$ as a function of ρ_c it is easy to check that such a function is concave and increasing, so we have that

$$\frac{1-\hat{\rho}_i}{1-\rho_i} \;=\; e^{\lambda_x^*}\frac{\hat{\rho}_i}{\rho_i} \;\geq\; e^{\lambda_x^*}\frac{\hat{\rho}_j}{\rho_j} \;=\; \frac{1-\hat{\rho}_j}{1-\rho_j}$$

whenever $\rho_i \geq \rho_j$. Thus, assuming that $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_C$— which can be done without loss of generality —it follows that

$$G_a(k) = \prod_{i=1}^{k} \frac{1-\hat{\rho}_i}{1-\rho_i}. \tag{43}$$

Additionally, note that by using this approximation constraint (41) now depends only on the number of connections routed through arc a (i.e. $\sum_{c=1}^{C} y_{c,a}$) and on the number of these connections that are active in each sample (i.e. $\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a}$), so we can formulate a model which is similar to the one constructed for the homogeneous case.

$$\min \sum_{a \in A} w_a \tag{44}$$

$$\mathcal{N} y^c = d^c \quad \forall c = 1, \ldots, C \tag{45}$$

$$\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a} \leq w_a + \sum_{k=1}^{C} k \cdot u_{a,s,k} \;\; \forall a \in A, \forall s = 1, \ldots, S \tag{46}$$

$$\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a} \geq \sum_{k=1}^{C} k \cdot u_{a,s,k} \;\; \forall a \in A, \forall s = 1, \ldots, S \tag{47}$$

$$\sum_{s=1}^{N} \sum_{k=1}^{C} e^{-k\lambda} u_{a,s,k} \leq \alpha N \sum_{k=0}^{C} G_a(k) v_{a,k} \;\; \forall a \in A \tag{48}$$

$$\sum_{c=1}^{C} y_{c,a} = \sum_{k=0}^{C} k v_{a,k} \;\; \forall a \in A \tag{49}$$

$$\sum_{k=0}^{C} v_{a,k} = 1 \;\; \forall a \in A \tag{50}$$

$$\sum_{k=1}^{C} u_{a,s,k} \leq 1 \;\; \forall a \in A, \forall s = 1, \ldots, S \tag{51}$$

$$w_a \in \mathbb{N}, \;\; y_{c,a} \in \{0,1\}, u_{a,s,k} \in \{0,1\}, v_{a,k} \in \{0,1\} \;\; \forall a \in A, \forall c = 1, \ldots, C, \forall s = 1, \ldots, S \tag{52}$$

Binary variables $v_{a,k}$, together with equations (50) and (49), satisfy that $v_{a,k} = 1$ if and only if $\sum_{c=1}^{C} y_{c,a} = k$. The role of binary variables u is explained in the following lemma

Lemma 7. *Let (x, w, u, v) be an optimal solution of previous formulation, then there exist an optimal solution $(x, w, \hat{u}, v)$ such that*

1. $\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a} \leq w_a$ *if and only if* $\hat{u}_{a,s,k} = 0$ *for all* $k = 1, \ldots, C$.
2. *if* $\hat{u}_{a,s,k} = 1$ *then* $\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a} = k$,

Hence,

$$\sum_{k=1}^{C} e^{-k\lambda} \hat{u}_{a,s,k} = e^{\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a}} \mathbb{1}_{(\sum\limits_{c=1}^{C} \hat{\xi}_c^s y_{c,a} > w_a)}$$

Proof. First, note that constraint (46) impose that if $u_{a,s,k} = 0$ for all k then $\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a} \leq w_a$. Suppose that $\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a} \leq w_a$ but $u_{a,s,k'} = 1$ for some k'. It is easy to see that defining $\hat{u}_{a,s,k} = 0$ for all k and $\hat{u} = u$ for the other variables, then $\hat{u}$ also satisfy Equation (46) and (47), and since $\hat{u} \leq u$ then it also satisfy Equations (51) and (48), hence $(x, w, \hat{u}, v)$ is also optimal. Repeating this procedure is easy to see that we obtain a solution that satisfies condition (1). For the second condition, suppose that $u_{a,s,k} = 1$ for some k but $\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a} > k$. Let $\hat{k} = \sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a}$ and define $\hat{u}_{a,s,\hat{k}} = 1$, $\hat{u}_{a,s,k} = 0 \; \forall k \neq \hat{k}$ and $\hat{u} = u$ for the other variables. By definition, $(w, x, \hat{u}, v)$ satisfies (46) and (51), and since $\hat{k} > k$ then it also satisfies (46). On the other hand, since $\lambda > 0$ then $e^{-\lambda k} > e^{-\lambda \hat{k}}$ so it also satisfy (48) and then $(x, w, \hat{u}, v)$ is also optimal. Repeating this procedure is easy to see that we obtain a solution that satisfies condition (2). □

Lemma 7 shows that the optimal solution (y, w) of this MIP formulation satisfies

$$\frac{1}{N} \sum_{s=1}^{S} e^{\sum_{c=1}^{C} \hat{\xi}_c^s y_{c,a}} \mathbb{1}_{(\sum\limits_{c \in C} \hat{\xi}_c^s y_{c,a} > w_a)} \leq \alpha \cdot G_a \left(\sum_{c=1}^{C} y_{c,a} \right) \qquad \forall a \in A$$

which is the desired approximation of equation $\hat{p}_a^{\mathrm{IS}_0} \leq \alpha$. By construction the proposed scheme is an outer approximation of constraint (41). Of course, by replacing max with min in (42) we can easily obtain an inner approximation in a similar fashion.

4 Choosing the importance sampling estimator

In this section we discuss how to choose an IS distribution that ensures ε-uniform variance reduction over a subset X_0 of the set of points $x = (y, w)$ satisfying (19) and (21), which we will denote by X henceforth. As we shall see shortly, it is not necessary — in fact, it is not even desirable — to attempt to obtain uniform reduction over the entire set X.

Since we want to estimate the blocking probability for each arc a in A, we can choose a different IS distribution for each arc. Nevertheless, as our analysis is the same for each arc we drop the subscripts a to simplify the notation.

Recall that $\xi = (\xi_c)_{c=1}^{C}$ is a vector of Bernoulli random variables modeling the presence or not of each connection c, with $\mathbb{E}[\xi_c] = \rho_c$. Throughout this section we will assume without loss of generality that the connections are numbered in such a way that the corresponding rates ρ_c satisfy $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_C$. As discussed before, the probability we want to estimate is

$$p(x) = \mathbb{P} \left\{ \sum_{c=1}^{C} y_c \xi_c > w \right\} = \sum_{\xi \in \{0,1\}^C} \mathbb{1}_{\{\sum_{c=1}^{C} y_c \xi_c > w\}} \prod_{c=1}^{C} \rho_c^{y_c \xi_c} (1 - \rho_c)^{y_c (1 - \xi_c)},$$

and to do that we will use a product of Bernoulli distributions with parameters $(\hat{\rho}_c)_{c=1}^{C}$ as the IS distribution, which results in the estimator $\hat{p}^{\mathrm{IS}_0}(x)$ defined in (27). The problem then is to find the optimal value of the parameters $(\hat{\rho}_c)_{c=1}^{C}$ such that the variance of IS estimator is uniformly minimized over an appropriate set X_0, that is, we will look for the best ε-uniform variance reduction over that set. We first discuss how to choose good IS parameters for a given solution $x = (y, w) \in X$, and then we combine the results to choose an overall IS distribution for the whole set.

