Adaptive Importance Sampling Based Surrogation Methods for Bayesian Hierarchical Models, via Logarithmic Integral Optimization[*]

Ziyu He Junyi Liu Jong-Shi Pang

Original May 2023

Abstract

This paper investigates computational algorithms for Maximum a Posteriori (MAP) and Maximum Likelihood Estimation (MLE) inference of Bayesian Hierarchical Models (BHMs), via a unified formulation as a nonconvex and nondifferentiable logarithmic integral optimization problem. Specifically, we explore situations where a BHM comprises density functions with intractable normalizers, a feature that can present significant computational obstacles, particularly when combined with nonconvexity and nondifferentiability, which are increasingly prevalent in contemporary applications of computational statistics. To deal with these challenges, we propose an efficient algorithmic approach, termed *Adaptive Importance Sampling-based Surrogation*, to simultaneously handle nonconvexity and nondifferentiability while also improving the sampling approximation of the intractable normalizer through variance reduction. Performance of this algorithm is guaranteed by our analysis which establishes an almost sure subsequential convergence to a necessary candidate for a local minimizer, referred to as a *surrogation stationary point*. We also demonstrate the effectiveness of our algorithm through extensive numerical experiments, verifying its efficiency and stability in enabling more advanced BHMs where intractable normalizers arise as a result of enhanced modeling capability.

Key words. Bayesian hierarchical models, intractable normarlizer, logarithmic integral optimization, adaptive importance sampling.

1 Introduction

In recent decades, *Bayesian Hierarchical Models* (BHMs) have emerged as a powerful tool for data science due to its flexibility in modeling complex data structures, while accounting for prior knowledge on the underlying data generating mechanism. A BHM achieves this dual benefit by proposing hierarchies of conditional probability distributions to capture the correlations among (random) data $\widetilde{y}$ and (random) model parameters $\widetilde{\xi}$. On one hand, $\widetilde{y}$ is connected to $\widetilde{\xi}$ via the conditional density function $p(y \mid \xi)$, called the *likelihood*; on the other hand, the *prior* distribution of $\widetilde{\xi}$ has density $q(\bullet)$, which may include additional dependencies among the components of $\widetilde{\xi}$. Given a BHM, a common approach to estimate the model parameter ξ from observed data y, is to solve the *Maximum a Posteriori* (MAP) problem:

$$\underset{\xi}{\textbf{maximize}} \quad p(y \mid \xi) q(\xi) \quad \Longleftrightarrow \quad \underset{\xi}{\textbf{minimize}} \quad -\log p(y \mid \xi) \, - \, \log q(\xi) \tag{1}$$

[*] He and Pang are affiliated with the Daniel J. Epstein Department of Industrial and Systems Engineering, University of Southern California, Los Angeles, U.S.A. 90089. Dr. Liu is affiliated with the Department of Industrial Engineering, Tsinghua University, Beijing, China. The work of Pang was based on research supported by the U.S. Air Force Office of Scientific Research under grant FA9550-18-1-0382. Emails: ziyuhe@usc.edu; junyiliu@tsinghua.edu.cn; jongship@usc.edu.

which maximizes the posterior density of $\widetilde{\xi}$ and reduces to *Maximum Likelihood Estimation* (MLE) when $\widetilde{\xi}$ is deterministic. Despite their popularity, MAP and MLE for BHMs with enhanced modeling capacities are often handicapped by the following computational challenges that are increasingly prevalent yet under-addressed in modern applications. These challenges provide the motivation for our study in this paper:

- **(Intractable Normalizers)** A basic BHM employs a conditional density function, denoted as a ratio, which in generic notation is given by: $\pi(\zeta \,|\, \chi) = \dfrac{\ell(\zeta,\, \chi)}{L(\chi)}$ where $\zeta \in \Xi$; this gives rise to the normalizer $L(\chi) \triangleq \displaystyle\int_{\Xi} \ell(z, \chi) dz$ whose closed form can be unavailable. This issue is pervasive in many interesting models where an un-normalized $\ell(\zeta,\, \chi)$ is designed to facilitate the modeling capabilities, albeit at the expense of a resulting normalizer that cannot be evaluated effectively. Such difficulty becomes a major handicap for MAP and MLE inference on models of this type, which can be found in a wide span of applications including machine learning [42], graphical models [3], social networks [23], population genetics [39], epidemiology [28, 33] and image processing [6, 25].
- **(Nonconvexity and Nondifferentiability)** To capture complex phenomena in realistic settings, BHMs often employ density functions that are nondifferentiable and potentially lead to a nonconvex problem (1). A common source is the family of indicator functions (see [29]) which can be approximated by piecewise affine functions (see Section 3). Nonconvexity also typically arises when prevsiously deterministic parameters are randomized to improve the flexibility of a benchmark model. An example exibiting such properties is a generalized *Markov Random Field* (MRF) with an unknown network topology; see Subsection 2.2.