4.1 Choosing the IS parameters for a fixed solution (y, w).

Given $x = (y, w) \in X$, let us denote $|y| = \sum_{c=1}^{C} y_c$. In order to reduce the variance of $\hat{p}^{\mathrm{IS}_0}(x)$, from expression (28) we should minimize the term

$$\begin{aligned}
&\mathbb{E}_\xi \left[\mathbb{1}_{\{\sum_{c=1}^{C} y_c\xi_c > w\}} \prod_{c=1}^{C} \left(\frac{\rho_c}{\hat{\rho}_c}\right)^{y_c\xi_c} \left(\frac{1-\rho_c}{1-\hat{\rho}_c}\right)^{y_c(1-\xi_c)} \right] \\
=&\mathbb{E}_\xi \left[\mathbb{1}_{\{\sum_{c=1}^{C} y_c\xi_c > w\}} \prod_{c=1}^{C} \left(\frac{\rho_c(1-\hat{\rho}_c)}{\hat{\rho}_c(1-\rho_c)}\right)^{y_c\xi_c} \prod_{c=1}^{C} \left(\frac{1-\rho_c}{1-\hat{\rho}_c}\right)^{y_c} \right].
\end{aligned} \tag{53}$$

For each $c \in C$, let

$$\lambda_c \ := \ \log\left(\frac{1/\rho_c - 1}{1/\hat{\rho}_c - 1}\right) \ = \ \log\left(\frac{\hat{\rho}_c(1-\rho_c)}{\rho_c(1-\hat{\rho}_c)}\right). \tag{54}$$

Note that $\lambda_c \geq 0$ when $\hat{\rho}_c \geq \rho_c$ — which is case since the IS distribution works by inflating the connection rates so the event $\sum_{c=1}^{C} y_c\xi_c > w$ happens more often. Moreover, a bit of algebra shows that $\frac{1-\rho_c}{1-\hat{\rho}_c} = e^{\lambda_c}\rho_c + (1-\rho_c)$. It follows that we can write expression (53) as

$$A_x(\vec{\lambda}) \ := \ \mathbb{E}_\xi \left[\mathbb{1}_{\{\sum_{c=1}^{C} y_c\xi_c > w\}} e^{-\sum_{c\in C} \lambda_c y_c \xi_c} \prod_{c=1}^{C} \left(e^{\lambda_c}\rho_c + (1-\rho_c)\right)^{y_c} \right] . \tag{55}$$

Minimizing $A_x(\vec{\lambda})$ requires solving a multidimensional stochastic nonlinear problem, which is a difficult task. Alternatively, our approach is to minimize the largest term inside the expected value, that is

$$B_x(\vec{\lambda}) \ := \ \max_{\xi\,:\,\sum_{c=1}^{C} y_c\xi_c > w} \left\{ e^{-\sum_{c\in C} \lambda_c y_c \xi_c} \prod_{c=1}^{C} \left(e^{\lambda_c}\rho_c + (1-\rho_c)\right)^{y_c} \right\}. \tag{56}$$

It is easy to see that $B_x(\vec{\lambda}) > 0$. Moreover, from Proposition 5 we obtain that $\mathrm{Var}\left(\hat{p}^{\mathrm{IS}_0}(x)\right) \ \leq \ B_x(\vec{\lambda})\mathrm{Var}\left(\hat{p}(x)\right)$ provided that $B_x(\vec{\lambda}) \leq 1$. Thus, we can reduce the variance of the IS estimator $\hat{p}^{\mathrm{IS}_0}(x)$ by minimizing $B_x(\vec{\lambda})$ over $\vec{\lambda} \geq 0$. By using KKT conditions, we obtain a characterization for the optimal $\vec{\lambda}$, which is given in the following Theorem 9. Furthermore, by using (54) we can compute the corresponding parameters $\hat{\rho}_c = \hat{\rho}_c(\lambda_c)$ as

$$\hat{\rho}_c(\lambda_c) \ = \ \frac{e^{\lambda_c}\rho_c}{e^{\lambda_c}\rho_c + (1-\rho_c)},$$

and we define $\hat{\rho}_c(\infty) := \lim_{\lambda_c\to\infty} \hat{\rho}_c(\lambda_c) = 1$.

Before we state the Theorem we present an auxiliary result — of independent interest — which gives a lower bound on the probability that a random variable defined as a sum of independent Bernoulli random variables exceeds its mean. Unlike upper bound inequalities such as Chebyshev's, which are valid for any distribution, lower bound inequalities typically require exploiting characteristics of the underlying distributions as we do below. The proposition will be used to verify the assumptions of Theorem 9.

Proposition 8. *Let $\zeta_1, \ldots, \zeta_m$ be $m \geq 1$ independent Bernoulli random variables with $\mathbb{P}\{\zeta_i = 1\} = p_i$, and suppose that $0 < p_i < 1$ for all i. Let $Z := \sum_{i=1}^{m} \zeta_i$, and define $\delta := \min_i p_i(1-p_i) > 0$. Then, we have*

$$\mathbb{P}\left\{Z > \mathbb{E}[Z]\right\} \ > \ \frac{\delta}{2m} \ . \tag{57}$$

Proof. Let $u : [0, m] \mapsto \mathbb{R}$ be the function defined as $u(t) := m^2 - t^2$. Since $u(\cdot)$ is nonnegative and decreasing on $[0, m]$, we have that

$$\mathbb{P}\left\{Z \leq \mathbb{E}[Z]\right\} \ = \ \mathbb{P}\left\{u(Z) \geq u(\mathbb{E}[Z])\right\}$$

$$
\begin{aligned}
&= \mathbb{P}\left\{m^2 - Z^2 \geq m^2 - (\mathbb{E}[Z])^2\right\} \\
&\leq \frac{\mathbb{E}\left[m^2 - Z^2\right]}{m^2 - (\mathbb{E}[Z])^2},
\end{aligned}
$$

where the last inequality follows from Markov's inequality. Thus, we have

$$
\begin{aligned}
\mathbb{P}\{Z > \mathbb{E}[Z]\} \;\geq\; 1 - \frac{\mathbb{E}\left[m^2 - Z^2\right]}{m^2 - (\mathbb{E}[Z])^2} \;&=\; \frac{\mathbb{E}\left[Z^2\right] - (\mathbb{E}[Z])^2}{m^2 - (\mathbb{E}[Z])^2} \\
&= \frac{\mathrm{Var}(Z)}{(m + \mathbb{E}[Z])(m - \mathbb{E}[Z])}.
\end{aligned} \tag{58}
$$

Next, notice that independence of $\{\zeta_i\}$ implies that $\mathrm{Var}(Z) = \sum_{i=1}^m p_i(1-p_i)$. Moreover, since $0 < \mathbb{E}[Z] < m$ we have that $m + \mathbb{E}[Z] < 2m$, $m - \mathbb{E}[Z] < m$ and thus from (58) we have that

$$
\mathbb{P}\{Z > \mathbb{E}[Z]\} \;>\; \frac{\sum_{i=1}^m p_i(1-p_i)}{2m^2} \;\geq\; \frac{\delta m}{2m^2} \;=\; \frac{\delta}{2m}.
$$

□

Theorem 9. *Suppose that $0 < \rho_c < 1$ for all $c = 1, \ldots, C$. Let $x = (y, w)$ be such that $w \in \mathbb{N}$ satisfies $\sum_{c=1}^C \rho_c y_c < w \leq \sum_{c=1}^C y_c - 1$. Then the function $B_x(\vec{\lambda})$ is convex and there exists $\lambda_x^* \in \mathbb{R}_+ \cup \{\infty\}$ such that the vector $\vec{\lambda}$ defined as $\lambda_c = \lambda_x^*\ \forall c \in C$ minimizes $B_x(\vec{\lambda})$. If $w = \sum_{c=1}^C y_c - 1$, then the optimal λ_x^* is $\lambda_x^* = \infty$ and $\hat{\rho}_c(\lambda_x^*) = 1$; otherwise, λ_x^* and $\hat{\rho}_c(\lambda_x^*)$ satisfy*

$$
\sum_{c=1}^C \hat{\rho}_c(\lambda_x^*) y_c = w + 1 \qquad \text{and} \qquad \hat{\rho}_c(\lambda_x^*) = \frac{e^{\lambda_x^*}\rho_c}{e^{\lambda_x^*}\rho_c + (1-\rho_c)}\ .
$$

To prove the theorem, we need the following lemma, the proof of which is shown after the proof of the theorem.

Lemma 10. *For $n \geq 1$, let ρ_i, $i = 1, \ldots, n$ be numbers such that $\rho_i \in (0,1)$ and $\rho_1 \geq \rho_2 \geq \ldots \geq \rho_n$. Given an integer w such that $0 \leq w \leq n-1$, consider problem (P) defined as follows:*

$$
\min_{\lambda \in \mathbb{R}_+^n} \max_{\substack{z_i \in \{0,1\}^n \\ \sum_i z_i = w+1}} -\sum_{i=1}^n z_i \lambda_i + \sum_{i=1}^n \log(e^{\lambda_i}\rho_i + (1-\rho_i)). \tag{P}
$$

Then, there exists an optimal solution to (P) that satisfies $\lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$.

of Theorem 9. Let $n = \sum_{c=1}^C y_c$. Without loss of generality, let us assume for the sake of simplifying notation that the set $\{c\ :\ y_c = 1\}$ corresponds to $\{1, \ldots, n\}$. Since the log function is increasing, we have that

$$
\log(B_x(\vec{\lambda})) \;=\; \max_{\substack{z_i \in \{0,1\}^n \\ \sum_i z_i = w+1}} -\sum_{i=1}^n z_i \lambda_i + \sum_{i=1}^n \log(e^{\lambda_i}\rho_i + (1-\rho_i))
$$

By Lemma 10, minimizing $\log(B_x(\vec{\lambda}))$ amounts to solving the following problem:

$$
\min_{\vec{\lambda} \in \mathbb{R}^n} \psi(\vec{\lambda}) := -\sum_{i=1}^{w+1} \lambda_i + \sum_{i=1}^n \log(e^{\lambda_i}\rho_i + (1-\rho_i)) \tag{59}
$$

$$
\lambda_i \leq \lambda_{i+1} \quad i = 1 \ldots n-1 \tag{60}
$$

$$
\lambda_1 \geq 0 \tag{61}
$$

Note that the objective function of the above problem is strictly convex in $\vec{\lambda}$. In fact, its second derivatives are

$$\frac{\partial^2 \psi}{\partial \lambda_i^2} = \frac{e^{\lambda_i}\rho_i(1-\rho_i)}{(e^{\lambda_i}\rho_i + (1-\rho_i))^2} > 0, \qquad \frac{\partial^2 \psi}{\partial \lambda_i \partial \lambda_j} = 0.$$

Since $B_x(\vec{\lambda}) = \exp(\log(B_x(\vec{\lambda}))$ and $\log(B_x(\vec{\lambda}))$ is convex — though not strictly convex due to the components λ_c such that $y_c = 0$ — it follows that B_x is convex in $\vec{\lambda}$. Of course, the components λ_c such that $y_c = 0$ do not affect the value of $B_x(\vec{\lambda})$.

Suppose first that $w = n-1$. Then, the first derivative of the objective function in (59) is given by

$$\frac{\partial \psi}{\partial \lambda_i} \;=\; -1 + \frac{e^{\lambda_i}\rho_i}{e^{\lambda_i} + (1-\rho_i)}, \quad i = 1, \ldots, n,$$

so we see that $\lim_{\vec{\lambda}\to\infty} \nabla\psi(\vec{\lambda}) = 0$. Notice that we can in particular interpret $\lim_{\vec{\lambda}\to\infty}$ as $\lim_{\lambda\to\infty}$ with $\lambda_i = \lambda$. That is, in that case the optimal solution of (59)-(61) is $\lambda_i = \infty$, $i = 1, \ldots, n$.

Consider now the case $w < n-1$. We will show that problem (59)-(61) has a unique optimal solution, which can be found by writing the Karush-Kuhn-Tucker conditions as follows:

$$-1_{(i \le w+1)} + \frac{e^{\lambda_i}\rho_i}{e^{\lambda_i}\rho_i + (1-\rho_i)} + \mu_i - \mu_{i-1} = 0 \qquad i = 1 \ldots n-1 \tag{62}$$

$$\frac{e^{\lambda_n}\rho_n}{e^{\lambda_n}\rho_n + (1-\rho_n)} - \mu_{n-1} = 0 \tag{63}$$

$$\mu_i(\lambda_{i+1} - \lambda_i) = 0 \qquad i = 1 \ldots n-1 \tag{64}$$

$$\mu_0 \lambda_1 = 0 \tag{65}$$

$$\mu_i \ge 0 \qquad i = 0 \ldots n-1 \tag{66}$$

where $\vec{\mu} = (\mu_i)$ is the vector of Lagrangean multipliers of constraints (60) and μ_0 is the Lagrangean multiplier of constraint (61).

Consider now a particular choice of vectors $\vec{\mu}$ and $\vec{\lambda}$ defined as follows. All components of $\vec{\lambda}$ are identical, with $\lambda_i = \lambda^*$, where $\lambda^* \in \mathbb{R}_+$ solves the equation

$$\varphi(\lambda^*) := \sum_{i=1}^{n} \frac{e^{\lambda^*}\rho_i}{e^{\lambda^*}\rho_i + (1-\rho_i)} \;=\; w+1. \tag{67}$$

Note that we can always find such λ^*, since the function $\varphi(\lambda)$ is continuous and increasing, and

$$\varphi(0) \;=\; \sum_{i=1}^{n} \rho_i \;< w \;<\; w+1 \tag{68}$$

$$\lim_{\lambda\to\infty} \varphi(\lambda) \;=\; n \;>\; w+1. \tag{69}$$

The inequalities in (68) follow from the assumptions of the theorem on w and the fact that we are analyzing the case $w < n-1$. The components of $\vec{\mu}$ are defined as

$$\mu_0 := 0 \tag{70}$$

$$\mu_i := \min\{i, w+1\} - \sum_{j=1}^{i} \frac{e^{\lambda^*}\rho_j}{e^{\lambda^*}\rho_j + (1-\rho_j)} \quad i = 1, \ldots, n-1. \tag{71}$$

We claim that $\vec{\mu}$ and $\vec{\lambda}$ satisfy the KKT conditions (62)-(66) laid out above. To see that, observe that equations (70)-(71) imply (62). Equation (63) follows from (67), since we have

$$\frac{e^{\lambda^*}\rho_n}{e^{\lambda^*}\rho_n + (1-\rho_n)} \;=\; w+1 - \sum_{i=1}^{n-1} \frac{e^{\lambda^*}\rho_i}{e^{\lambda^*}\rho_i + (1-\rho_i)}$$

and the latter term coincides with μ_{n-1} defined in (71). Equations (64) and (65) are trivially satisfied. Finally, we show that (66) holds, with strict inequality if $i \geq 1$. Indeed, (67) implies that

$$\sum_{j=1}^{i} \frac{e^{\lambda^*}\rho_j}{e^{\lambda^*}\rho_j + (1-\rho_j)} \; < \; w+1 \quad i = 1, \ldots, n-1$$

and clearly have

$$\sum_{j=1}^{i} \frac{e^{\lambda^*}\rho_j}{e^{\lambda^*}\rho_j + (1-\rho_j)} \; < \; i \quad i = 1, \ldots, n$$

as each term in the summand is less than 1. □

of Lemma 10. Suppose that $\vec{\lambda} = (\lambda_1, \ldots, \lambda_n)$ is an optimal solution and there exists some $j < n$ such that $\lambda_j > \lambda_{j+1}$. We will show that $\vec{\bar{\lambda}}$ defined as $\bar{\lambda}_j = \lambda_{j+1}$, $\bar{\lambda}_{j+1} = \lambda_j$ and $\bar{\lambda}_i = \lambda_i$ for $i \neq \{j, j+1\}$ has no worse objective function than $\vec{\lambda}$. Let Δ be defined as the difference in objective function between $\vec{\lambda}$ and $\vec{\bar{\lambda}}$, i.e.,

$$\Delta = \max_{\substack{z_i \in \{0,1\}^n \\ \sum_i z_i = w+1}} -\sum_{i=1}^{n} z_i\lambda_i + \sum_{i=1}^{n} \log(e^{\lambda_i}\rho_i + (1-\rho_i)) \tag{72}$$

$$-\left(\max_{\substack{z_i \in \{0,1\}^n \\ \sum_i z_i = w+1}} -\sum_{i=1}^{n} z_i\bar{\lambda}_i + \sum_{i=1}^{n} \log(e^{\bar{\lambda}_i}\rho_i + (1-\rho_i)) \right). \tag{73}$$