BHMs with intractable normalizers are also referred to as being *doubly intractable* in the literature, a substantial amount of which is dedicated to resolving this issue for *Markov Chain Monte Carlo* inference that aims to sample from the posterior distribution of model parameters, see for instance [17, 34, 35]. In contrast, studies on MAP and MLE under the doubly intractable settings are relatively limited. Early approaches, including *pseudo likelihood* [4] and *stochastic gradient descent* [45] are not adequate as they lack theoretical guarantees in nonconvex and nondifferentiable cases. While *Sample Average Approximation* (SAA) [18, 20, 21] has been shown to be consistent for MLE, it is not practically an algorithm for such cases as well.

To date there has been an absence of rigorous investigations into computational algorithms for MAP/MLE of BHMs with intractable normalizers that are additionally nonconvex and nondifferentiable, hindering their utilities in modern data science. In response, the primary purpose of this paper is to bridge such a gap with the following contributions:

1. Offering a unified treatment for MAP/MLE, we propose a practical algorithmic framework termed *Adpative Importance Sampling (AIS)-based Surrogation* to tackle the following *Logarithmic Integral Optimization* problem:

$$\underset{x \in X}{\textbf{minimize}}\; c(x) \,+\, \log Z(x) \qquad \text{where} \;\; Z(x) \,\triangleq\, \int_{\Xi} r(x,z) dz, \tag{2}$$

where X and Ξ are given sets in their respective spaces (to be specified later). This algorithm employs two major techniques. First, we introduce *surrogation* [11, Chapter 7] to address the challenge of nonconvexity and nondifferentiability. Second, an adaptive importance sampling (AIS) scheme is incorporated to allow for improved control over the variances of sampling-based approximations for intractable Z during the iterations.

2. As a practical solution to accommodate applications with discrete parameters $\widetilde{\xi}$, we present a systematic approach to approximate the inverse of cumulative distribution functions to the discrete priors and convert an otherwise mixed integer programming formulation of the MAP into a continuous formulation that can be handled by our proposed algorithm.

3. Via a rigorous SAA consistency analysis on the non-*independent and identically distributed* (non-iid) triangular array produced by the iterate-dependent sampling process in the AIS-based surrogation method, we present the convergence of our algorithm to a kind of surrogation stationary point, which serves as a computable candidate for a local minimizer. To further support the numerical stability and efficiency of the algorithm in practice, we conduct extensive numerical experiments and compare its performance with related approaches on some realistic BHMs.

By pursuing these endeavors, we anticipate that our research in this paper will not only enhance the field of computational statistics, but also extend the literature on nonconvex stochastic optimization by offering an effective variance reduction algorithm for the logarithmic integral optimization problem, which is interesting in its own right. Before delving into further details, we would like to provide a brief overview on previous works that are relevant to our AIS scheme, which in high level iteratively alternates between sampling approximation of the integral function and updating the sampling distribution. Emerged from the simulation community, the concept of AIS was initially proposed as a variance reduction technique for estimating a scalar-valued integral through sampling, cf. [36] and [7]. Early approaches for stochastic optimization using the AIS scheme that are more related to ours were introduced in [12] and [26]. These studies focused on decomposition methods for multistage stochastic programs and were followed up by a few subsequent studies like [38]. While more recent works have explored AIS in conjunction with stochastic (proximal) gradient descent [1, 27] and coordinate primal descent (dual ascent) methods [8, 44] on problems with convexity or smoothness, the utility of AIS for the challenging type of problem (2) we consider here still remains open, which further motivates the proposal and analysis of our algorithm.

The rest of the paper is organized as follows. In Section 2 we provide some relevant background and motivating examples for our investigation. Section 3 introduces our approach to handle models with discrete priors. Our AIS-based surrogation algorithm is formalized in Section 4 and analyzed in Section 5. Finally, Section 6 reports on the numerical performance of our algorithm.

2 The 3-Layer Bayesian Hierarchical Model

In general, a BHM involves a finite number of uncertainty layers wherein the randomness of a layer is conditional on that of the lower layers. The treatment of this general model involves significant notational complications while the methodology can be based on and extended from that of a 3-layer model, which we present below. Specifically, a 3-level BHM is described by:

$$\begin{array}{llll} \text{level 2}: & \widetilde{y}_i \,|\, \widetilde{\theta} \overset{\text{ind.}}{\sim} & p_i^y(y_i \,|\, \widetilde{\theta}), & i = 1, \ldots, M \\ \text{level 1}: & \widetilde{\theta} \,|\, \widetilde{\gamma} \quad \sim & q(\theta \,|\, \widetilde{\gamma}), & \widetilde{\gamma} \triangleq (\widetilde{\gamma}_\ell)_{\ell=1}^L \\ \text{level 0}: & \quad \widetilde{\gamma}_\ell \overset{\text{ind.}}{\sim} & p_\ell^0(\gamma_\ell), & \ell = 1, \ldots, L; \end{array} \tag{3}$$