We will prove that $\Delta \geq 0$, showing that $\vec{\bar{\lambda}}$ is no worse than $\vec{\lambda}$. Note initially that

$$\max_{\substack{z_i \in \{0,1\}^n \\ \sum_i z_i = w+1}} -\sum_{i=1}^{n} z_i\lambda_i \; = \; \max_{\substack{z_i \in \{0,1\}^n \\ \sum_i z_i = w+1}} -\sum_{i=1}^{n} z_i\bar{\lambda}_i,$$

since the maximum value on both sides is equal to the sum of the smallest $w+1$ components of the vector $\vec{\lambda}$. Thus, we only need to compare remaining part of the objective function, i.e., we have

$$\begin{aligned}\Delta \; &= \; \sum_{i=1}^{n} \log(e^{\lambda_i}\rho_i + (1-\rho_i)) - \sum_{i=1}^{n} \log(e^{\bar{\lambda}_i}\rho_i + (1-\rho_i)) \\ &= \; \log(e^{\lambda_j}\rho_j + (1-\rho_j)) + \log(e^{\lambda_{j+1}}\rho_{j+1} + (1-\rho_{j+1})) \\ &\quad - \log(e^{\bar{\lambda}_j}\rho_j + (1-\rho_j)) - \log(e^{\bar{\lambda}_{j+1}}\rho_{j+1} + (1-\rho_{j+1})).\end{aligned}$$

Since $\bar{\lambda}_j = \lambda_{j+1}$ and $\bar{\lambda}_{j+1} = \lambda_j$, it follows that

$$\begin{aligned}\Delta \; &= \; \log\left(\frac{e^{\lambda_j}\rho_j + (1-\rho_j)}{e^{\lambda_{j+1}}\rho_j + (1-\rho_j)}\right) - \log\left(\frac{e^{\lambda_j}\rho_{j+1} + (1-\rho_{j+1})}{e^{\lambda_{j+1}}\rho_{j+1} + (1-\rho_{j+1})}\right) \\ &= \; \log\left(\frac{e^{\lambda_j} - e^{\lambda_{j+1}}}{e^{\lambda_{j+1}} + \frac{1}{\rho_j} - 1} + 1\right) - \log\left(\frac{e^{\lambda_j} - e^{\lambda_{j+1}}}{e^{\lambda_{j+1}} + \frac{1}{\rho_{j+1}} - 1} + 1\right).\end{aligned}$$

Note that the argument inside the log is positive, since $\lambda_j > \lambda_{j+1}$. Moreover, since $\rho_j \geq \rho_{j+1}$, we see that $1/\rho_j - 1 \leq 1/\rho_{j+1} - 1$ and hence we conclude that $\Delta \geq 0$. □

Theorem 9 reduces the minimization problem to a one-dimensional problem, which can be efficiently solved numerically. Note that if x is such that $|y| = w+1$, then $\hat{\rho}_c = 1$. In this case $\hat{p}^{\mathrm{IS}_0}(x)$ is constant and equal to $p(x)$. Regarding the assumptions of Theorem 9, the condition $\sum_{c=1}^{C} \rho_c y_c < w$ is satisfied in our routing and dimensioning problem under mild assumptions. Indeed, since $\sum_{c=1}^{C} y_c \leq C$, it follows from Proposition 8 that $\mathbb{P}\left\{\sum_c y_c\xi_c > \sum_c y_c\rho_c\right\} > \delta/(2C) > 0$, provided that $0 < \rho_c < 1$ for all c. Therefore, if $\delta/(2C) \geq \alpha$ — which is typically the case since α is very small — then w must be bigger than $\sum_{c=1}^{C} y_c\rho_c$ to be feasible. Similarly, the condition $w \leq \sum_{c=1}^{C} y_c - 1$ is harmless, since if $w \geq \sum_c y_c$ then $p(x) = 0$ regardless of the value of the parameters ρ_c, i.e, there is nothing to estimate.

4.2 Uniform variance reduction for the (CC-JRD) formulation

The results in Section 4.1 show how to control the variance of the estimator $\hat{p}^{\mathrm{IS}_0}(x)$ for a fixed x. We would like to find an IS distribution that reduces the variance uniformly for all $x \in X$ (recall that X is set of points $x = (y, w)$ satisfying (19) and (21)). As it turns out, such a requirement is too strong. To understand why, consider a point $x \in X$ such that $p(x) \gg \alpha$. Not only is x infeasible for problem Problem **(CC-JRD)**, but also its variance may be large. On the other hand, there is no need to obtain a precise estimation of that quantity in order to check its infeasibility. So, trying to obtain uniform variance reduction over X may not be desirable, since by requiring such uniformity over the whole set X we would be sacrificing the quality of the estimators at the points that really matter, i.e. where $p(x) \approx \alpha$. The numerical experiments in Section 5 will illustrate this issue. Of course, characterizing exactly the feasibility set $\{x \in X \,:\, p(x) \leq \alpha\}$ is impractical, since such a task is as difficult as solving the original problem. Our approach is then to construct an *outer approximation* X_0 of the feasibility set such that variance is reduced uniformly over X_0. In what follows we proceed in that direction.

As discussed in Section 4.1, a necessary condition for feasibility of x (when α is sufficiently small) is that $w > \sum_{c=1}^{C} \rho_c y_c$. Thus, we can replace the original set X with the set

$$X_0 \;:=\; \left\{ x \in X \;:\; w > \sum_{c=1}^{C} \rho_c y_c \right\} \tag{74}$$

since this entails simply adding a linear inequality to the original problem. Our initial goal is to ensure uniform reduction variance for all x in X_0.

The following result shows how to choose IS parameters in order to guarantee variance reduction for all $x \in X_0$. Recall from Theorem 9 that there is a minimizer of the function $B_x(\vec{\lambda})$ such that all components of $\vec{\lambda}$ have the same value. Consider now the restriction of B_x to the set $\{\vec{\lambda} \in \mathbb{R}_+^C \,:\, \lambda_1 = \ldots = \lambda_C\}$. To abbreviate the notation, we shall write this function as $B_x(\lambda)$, where $\lambda \in \mathbb{R}_+$. Similarly, we will write the corresponding IS parameter as $\hat{\rho}_c(\lambda)$ for all $c \in 1 \ldots C$.

Proposition 11. *Let*

$$\varepsilon^{\mathrm{IS}_0}(\lambda) \;:=\; \max_{x \in X_0} B_x(\lambda) \tag{75}$$

Then, $\varepsilon^{\mathrm{IS}_0}(\cdot)$ is a convex function and there is a $\bar{\lambda} \in \mathbb{R}_+ \cup \{\infty\}$ that minimizes that function. Moreover, $\varepsilon^{\mathrm{IS}_0}(\bar{\lambda}) < 1$ and $\hat{p}^{\mathrm{IS}_0}$ has $\varepsilon^{\mathrm{IS}_0}(\bar{\lambda})$-uniform variance reduction with respect to the standard Monte Carlo estimator $\hat{p}$ on the set X_0.

Proof. By Theorem 9, $B_x(\cdot)$ is a continuous convex function, so $\varepsilon^{\mathrm{IS}_0}(\cdot)$ is also a continuous convex function. Moreover, using an argument similar to that used in the proof of Theorem 9 we find that the derivative of the function $\psi_x(\lambda) := \log B_x(\lambda)$ is

$$\frac{d\psi_x}{d\lambda} \;=\; -(w+1) + \sum_{c=1}^{C} \frac{e^{\lambda}\rho_c y_c}{e^{\lambda}\rho_c + (1-\rho_c)}\,.$$

Since $\frac{dB_x}{d\lambda} = B_x(\lambda)\frac{d\psi_x}{d\lambda}$ and $B_x(0) = 1$, it follows that $\frac{dB_x}{d\lambda}(0) = \sum_{c=1}^{C} \rho_c y_c - (w+1) < 0$ for all $x \in X_0$. Therefore, the $\bar{\lambda}$ that minimizes $\varepsilon^{\mathrm{IS}_0}(\cdot)$ is such that $\bar{\lambda} > 0$ (possibly $\bar{\lambda} = \infty$) and $\varepsilon^{\mathrm{IS}_0}(\bar{\lambda}) < 1$. Finally, from Proposition 5 we conclude that $\hat{p}^{\mathrm{IS}_0}$ has $\varepsilon^{\mathrm{IS}_0}$-uniform variance reduction with respect to $\hat{p}$ on the set X_0. □

Further enhancement in variance reduction can be obtained as follows. As we have seen in Section 4.1, for the points $x = (y, w)$ such that $w \geq |y|$ we have $p(x) = 0$ and so any sampling approximation will yield the correct value. Thus, we can assume that $w \leq |y| - 1$, which implies that

$$\begin{aligned} p(x) &= \mathbb{P}\left\{\sum_{c=1}^{C} \xi_c y_c > w\right\} \;\geq\; \mathbb{P}\left\{\sum_{c=1}^{C} \xi_c y_c > |y| - 1\right\} \;=\; \mathbb{P}\left\{\sum_{c=1}^{C} \xi_c y_c = |y|\right\} \\ &= \mathbb{P}\{\xi_c = 1, \text{ for all } c \text{ such that } y_c = 1\} \;\geq\; \prod_{c=C-|y|+1}^{C} \rho_c \ , \end{aligned}$$

where the last inequality arise from our stated assumption that $\rho_1 \geq \ldots \geq \rho_C$ and the independence of $\{\xi_c\}$. It follows that if

$$|y| \;\leq\; n_0 := \max\left\{k : \prod_{c=C-k+1}^{C} \rho_c \geq K\alpha\right\} \tag{76}$$

for some reasonably large K (say, $K = 10$), then we have that $p(x) \geq K\alpha$, in which case we say that x is sufficiently far from the feasibility set and therefore variance need not be reduced for such x. In addition, let $w_0(y)$ be a valid lower bound for w for any feasible x, such that $w_0(y) > \sum_{c=1}^{C} \rho_c y_c$ (later we will see how to construct one such bound). Define the set

$$X_0' \;:=\; \{x \in X_0 \,:\, w \geq w_0(y) \text{ and } |y| \geq n_0\} \ . \tag{77}$$

Since $X_0' \subseteq X_0$, from Proposition 11 it is clear that by re-defining $\varepsilon^{\mathrm{IS}_0}$ as $\varepsilon^{\mathrm{IS}_0}(\lambda) := \max_{x \in X_0'} B_x(\lambda)$ we also obtain uniform variance reduction on X_0'. The advantage of working with the reduced set X_0' instead of X_0 is that, by discarding points that are "too far" from the feasibility set (in the sense that $p(x) \geq K\alpha$) and by cutting further the feasibility set via the sharper lower bound $w_0(y)$, larger variance reduction can be obtained; as we shall see in Section 5, such intuition is confirmed by the numerical experiments.

It remains to discuss how to calculate $\varepsilon^{\mathrm{IS}_0}(\lambda)$ in (75) (using X_0' in (77) in place of X_0) since the maximum is computed over an uncountable set. In the particular case where $w_0(y)$ only depends on $|y|$, let x_k be the routing such that the k connections with highest rates are routed through the link, i.e., $y_c = \mathbb{1}_{c \leq k}$, and let $w_k := w_0(y)$ for y such that $|y| = k$. Then we have

$$\begin{aligned} \varepsilon^{\mathrm{IS}_0}(\lambda) &= \max_{x \in X_0} \left\{\exp\big(-\lambda(w_0(y)+1)\big) \prod_{c=1}^{C} \big(1 + \rho_c(e^{\lambda} - 1)\big)^{y_c}\right\} \\ &= \max_{k=n_0,\ldots,C} \max_{x : |y| = k} \left\{\exp\big(-\lambda(w_0(y)+1)\big) \prod_{c=1}^{C} \big(1 + \rho_c(e^{\lambda} - 1)\big)^{y_c}\right\} \\ &= \max_{k=n_0,\ldots,C} \left\{\exp\big(-\lambda(w_k+1)\big) \prod_{c=1}^{k} \big(1 + \rho_c(e^{\lambda} - 1)\big)\right\} \\ &= \max_{k=n_0,\ldots,C} \{B_{x_k}(\lambda)\} \ . \end{aligned}$$

Hence, in order to find $\bar{\lambda}$ that minimizes $\varepsilon^{\mathrm{IS}_0}(\cdot)$ we only need to consider the maximum among $C - n_0 + 1$ convex functions.

To illustrate the calculation, in Figure 2 we plot B_{x_k} for a case of $C = 21$ connections with rates ρ_c randomly chosen between 0.1 and 0.3 and $\alpha = 10^{-6}$. We use $K = 10$ in (76), which yields $n_0 = 6$, and the lower bound $w_0(y)$ constructed as the $(1-\alpha)$-quantile of a binomial distribution with parameters $|y|$ and ρ_C. It can be seen that B_{x_k} has a steep decrease near zero, and $\varepsilon^{\mathrm{IS}_0}(\lambda)$ can reach very low values, resulting in considerable variance reduction.

Finally, note that all of the results in this section can be naturally extended to the estimator $\hat{p}^{\mathrm{IS}}(x)$ defined in (25). In fact, in that case the calculation is much easier, due to the fact that the likelihood function does not depend on x. We state the result in Proposition 12 below.

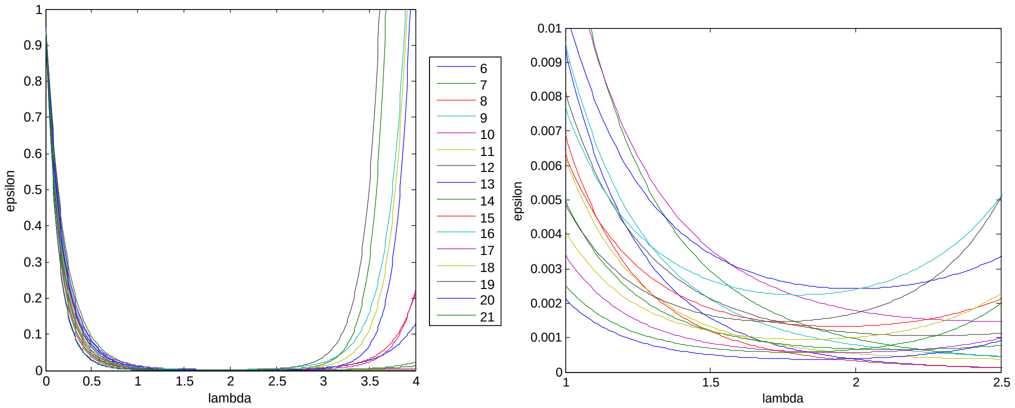

Figure 2: Example of the values of $B_{x_k}(\lambda)$ for $C = 21$ connections with rates ρ_c randomly chosen between 0.1 and 0.3 and $\alpha = 10^{-6}$ for $\lambda \in [0, 4]$ (left) and an enlarged version for $\lambda \in [1, 2.5]$ (right).