where "ind." stands for "independently distributed"; the tilde notation denotes a random variable whose realizations are written without the tilde; the vertical notation denotes conditioning, with $p_i^y(y_i \,|\, \widetilde{\theta})$, $q(\theta \,|\, \widetilde{\gamma})$, and $p_\ell^0(\gamma_\ell)$ denoting the density functions of the respective random variables; the random vector $\widetilde{y}_i$ (the data) is m-dimensional, whereas $\widetilde{\theta}$ (the intermediate) and $\widetilde{\gamma}_\ell$ (the prior) have support $\Theta \subseteq \mathbb{R}^d$ and $\Gamma_\ell \subseteq \mathbb{R}$ respectively. Notice that level 2 of the model as stated assumes that the output $\{\widetilde{y}_i\}_{i=1}^M$ are conditionally independent given $\widetilde{\theta}$ with conditional density $p_i^y(y_i \,|\, \widetilde{\theta})$ and are related to $\widetilde{\gamma} = (\widetilde{\gamma}_\ell)_{\ell=1}^L$, which has independent components, only through the intermediate $\widetilde{\theta}$.

To determine the unknown parameters θ and γ given observations $\{y_i\}_{i=1}^M$ for some positive integer M, the (empirical) MAP problem associated with the BHM (3) maximizes the posterior density of parameters $\widetilde{\theta}$ and $\widetilde{\gamma}$. Termed the o-BHOP (for *original-Bayesian Hierarchical Optimization Problem*), this problem can formally be stated as:

$$\underset{\theta \in \Theta;\ \gamma \in \Gamma}{\textbf{minimize}} \quad -\sum_{i=1}^M \log p_i^y(y_i \,|\, \theta) - \log q(\theta \,|\, \gamma) - \sum_{\ell=1}^L \log p_\ell^0(\gamma_\ell) \qquad \textbf{where} \quad \Gamma \triangleq \prod_{\ell=1}^L \Gamma_\ell. \tag{4}$$

The above formulation has a major deficiency when $\widetilde{\gamma}$ is discretely distributed; i.e., when Γ_ℓ is a discrete set for all ℓ. In this case, the last logarithmic term in the objective function is extended valued whenever γ_ℓ is outside this finite set. Thus any approximation/relaxation-based solution methods can be expected to encounter great difficulty. It turns out that by a well-known transformation as shown below, the BHM model (3) can be reformulated so that this difficulty vanishes, which can serve as the basis for the design of a continuous optimization algorithm.

Fact: Let χ be a random variable with cumulative distribution function (cdf) $F_\chi(t) \triangleq \mathbb{P}(\chi \leq t)$ whose generalized inverse $F_\chi^{-1} : [0,1] \to [\,\inf \chi,\, \sup \chi\,]$ is given by

$$F_\chi^{-1}(s) \triangleq \begin{cases} \inf \left\{ t \mid F_\chi(t) \geq s \right\} & \text{if } s \in (0,1) \\ \inf \chi & \text{if } s = 0 \\ \sup \chi & \text{if } s = 1. \end{cases}$$

so that $Y \triangleq F_\chi^{-1}(U)$, where U is uniformly distributed in $[0,1]$, has the same distribution as χ.

Applying the above transformation to each random variable $\widetilde{\gamma}_\ell$ and letting $\widetilde{U} = (\widetilde{u}_\ell)_{\ell=1}^L$ be the vector of associated uniformly distributed random variables $\widetilde{u}_\ell$ with realizations u_ℓ, we may consider, as an alternative to (3), the uniformly-transformed BHM:

$$\begin{array}{lccl} \text{level 2 :} & \widetilde{y}_i \,|\, \widetilde{\theta} & \overset{\text{ind.}}{\sim} & p_i^y(y_i \,|\, \widetilde{\theta}), \quad i = 1, \ldots, M \\ \text{level 1 :} & \widetilde{\theta} \,|\, \widetilde{U} & \sim & \widehat{q}(\theta \,|\, \widetilde{U}) \triangleq q(\theta \,|\, \phi(\widetilde{U})), \quad \text{where} \ \ \phi(U) \triangleq \left(F_{\widetilde{\gamma}_\ell}^{-1}(u_\ell) \right)_{\ell=1}^L \\ \text{level 0 :} & \widetilde{U} = (\,\widetilde{u}_\ell\,)_{\ell=1}^L & : & L \text{ independent uniform random variables in } [0,1]. \end{array} \tag{5}$$

This leads to the uniform-transformed (empirical) MAP problem, which we term the u-BHOP (for *uniform-Bayesian Hierarchical Optimization Problem*):

$$\underset{\theta \in \Theta;\ u \in [0,1]^L}{\textbf{minimize}} \quad -\sum_{i=1}^M \log p_i^y(y_i \,|\, \theta) - \log \widehat{q}(\theta \,|\, u). \tag{6}$$

Since the density function of $\widetilde{U}$ is constant, its associated term is omitted from the objective of (6).