Proposition 12. *Suppose that the lower bound $w_0(y)$ is a non-decreasing function that only depends on $|y|$, let $w_k := w_0(y)$ for y s.t. $|y| = k$, and let*

$$\varepsilon^{\mathrm{IS}}(\lambda) \;:=\; e^{-\lambda(w_{n_0}+1)} \prod_{c=1}^{C} (e^{\lambda}\rho_c + (1-\rho_c)) \;, \tag{78}$$

where n_0 is defined in (76). If $\sum_{c=1}^{C} \rho_c \le w_{n_0} + 1$ then the optimal $\bar{\lambda}$ that minimizes $\varepsilon^{\mathrm{IS}}(\cdot)$ satisfies

$$\sum_{c=1}^{C} \hat{\rho}_c(\bar{\lambda}) = w_{n_0} + 1 \qquad \text{and} \qquad \hat{\rho}_c(\bar{\lambda}) = \frac{e^{\bar{\lambda}}\rho_c}{e^{\bar{\lambda}}\rho_c + (1-\rho_c)} \;.$$

Moreover, $\varepsilon^{\mathrm{IS}}(\bar{\lambda}) < 1$ and $\hat{p}^{\mathrm{IS}}$ has $\varepsilon^{\mathrm{IS}}(\bar{\lambda})$-uniform variance reduction with respect to the standard Monte Carlo estimator $\hat{p}$ on the set X_0'.

Proof. By doing a similar construction to that in Section 4.1 for the estimator $\hat{p}^{\mathrm{IS}}(x)$, we obtain that the quantity $B_x(\vec{\lambda})$ defined in (56) is the same except that it does not have the exponent y_c. Thus, Theorem 9 applies, with $y_c = 1$ for all c. It follows that the quantity analogous to (75) can be written as

$$\varepsilon(\lambda) \;:=\; \max_{x \in X_0'} \left\{ e^{-\lambda(w+1)} \prod_{c=1}^{C} (e^{\lambda}\rho_c + (1-\rho_c)) \right\}.$$

Note that the quantity inside the braces is a decreasing function of w, so the max is achieved at the smallest value of $w+1$ in X_0', which is $w_{n_0} + 1$. This leads to the expression (78). By applying logarithm and minimizing over λ we reach the desired conclusion. □

We conclude this section by recalling that the above calculations have been conducted for an arbitrary arc a. While the calculations are similar for other arcs, the obtained IS parameter values for the estimator $\hat{p}^{\mathrm{IS}_0}$ may of course be different, as they depend on which and how many connections are routed through that particular arc. In the case of the estimator $\hat{p}^{\mathrm{IS}}$, however, we see from Proposition 12 that the parameter values depends on all the connections that can be potentially routed through that arc.

5 Computational experiments

In this section we compare the performance of the different approaches for solving problem **(CC-JRD)** through computational experiments. When we use estimators $\hat{p}_a(x), \hat{p}_a^{\mathrm{IS}}(x)$ and $\hat{p}_a^{\mathrm{IS}_0}(x)$, for each arc $a \in A$, we will refer to the corresponding approximation as SAA, SAA-IS and SAA-IS$_0$, respectively. We are interested in evaluating the quality of the solutions obtained, in particular when the sample size N is significantly smaller than $1/\alpha$.

We test the three approaches over a ring topology, which is one of the most common real-world topologies for optical networks. We study rings with 7 and 9 nodes for the homogeneous case, in which all connection rates ρ_c are identical and equal to 0.1. As mentioned earlier, in that case the problem can be solved directly by inverting binomial distributions; nevertheless, we apply our SAA methodology to that problem in order to verify the quality of the solutions obtained with the importance sampling approach. We later report experiments conducted for the heterogeneous case, for which there is no analytical solution.

As previously explained, we need to consider a smaller subset X_0' where the variance of $\hat{p}^{\mathrm{IS}_0}$ is uniformly reduced, which is given by the choice of the function $w_0(y)$. Table 1 shows the obtained IS measure $\hat{\rho}$ and the resulting theoretical variance reduction obtained by applying different IS estimators, with different functions w_0. All IS parameters were computed using Propositions 11 and 12, for n_0 computed using (76) with $K = 10$. In column SAA-IS$_0$-1 we use the best possible lower bound $w_0(y)$, obtained by computing the $(1-\alpha)$-quantile of a binomial distribution with parameters $|y|$ and ρ. We see that we obtain enormous variance reduction in this case for all $x \in X_0'$, with values similar to the value of α. However, note that by adding this constraint to the problem we generate exactly the feasibility set of problem **(CC-JRD)**, without requiring to use a sampling approximation of the chance constraints; of course, we can do that in this case because we are dealing with a homogeneous setting, which as we mentioned before can be solved without sampling—recall that our goal in studying the homogeneous case is just to test our procedure. In column SAA-IS$_0$-2 we use the worst lower bound that satisfy the conditions of Theorem 9, that is, $w_0(y) = \sum_{c=1}^{C} \rho_c y_{a,c}$. In this case, the subset X_0' is too large, resulting in negligible variance reduction. In order to mimic an intermediate situation, in column SAA-IS$_0$-3 we present the results obtained using a linear function $w_0 = m|y|$, where m is the maximum scalar such that w_0 is a valid lower bound. As we can see, by using this restriction of the feasible set we still can decrease the variance of the estimator by 2 to 3 orders of magnitude. Hence, in the following experiments we will use this last IS estimator, and we include the constraint $w_a \geq m \sum_{c=1}^{C} y_{a,c}$ for all $a \in A$ in all formulations.

Instance		SAA-IS		SAA-IS$_0$-1		SAA-IS$_0$-2		SAA-IS$_0$-3	
Size	α	$\hat{\rho}$	ε^{IS}	$\hat{\rho}$	ε^{IS_0}	$\hat{\rho}$	ε^{IS_0}	$\hat{\rho}$	ε^{IS_0}
7	10^{-3}	0.142	8.3E-01	0.524	6.9E-03	0.147	7.9E-01	0.644	6.E-02
7	10^{-6}	0.285	5.9E-02	0.693	9.0E-06	0.147	7.9E-01	0.667	4.E-03
9	10^{-3}	0.100	1.0E+00	0.437	1.2E-02	0.128	8.7E-01	0.523	1.E-01
9	10^{-6}	0.166	4.7E-01	0.575	2.8E-05	0.128	8.7E-01	0.555	2.E-02

Table 1: Optimal IS parameter $\hat{\rho}$ and resulting variance reduction ε for different IS.

We solve the corresponding MIP formulations, SAA, SAA-IS and SAA-IS$_0$, with sample sizes $N = 20$ and $N = 50$ for $\rho = 0.1$ and $\alpha = 10^{-3}$ and $\alpha = 10^{-6}$. Additionally, we solve the SAA formulation with $N = 1000$ samples. Each instance is solved 100 times with different random seeds.

Instance			SAA			SAA-IS		SAA-IS$_0$	
Size	α	Opt.	$N = 20$	$N = 50$	$N = 1000$	$N = 20$	$N = 50$	$N = 20$	$N = 50$
7	10^{-3}	47	31.87	35.22	44.56	35.62	40.41	44.36	46.26
7	10^{-6}	68	42.34	42.77	48.10	52.01	57.26	67.03	67.65
9	10^{-3}	81	57.57	62.01	76.93	57.57	62.01	78.73	OOM
9	10^{-6}	117	72.64	73.51	84.07	77.62	82.61	114.94	117.29

Table 2: Computational results for the homogeneous case $\rho = 0.1$.

The average value of the obtained objective function is presented in Table 2, and the true optimal value is presented in column *Opt*. As expected, we see that the SAA formulation only provides a reasonably good approximation of the true optimum value when the number of samples is of order $1/\alpha$. A striking feature of the SAA-IS$_0$ is that it generates good approximation of the real optimal value of the problem even with a very small number of samples. However, this IS requires a more complex formulations, that runs out of memory (OOM) for one of these instances. For the case of SAA-IS, the approximation is not as good as in the previous case, but still better than the traditional SAA formulation. Those results strongly support the use of importance sampling in chance-constrained problems with very small α.