2.1 Exponential families

We focus on a class of BHMs wherein with $\widetilde{\theta} = (\widetilde{\theta}_i)_{i=1}^M$, the original conditional distributions $\widetilde{y}_i \,|\, \widetilde{\theta}$ and $\widetilde{\theta} \,|\, \widetilde{\gamma}$ have densities belonging to the exponential family so that:

$$p_i^y(y_i \,|\, \theta) = \exp\left[\, g(y_i, \theta_i) - a(\theta_i) \,\right] \quad \text{and} \quad q(\theta \,|\, \gamma) = \exp\left[\, h(\theta, \gamma) - b(\gamma) \,\right] \tag{7}$$

for some bivariate functions $g(y_i, \theta_i)$ and $h(\theta, \gamma)$ with $a(\theta_i)$ and $b(\gamma)$ being the following normalizing factors that ensure the integration of $p_i^y(y_i \,|\, \theta)$ and $q(\theta \,|\, \gamma)$ equal to unity:

$$a(\theta_i) \;=\; \log \int_{\mathbb{R}^m} \exp(g(y', \theta_i))\, dy', \qquad b(\gamma) \;=\; \log \int_{\Theta} \exp(h(\theta', \gamma))\, d\theta'.$$

Thus, under the framework of the exponential families (7), we obtain the associated u-BHOP (6) as:

$$\begin{aligned} \underset{\theta \in \Theta;\; u \in [0,1]^L}{\textbf{minimize}} \quad & -\sum_{i=1}^{M} g(y_i, \theta_i) + \sum_{i=1}^{M} \log \int_{\mathbb{R}^m} \exp\left(\, g(y', \theta_i) \,\right)\, dy' \\ & - h\left(\theta, \phi(u)\right) + \log \int_{\Theta} \exp\left(\, h(\theta', \phi(u)) \,\right) d\theta'. \end{aligned} \tag{8}$$

We summarize the computational challenges of this problem as follows:

• **Intractability** of the logarithmic integral function(s): there are two of these in general; nevertheless, in the case where $p_i^y(y_i \,|\, \theta)$ is the density function of a multivariate normal distribution with mean θ_i and covariance matrix V:

$$p_i^y(y_i \,|\, \theta) \;=\; \frac{1}{(2\pi)^{d/2} \sqrt{\det V}} \exp\left[\, -\tfrac{1}{2} \,(\, y_i - \theta_i \,)^\top V^{-1} (\, y_i - \theta_i \,) \,\right],$$

the function $a(\theta_i)$ is the constant $\log\left((2\pi)^{d/2} \sqrt{\det V} \right)$ that can removed from (8); yet the other integral function remains. While being intractable, the logarithmic function turns out to play an important role in the proposed solution method for the problem (8).

• **Nonconvexity** of the functions $g(y_i, \theta_i)$ and $h(\theta, \gamma)$: often, these functions are products of functions of their arguments; for example, $h(\theta, \gamma) = \gamma^\top v(\theta)$, for some $v : \Theta \to \mathbb{R}^L$, in the application to Markov Random Fields (MRFs); see the next subsection.

• **Nondifferentiability** of the functions $g(y_i, \theta_i)$ and $h(\theta, \gamma)$: when they involve distance functions as measures of deviation/similarity, as in the MRFs; see the next subsection.

• **Discontinuity** of the transformation function $\phi(u)$: when $\widetilde{\gamma}$ is a discrete random variable, the inverse cdf $F_{\widetilde{\gamma}_\ell}^{-1}$ is a step function; see Section 3. The approximation of such a discontinuous function by continuous functions easily leads to another source that renders (8) a nonconvex nondifferentiable optimization problem; added to the challenge is the fact that these two "non"-properties are coupled and embedded in the integral function.

2.2 A source application: Markov random fields with unknown topologies

In what follows, we use a graphical model of generalized Markov Random Fields (MRFs) to illustrate how BHM can easily give rise to the computational challenges we just introduced. Specifically, consider a MRF [3, 5] defined on an undirected graph with node set $\mathcal{V}$, where each node $i \in \mathcal{V}$ carries an m-dimensional random vector $\widetilde{\theta}_i$ whose components represent the features of an object of interest (e.g., pixels of an image) at this location. Thus here we have $M = |\mathcal{V}|$ and $d = M \times m$ using the previous notations. Conventionally, with a given edge set $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ and fixed nonnegative weights $\{\bar{\gamma}_{ij}\}_{(i,j) \in \mathcal{E}}$, an MRF represents our beliefs that features on pairs of nodes incident to a common edge should be similar, which are modeled through the following density function $\bar{q}$:

$$\begin{aligned} \bar{q}(\theta) \;&\triangleq\; \exp\left[\, \bar{h}(\theta) - \beta \,\right], \\ \text{where} \quad \bar{h}(\theta) \;&\triangleq\; -\sum_{(i,j) \in \mathcal{E}} \bar{\gamma}_{ij}\, h_{ij}(\theta_i, \theta_j) \quad \text{and} \quad \beta \triangleq \log \int_{\Theta} \exp\left(\bar{h}(\theta') \right)\, d\theta', \end{aligned} \tag{9}$$

and $h_{ij}(\theta_i, \theta_j)$ is a metric measuring the "similarity" between θ_i and θ_j, e.g., $h_{ij}(\theta_i, \theta_j) = \|\theta_i - \theta_j\|_2$, whose "strength" is encoded in $\bar{\gamma}_{ij}$. Thus through (9), the probability distribution of $\{\widetilde{\theta}_i\}_{i\in\mathcal{V}}$ is more concentrated if $\widetilde{\theta}_i$ and $\widetilde{\theta}_j$ take similar values when $(i,j) \in \mathcal{E}$. Based on (9), we have the following benchmark model which intends to recover the underlying $\widetilde{\theta}$ from observation $\{y_i\}_{i\in\mathcal{V}}$

$$\begin{array}{llcll} \text{level 1 :} & \widetilde{y}_i \,|\, \widetilde{\theta} & \overset{\text{ind.}}{\sim} & p_i^y(\, y_i \,|\, \widetilde{\theta}\,), & \text{for } i \in \mathcal{V} \\ \text{level 0 :} & \widetilde{\theta} & \sim & \overline{q}(\theta), & \text{as defined in (9)} \end{array} \tag{10}$$

Models of type (10) have broad applications in domains including image processing [6, 19] and disease mapping [28, 33]. However, the graph topology $\mathcal{E}$ and weights $\{\bar{\gamma}_{ij}\}_{(i,j)\in\mathcal{E}}$ in (10) can easily be misspecified if they are heuristically prescribed from data $\{y_i\}_{i\in\mathcal{V}}$. This can lead to an unreliable recovery of $\widetilde{\theta}$, especially when $\{y_i\}_{i\in\mathcal{V}}$ are considered as the noisy version of $\widetilde{\theta}$.

In light of this issue, a more reasonable approach is to incorporate $\mathcal{E}$ and $\{\bar{\gamma}_{ij}\}_{(i,j)\in\mathcal{E}}$ as part of inference via the proposal of the following generalized MRF using the notations in (7),

$$q(\theta \,|\, \gamma) \;=\; \exp\left[\, h(\theta, \gamma) - b(\gamma) \,\right], \quad \text{with} \quad h(\theta, \gamma) \;=\; -\sum_{i<j} \gamma_{ij}\, h_{ij}(\theta_i, \theta_j), \tag{11}$$

which gives us the following generalizations of the benchmark model (10):

- **MRF with unknown edges**

$$\begin{array}{llcll} \text{level 2} & \widetilde{y}_i \,|\, \widetilde{\theta} & \overset{\text{ind.}}{\sim} & p_i^y(y_i \,|\, \widetilde{\theta}), & \text{for } i \in \mathcal{V} \\ \text{level 1} & \widetilde{\theta} \,|\, \widetilde{\gamma} & \sim & q(\theta \,|\, \widetilde{\gamma}), & \text{as defined in (11)} \\ \text{level 0} & \widetilde{\gamma}_{ij} & \overset{\text{ind.}}{\sim} & \text{Bernoulli}(p_{ij}), & \text{where } p_{ij} \in [0,1] \text{ are fixed for } i < j. \end{array} \tag{12}$$

- **MRF with unknown weights**

$$\begin{array}{llcll} \text{level 2} & \widetilde{y}_i \,|\, \widetilde{\theta} & \overset{\text{ind.}}{\sim} & p_i^y(y_i \,|\, \widetilde{\theta}), & \text{for } i \in \mathcal{V} \\ \text{level 1} & \widetilde{\theta} \,|\, \widetilde{\gamma} & \sim & q(\theta \,|\, \widetilde{\gamma}), & \text{as defined in (11)} \\ \text{level 0} & \widetilde{\gamma}_{ij} & \overset{\text{ind.}}{\sim} & \text{Uniform}\left(\left[\, \underline{\gamma}_{ij}, \, \overline{\gamma}_{ij} \,\right]\right), & \text{where } \left(\underline{\gamma}_{ij}, \, \overline{\gamma}_{ij}\right) \in \mathbb{R}^2_+ \text{ are fixed for } i < j. \end{array} \tag{13}$$

Model (12) makes the assumption that the occurrence of an edge between (i,j) has probability p_{ij}, while (13) assumes that the "strength" of connection between (i,j) is unknown and to be estimated. These frameworks indeed generalize (10), e.g., if $\bar{\gamma}_{ij} = 1$ for all $(i,j) \in \mathcal{E}$ in (10), then model (12) is apparently more flexible by allowing us to learn the appearance of edges $\mathcal{E}$ from solving the corresponding o-BHOP. However, despite their enhanced modeling capability, the generalized MRF models (12) and (13) clearly introduce additional nonconvexity (from $h(\theta, \gamma)$ in (11)) and intractability (from $b(\gamma)$ in (11)) to the original o-BHOP associated with (10). Additionally, the o-BHOP associated with (12) will be a mixed zero-one program since $\widetilde{\gamma}_{ij}$ therein is binary valued. When coupled together, these challenges call for advanced computational methods for their resolution.