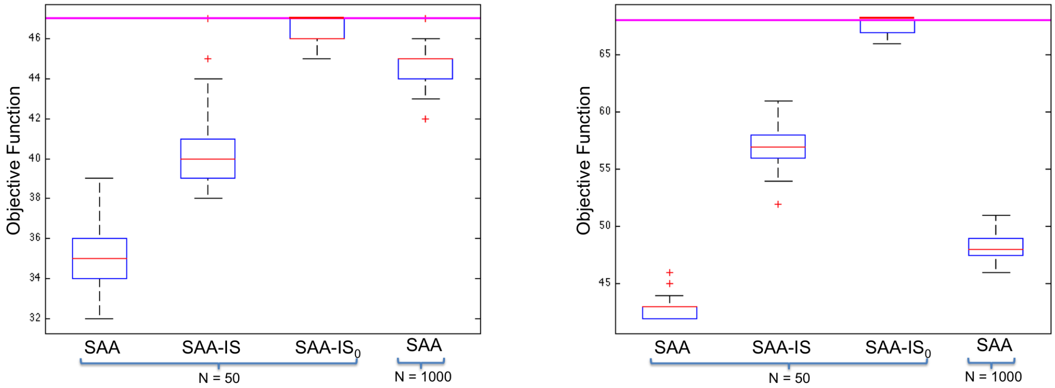

Figure 3: Homogeneous case: boxplot of the different estimators for a ring of size 7 and $\rho = 0.1$, for $\alpha = 10^{-3}$ and $\alpha = 10^{-6}$.

Table 2 illustrates our claim that by using importance sampling it is possible to obtain good approximations with very reduced sample sizes, that is, much smaller than $O(1/\alpha)$. In order to study the dispersion of these values, in Figure 3 we present a boxplot of the resulting objective values for each approach in the case of a ring of size 7 where $\alpha = 10^{-3}$ (left) and 10^{-6} (right). The first three columns show SAA, SAA-IS and SAA-IS$_0$ for $N = 50$, respectively. The last column exhibit the values for SAA with $N = 1000$ samples. As we can see, IS estimators produce better results than SAA in almost all runs, particularly when $\alpha = 10^{-6}$.

It is worth noting that in the majority of cases the resulting objective values are smaller than the real optimal value in almost every run. Therefore, the resulting solution of each problem must be infeasible for the original problem. This infeasibility comes from an underestimation of the capacity w_a required at each link $a \in A$, given the corresponding routing decision of each solution (i.e., the y variables). In fact, all runs of SAA and SAA-IS returns an infeasible solution, but for SAA-IS$_0$ that is not the case. For the ring of size 7, $N = 50$ samples and $\alpha = 10^{-3}$ and 10^{-6}, formulation SAA-IS$_0$ returns 4 and 28 times (of the 100 runs) the true optimum, respectively. For the ring of size 9, $N = 50$ samples and $\alpha = 10^{-6}$, this formulation returns 19 times the true optimum.

For the heterogeneous case, we use a traffic rate ρ_c randomly chosen according to a uniform distribution between 0.1 and 0.3. Recall that the exact optimal solution for these instances is not known. Nevertheless, we can obtain a lower bound for variables w by computing the $(1-\alpha)$-quantile of a binomial distribution with parameters $|y|$ and ρ_C, that is, the smallest rate. Then, we use this lower bound as our function $w_0(y)$ to define the set X_0' where variance will be uniformly reduced. Note that each link has a different set of connections that can potentially be routed through them, so a different IS parameter may be obtained for each link. We choose the IS parameter $\hat{\rho}_c$ for each link as explained in Propositions 11 and 12. The range of theoretical ε-variance reduction among all links is presented in Table 3.

In Table 4 we show results for the heterogeneous case. As before, each column shows the average obtained objective value among 100 runs of each configuration. Since the exact optimal solution for these instances is not known, we present an optimality range, constructed assuming that for each link all connections have an homogeneous rate equal to the minimum (maximum) rate among all connections that can be routed though

Size	α	ε^{IS}	ε^{IS_0}
7	10^{-3}	1.000	0.1589 - 0.2994
7	10^{-6}	0.1370 - 0.2821	0.0024 - 0.0085
9	10^{-3}	1.000	0.2081 - 0.4366
9	10^{-6}	0.8441 - 0.9121	0.0036 - 0.0177

Table 3: Range of optimal ε-variance reduction for different IS, among all links.

them, in order to obtain a lower (upper) bound of the optimal value.

Instance		Optim.	SAA			SAA-IS		SAA-IS$_0$	
N	α	range	$N = 20$	$N = 50$	$N = 1000$	$N = 20$	$N = 50$	$N = 20$	$N = 50$
7	10^{-3}	48-70	49.29	50.93	60.38	49.29	50.93	57.67	OOM
7	10^{-6}	69-84	69.00	69.00	69.18	70.03	71.40	76.60	OOM
9	10^{-3}	84-135	85.34	87.92	OOM	85.34	87.92	OOM	OOM
9	10^{-6}	122-171	122.07	122.14	123.44	122.24	122.51	OOM	OOM

Table 4: Computational results for the heterogeneous case $\rho_c \in [0.1, 0.3]$.

It can be seen that for SAA, the resulting objective values are very close to the lower bound $w_0(y)$, particularly for $\alpha = 10^{-6}$. For SAA-IS, the results are similar to the SAA, which is explained by the small variance reduction of this estimator, as presented in Table 3. However, for SAA-IS$_0$ we obtain higher values than the other estimators using only $N = 20$ samples. Since rates are chosen uniformly between $[0.1, 0.3]$, it is reasonable to expect that the optimal value should be near the middle of the optimality range, which is the case for the values obtained by SAA-IS$_0$ for the 7-node ring instances. Nevertheless, it is hard to conclude further about the quality of the approximation of our IS estimators without knowing the true optimal value. Unfortunately the formulation SAA-IS$_0$ could not be solved for the 9-node ring and for 7-node ring with $N = 50$, since the resulting mixed-integer programs are too big. Still, our goal in this paper is to provide a "proof of concept" to demonstrate that the use of importance sampling in rare-event chance-constrained problems has the potential to allow for the solution of such problems with small sample sizes; conceivably, a more efficient formulation of the mixed integer program can be derived by exploiting characteristics of the problem, but such a task is out of the scope of this paper.

6 Conclusions

Sampling methods for chance-constrained programming (CCP) problems are extremely popular and have been used extensively lately. Our main contribution in this paper is to address the situation in which the desired reliability level is very close to one, e.g. $1-10^{-6}$. The results and algorithms available in the literature cannot cope with this situation, and we showed that importance sampling is a provably convergent tool to solve rare-event CCP problems. Importance sampling has been extensively used in simulation to estimate rare events, but in an optimization context several difficulties arise. The main problem is the dependence of the estimator on the decision variables, which motivates us to look for estimators that uniformly reduce the variance over the decision space.

We studied a problem in telecommunications and wrote explicit formulations that use IS techniques. We constructed IS distributions for which we can theoretically guarantee uniform variance reduction over an outer approximation of the feasibility set. For the homogeneous case our experiments showed that sample sizes much smaller that $1/\alpha$, e.g. $\alpha = 10^{-6}$ with a sample size of 50, can yield excellent approximations to the true optimum. In the heterogeneous case the optimal solution is not known but we showed that small sample sizes can yield good solutions as in the homogeneous case. For the SAA-IS$_0$ estimator the computational burden is significant and better formulations need to be derived in order to obtain solutions for the most demanding cases.

We would like to highlight the problem-dependent characteristic of importance sampling problems. We believe that in order to use IS techniques in other CCP problems with rare events two requirements must be satisfied: on one hand, IS estimators must be customized for each decision and, at the same time, lead to a tractable optimization problem. On the other hand, one must find an outer approximation of the feasibility set in which variance can be significantly reduced. It is worth pointing out that for some problems using a non adapted IS estimator such as SAA-IS can be a good compromise between a simple mathematical programming formulation and a good enough variance reduction, although it does not perform as well as an adapted one such as SAA-IS$_0$.