3 Discrete Priors

Motivated by and generalizing a Bernoulli random variable for $\widetilde{\gamma}_{ij}$ in the generalized MRF model (12) with unknown edges, we describe the transformation function $\phi(u)$ in (5) when each $\widetilde{\gamma}_\ell$ is a discrete random variable whose range $\{ \gamma_\ell^k \}_{k=1}^{K_\ell}$ satisfies

$$0 \triangleq \gamma_\ell^0 \leq \underbrace{\gamma_\ell^1 < \gamma_\ell^2 < \cdots < \gamma_\ell^{K_\ell}}_{\text{values of the random variable } \widetilde{\gamma}_\ell} \tag{14}$$

and the associated probabilities $\{ p_\ell^k \}_{k=1}^{K_\ell}$ are positive and sum up to unity. The cdf of $\widetilde{\gamma}_\ell$ is given by:

$$F_{\widetilde{\gamma}_\ell}(t) = \sum_{k\,:\,\gamma_\ell^k \leq t} p_\ell^k, \quad t \in (-\infty, \infty). \tag{15}$$

Moreover, we have

$$F_{\widetilde{\gamma}_\ell}^{-1}(s) \;=\; \gamma_\ell^1 + \sum_{k=2}^{K_\ell} \widehat{\gamma}_\ell^k \, \mathbf{1}_{(0,\infty)}(s - \widehat{p}_\ell^{k-1}), \quad \text{for } s \in [0,1], \tag{16}$$

where $\mathbf{1}_{(0,\infty)}(\bullet)$ is the indicator function of $(0,\infty)$ and

$$\widehat{\gamma}_\ell^k \triangleq \gamma_\ell^k - \gamma_\ell^{k-1} \geq 0 \text{ by (14)} \quad \text{and} \quad \widehat{p}_\ell^k \triangleq \sum_{i=1}^k p_\ell^i \quad \text{for} \quad k \in [K_\ell].$$

Note that $F_{\widetilde{\gamma}_\ell}^{-1}(\bullet)$ is lower semicontinuous (i.e., left-continuous), nondecreasing, piecewise constant on $(0,1)$, and by definition, right-continuous and left-continuous at the left and right end point of the interval, respectively. Following the schemes in the two references [9, 10], we approximate $F_{\widetilde{\gamma}_\ell}^{-1}(s)$ by continuous piecewise function employing

- a convex function $\widehat{\varphi}_{\text{cvx}} : \mathbb{R} \to \mathbb{R}$ and a concave function $\widehat{\varphi}_{\text{cve}} : \mathbb{R} \to \mathbb{R}$ satisfying

$$\widehat{\varphi}_{\text{cvx}}(0) = 0 = \widehat{\varphi}_{\text{cve}}(0) \quad \text{and} \quad \widehat{\varphi}_{\text{cvx}}(1) = 1 = \widehat{\varphi}_{\text{cve}}(1),$$

and with both (continuous) functions being increasing in the interval $[0,1]$ and nondecreasing outside.

Truncating these two functions to the range $[\,0,\,1\,]$, we obtain the upper and lower bounds of the two indicator functions $\mathbf{1}_{[\,0,\,\infty\,)}(t)$ and $\mathbf{1}_{(\,0,\,\infty\,)}(t)$ as follows: for any $(t,\delta) \in \mathbb{R} \times \mathbb{R}_{++}$,

$$\begin{aligned} \varphi_{\text{ub}}(t,\delta) &\triangleq \min\left\{ \max\left(\widehat{\varphi}_{\text{cvx}}\left(1 + \frac{t}{\delta}\right),\, 0 \right),\, 1 \right\} \\ &\geq \mathbf{1}_{[\,0,\,\infty\,)}(t) \geq \mathbf{1}_{(\,0,\,\infty\,)}(t) \\ &\geq \max\left\{ \min\left(\widehat{\varphi}_{\text{cve}}\left(\frac{t}{\delta}\right),\, 1 \right),\; 0 \right\} \triangleq \varphi_{\text{lb}}(t,\delta). \end{aligned} \tag{17}$$

we have the following result stated and proved in [9, Proposition 2].

Proposition 1. The bivariate functions φ_{ub} and φ_{lb} defined above have the following properties:

(a) For any $t \in \mathbb{R}$, $\varphi_{\text{ub}}(t,\delta)$ is a nondecreasing function in γ on $\mathbb{R}_{++}$ and $\varphi_{\text{lb}}(t,\delta)$ is a nonincreasing function in γ on $\mathbb{R}_{++}$. Both functions φ_{ub} and φ_{lb} are Lipschitz continuous on every compact set $T \times \Delta \subseteq \mathbb{R} \times \mathbb{R}_{++}$.

(b) The following equalities hold:

$$\begin{aligned} \mathbf{1}_{[\,0,\infty\,)}(t) &= \underset{\delta > 0}{\textbf{infimum}}\; \varphi_{\text{ub}}(t,\delta) = \lim_{\delta \downarrow 0} \varphi_{\text{ub}}(t,\delta), \quad \forall\, t \in \mathbb{R} \\ \text{and } \mathbf{1}_{(\,0,\infty\,)}(t) &= \underset{\delta > 0}{\textbf{supremum}}\; \varphi_{\text{lb}}(t,\gamma) = \lim_{\delta \downarrow 0} \varphi_{\text{lb}}(t,\delta), \quad \forall\, t \in \mathbb{R}. \end{aligned} \tag{18}$$

Applying $\varphi_{\rm ub}$ and $\varphi_{\rm lb}$ as the basic ingredients, we obtain the following approximations:

$$\begin{aligned} F^{-1}_{\widetilde{\gamma}_\ell}(u_\ell) &= \gamma^1_\ell + \sum_{k=2}^{K_\ell} \widehat{\gamma}^k_\ell \, \mathbf{1}_{(\,0,\infty\,)}(u_\ell - \widehat{p}^{\,k-1}_\ell) \\ &\geq \gamma^1_\ell + \sum_{k=2}^{K_\ell} \widehat{\gamma}^k_\ell \, \varphi_{\rm lb}\left(u_\ell - \widehat{p}^{\,k-1}_\ell, \delta \right) \triangleq \widehat{\varphi}_{\rm lb}(u_\ell, \delta) \end{aligned}$$

and

$$F^{-1}_{\widetilde{\gamma}_\ell}(u_\ell) \leq \gamma^1_\ell + \sum_{k=2}^{K_\ell} \widehat{\gamma}^k_\ell \, \varphi_{\rm ub}\left(u_\ell - \widehat{p}^{\,k-1}_\ell, \delta \right) \triangleq \widehat{\varphi}_{\rm ub}(u_\ell, \delta).$$

Therefore, for a function: $h(\theta, \gamma) = \gamma^\top \chi(\theta) = \sum_{\ell=1}^{L} \gamma_\ell v_\ell(\theta)$ where each v_ℓ is a nonpositive function (cf. (11)), it follows that its uniform-transformation $h(\theta, \phi(u))$ (as in (5) and (8)) satisfies

$$\underbrace{\sum_{\ell=1}^{L} \widehat{\varphi}_{\rm ub}(u_\ell, \delta) \, v_\ell(\theta)}_{\text{denoted } h_{\rm lb}(\theta, u)} \leq h(\theta, \phi(u)) = \sum_{\ell=1}^{L} F^{-1}_{\widetilde{\gamma}_\ell}(u_\ell) \, v_\ell(\theta) \leq \underbrace{\sum_{\ell=1}^{L} \widehat{\varphi}_{\rm lb}(u_\ell, \delta) \, v_\ell(\theta)}_{\text{denoted } h_{\rm ub}(\theta, u)}$$

More generally, provided that the function $h(\theta, \bullet)$ is *isotone* in the second argument (cf. e.g. (11)),

$$\gamma \leq \gamma' \Rightarrow h(\theta, \gamma) \leq h(\theta, \gamma')$$

we can bound the discontinuous function $h(\theta, \phi(u))$ by continuous functions. In the rest of the paper, we assume that such continuous bounding functions $h_{\rm lb}(\theta, u)$ and $h_{\rm ub}(\theta, u)$ have been derived for $h(\theta, \phi(u))$; needless to say, if the latter function is already continuous (for instance, if the cdf of each $\widetilde{\gamma}_\ell$ has a continuous inverse), then there is no need for such bounds.

Summarizing the above discussion, and assuming that $a(\theta_i)$ in (7) is a constant for notational simplicity, we obtain the following two problems that provide upper and lower bounds for (8):

$$\begin{aligned} & \underset{\theta \in \Theta;\ u \,\in\, [\,0,1\,]^L}{\textbf{minimum}} \ -\sum_{i=1}^{M} g(y_i, \theta_i) - h_{\rm ub}(\theta, u) + \log \int_\Theta \exp\left(h_{\rm lb}(\theta', u) \right) d\theta' \\ \leq \ & \underset{\theta \in \Theta;\ u \,\in\, [\,0,1\,]^L}{\textbf{minimum}} \ -\sum_{i=1}^{M} g(y_i, \theta_i) - h(\theta, \phi(u)) + \log \int_\Theta \exp\left(h(\theta', \, \phi(u) \right) d\theta' \quad \text{(problem (8))} \\ \leq \ & \underset{\theta \in \Theta;\ u \,\in\, [\,0,1\,]^L}{\textbf{minimum}} \ -\sum_{i=1}^{N} g(y_i, \theta_i) - h_{\rm lb}(\theta, u) + \log \int_\Theta \exp\left(h_{\rm ub}(\theta', u) \right) d\theta' \end{aligned}$$

The above inequalities are in terms of the global minimum objective values of the respective problems. Yet, the two bounding problems remain (highly) nonconvex and (often) nondifferentiable; moreover, they still contain the practically intractable logarithmic integral functions. To address the former "non"-features, we settle for a practically computable solution that satisfies a stationarity condition of some sort (to be defined in the next section). This modest goal is the general principle of our modern point of view of nonconvex nondifferentiable optimization detailed in the recent monograph [11]; that is, instead of the impossible task of computing global minimizers, we emphasize practical computability along with the validation of some stationarity conditions (i.e., necessary conditions for local optimality) satisfied by the computed solutions.