Therefore, we believe the results and ideas presented here could serve as a guidance for choosing the appropriate IS estimator for other problems. We hope that this work will foster further research on rare-event chance constrained problems, which seem to have been neglected so far in the literature.

Acknowledgements. Authors acknowledge the financial support of grant Anillo ACT-88, Basal Center CMM-UCh, CIRIC-INRIA Chile (J.B., E.M., G.C.) and FONDECYT grants 1100618 (J.B.), 1120244 (T.H., B.P.) and 1130681 (E.M.).

References

[1] Adas, A.: Traffic models in broadband networks. Communications Magazine, IEEE **35**(7), 82–89 (1997)

[2] Andrieu, L., Henrion, R., Römisch, W.: A model for dynamic chance constraints in hydro power reservoir management. European Journal of Operational Research **207**(2), 579–589 (2010)

[3] Artstein, Z., Wets, R.J.B.: Consistency of minimizers and the slln for stochastic programs. J. Convex Anal **2**(1-2), 1–17 (1996)

[4] Asmussen, S., Glynn, P.: Stochastic Simulation. Springer, New York (2007)

[5] Beraldi, P., Ruszczyński, A.: The probabilistic set-covering problem. Operations Research **50**(6), 956–967 (2002)

[6] Bonami, P., Lejeune, M.: An exact solution approach for portfolio optimization problems under stochastic and integer constraints. Operations Research **57**(3), 650–670 (2009)

[7] Calafiore, G., Campi, M.C.: Uncertain convex programs: randomized solutions and confidence levels. Mathematical Programming **102**(1), 25–46 (2005)

[8] Campi, M.C., Garatti, S.: The exact feasibility of randomized solutions of uncertain convex programs. SIAM Journal on Optimization **19**(3), 1211–1230 (2008)

[9] Campi, M.C., Garatti, S.: A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. Journal of Optimization Theory and Applications **148**(2), 257–280 (2011)

[10] Campi, M.C., Garatti, S., Prandini, M.: The scenario approach for systems and control design. Annual Reviews in Control **33**(2), 149–157 (2009)

[11] Carniato, A., Camponogara, E.: Integrated coal-mining operations planning: Modeling and case study. International Journal of Coal Preparation and Utilization **31**(6), 299–334 (2011)

[12] Charnes, A., W., C.W., Symmonds, G.: Cost horizons and certainty equivalents: and approach to stochastic programming of heating oil. Management Science **4**, 235–263 (1958)

[13] Chung, K.L.: A Course in Probability Theory, 2nd. edn. Academic Press, New York, NY (1974)

[14] Dantzig, G.B., Glynn, P.W.: Parallel processors for planning under uncertainty. Annals of Operations Research **22**(1), 1–21 (1990)

[15] Dentcheva, D., Prékopa, A., Ruszczynski, A.: Concavity and efficient points of discrete distributions in probabilistic programming. Mathematical Programming **89**(1), 55–77 (2000)

[16] Dorfleitner, G., Utz, S.: Safety first portfolio choice based on financial and sustainability returns. European Journal of Operational Research **221**(1), 155–164 (2012)

[17] Duckett, W.: Risk analysis and the acceptable probability of failure. Structural Engineer **83**(15), 25–26 (2005)

[18] Homem-de-Mello, T., Bayraksan, G.: Monte Carlo methods for stochastic optimization (2013). Manuscript, submitted to *Surveys in OR/MS*. Available on Optimization Online, http://www.optimization-online.org/DB_HTML/2013/06/3920.html.

[19] Infanger, G.: Monte Carlo (importance) sampling within a Benders decomposition algorithm for stochastic linear programs. Annals of Operations Research **39**(1), 69–95 (1992)

[20] Jiang, R., Guan, Y.: Data-driven chance constrained stochastic program (2012). Available at www.optimization-online.org

[21] Kahn, H., Harris, T.: Estimation of particle transmission by random sampling. National Bureau of Standards applied mathematics series **12**, 27–30 (1951)

[22] L'Ecuyer, P., Mandjes, M., Tuffin, B.: Importance sampling in rare event simulation. In: G. Rubino, B. Tuffin (eds.) Rare Event Simulation using Monte Carlo Methods, chap. 2. John Wiley & Sons, Inc., (2009)

[23] Lejeune, M.: Pattern definition of the p-efficiency concept. Annals of Operations Research **200**(1), 23–36 (2012)

[24] Li, W.L., Zhang, Y., So, A.C., Win, Z.: Slow adaptive ofdma systems through chance constrained programming. Signal Processing, IEEE Transactions on **58**(7), 3858–3869 (2010)

[25] Liu, Y., Guo, H., Zhou, F., Qin, X., Huang, K., Yu, Y.: Inexact chance-constrained linear programming model for optimal water pollution management at the watershed scale. Journal of Water Resources Planning and Management **134**(4), 347–356 (2008)

[26] Luedtke, J., Ahmed, S.: A sample approximation approach for optimization with probabilistic constraints. SIAM Journal on Optimization **19**(2), 674–699 (2008)

[27] Minoux, M.: Discrete cost multicommodity network optimization problems and exact solution methods. Annals of operations research **106**(1-4), 19–46 (2001)

[28] Minoux, M.: Multicommodity network flow models and algorithms in telecommunications. In: M. Resende, P. Pardalos (eds.) Handbook of optimization in telecommunications, pp. 163–184. Springer (2006)

[29] Nemirovski, A., Shapiro, A.: Convex approximations of chance constrained programs. SIAM Journal on Optimization **17**(4), 969–996 (2006)

[30] Pagnoncelli, B., Ahmed, S., Shapiro, A.: Sample average approximation method for chance constrained programming: theory and applications. Journal of optimization theory and applications **142**(2), 399–416 (2009)

[31] Pagnoncelli, B.K., Reich, D., Campi, M.C.: Risk-return trade-off with the scenario approach in practice: a case study in portfolio selection. Journal of Optimization Theory and Applications **155**(2), 707–722 (2012)

[32] Prékopa, A.: Probabilistic programming. In: A. Ruszczyński, A. Shapiro (eds.) Stochastic Programming, vol. 10, pp. 267–351. Elsevier (2004)

[33] Rockafellar, R.T., Wets, R.J.B.: Variational Analysis, *A Series of Comprehensive Studies in Mathematics*, vol. 317. Springer (1998)

[34] Rosenbluth, M.N., Rosenbluth, A.W.: Monte carlo calculation of the average extension of molecular chains. The Journal of Chemical Physics **23**, 356 (1955)

[35] Rubinstein, R.Y.: Cross-entropy and rare events for maximal cut and partition problems. ACM Transactions on Modeling and Computer Simulation **12**(1), 27–53 (2002)

[36] Rubinstein, R.Y., Shapiro, A.: Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method. J. Wiley & Sons, Chichester, England (1993)

[37] Shapiro, A.: Monte Carlo sampling methods. In: A. Ruszczynski, A. Shapiro. (eds.) Stochastic Programming, *Handbooks in Operations Research and Management Science*, vol. 10. Elsevier Science Publishers B.V., Amsterdam, Netherlands (2003)

[38] Shapiro, A., Dentcheva, D., Ruszczyński, A.: Lectures on stochastic programming: modeling and theory, vol. 9. SIAM (2009)

[39] Soekkha, H.M.: Aviation Safety: Human Factors, System Engineering, Flight Operations, Economics, Strategies, Management. VSP (1997)

[40] Thieu, Q.T., Hsieh, H.Y.: Use of chance-constrained programming for solving the opportunistic spectrum sharing problem under rayleigh fading. In: Wireless Communications and Mobile Computing Conference (IWCMC), 2013 9th International, pp. 1792–1797. IEEE (2013)

[41] Vallejos, R., Zapata-Beghelli, A., Albornoz, V., Tarifeño, M.: Joint routing and dimensioning of optical burst switching networks. Photonic Network Communications **17**(3), 266–276 (2009)