4 The Logarithmic Integral Optimiation Problem

As a unification of the upper and lower bounding minimization problems of (8), we consider the following nonconvex nondifferentiable optimization problem whose objective contains an intractable logarithmic integral function, which is the vanilla problem (2) with $r(x,z) = \exp(H(x,z))$ to highlight its connection to exponential families, although our methods in this section apply to general r:

$$\underset{x \in X}{\textbf{minimize}} \; c(x) + \log Z(x), \quad \text{where } \; Z(x) \triangleq \int_{\Xi} \exp(H(x,z))dz. \tag{19}$$

We impose the following blanket assumptions on (19):

1. **(Sets)** $X \subseteq \mathbb{R}^n$ is compact and convex; $\Xi \subseteq \mathbb{R}^d$ is compact.

2. **(Continuity)** Both $c : \mathcal{O} \to \mathbb{R}$ and $H : \mathcal{O} \times \mathcal{Z} \to \mathbb{R}$ are continuous functions, where $\mathcal{O}$ and $\mathcal{Z}$ are open sets containing X and Ξ. Hence $H(x,z)$ is Borel measurable in $z \in \Xi$ for all $x \in X$.

3. **(B-differentiability)** Functions c and $H(\bullet,z)$ (for all $z \in \Xi$) are *Bouligand differentiable* (B-differentiable); that is, they are locally Lipschitz continuous on $\mathcal{O}$ and their directional derivatives

$$c'(x;dx) \triangleq \lim_{\tau \downarrow 0} \frac{c(x+\tau dx) - c(x)}{\tau} \quad \text{and} \quad H(\bullet,z)'(x;dx) \triangleq \lim_{\tau \downarrow 0} \frac{H(x+\tau dx,z) - H(x,z)}{\tau}$$

exist for all $(x,dx) \in \mathcal{O} \times \mathbb{R}^n$. Lastly, we assume that a function $\text{Lip}_H : \mathcal{Z} \to \mathbb{R}_{++}$ exists satisfying

$$|\, H(x,z) - H(x',z) \,| \;\leq\; \text{Lip}_H(z) \,\|\, x - x' \,\|, \quad \forall\, x, x' \in \mathcal{O} \;\text{ and }\; z \in \mathcal{Z}$$

and $\displaystyle\int_{\Xi} \text{Lip}_H(z)\, dz < \infty$.

By the Dominated Convergence Theorem and Theorem 7.44 of [43], we obtain the following basic properties on function Z:

- Z is continuous on X and $Z(x) < \infty$ for all $x \in X$.

- Z, hence $\log Z$, is B-differentiable on X.

In terms of the directional derivative, we recall that a local minimizer of (19) must be directionally stationary; that is, if $\bar{x}$ is such a minimizer, then

$$c'(\bar{x}; x - \bar{x}) + (\, \log Z \,)'(\bar{x}; x - \bar{x}) \;\geq\; 0, \quad \forall\, x \,\in\, X$$

In the absence of a practical way to provably compute a local minimizer of a nonconvex nondifferentiable optimization problem, the practical goal of "solving" such a problem is to settle for the computation of a directional stationary solution. Even this less demanding task is not easy in general (see [11, Chapter 7]), and for (19) in particular. This task is complicated by the taunting, if not impossible, evaluation of $Z(x)$. As a necessary step to alleviate the latter, we propose a combination of the methods of *surrogation* [11, Chapter 7] and *adaptive importance sampling* [41]. The former substitutes the nonconvex functions c and $H(\bullet,z)$ at an arbitrary reference vector $\bar{x}$ by respective convex majoring functions that are the basis for the development of effective algorithms; the latter replaces the integral function by an equivalent expectation function which is then discretized into a finite sum via sampling from a judiciously chosen density function, and this is adaptively employed to control the variance of such discretization at each iteration. The end result of this fusion of deterministic surrogation and stochastic sampling allows us to design a practically implementable algorithm for computing a "stationary solution" of (19) of a certain kind.

4.1 Surrogation and stationarity

We assume that for every $(\bar{x}, z) \in X \times \Xi$, there exist convex $\widehat{c}(\bullet; \bar{x})$ and $\widehat{H}(\bullet, z; \bar{x})$ such that

1. **(Majorization)** $c(x) \leq \widehat{c}(x; \bar{x})$ and $H(x, z) \leq \widehat{H}(x, z; \bar{x})$ for all $x \in X$;
2. **(Touching)** $c(\bar{x}) = \widehat{c}(\bar{x}; \bar{x})$ and $H(\bar{x}, z) = \widehat{H}(\bar{x}, z; \bar{x})$;
3. **(Upper semicontinuity)** $\widehat{c}(\bullet, \bullet)$ and $\widehat{H}(\bullet, z; \bullet)$ are upper semicontinuous on $\mathcal{O} \times \mathcal{O}$;
4. **(Continuity)** $\widehat{H}(x, \bullet; \bar{x})$ is continuous hence Borel measurable on $\mathcal{Z}$ for all $(x, \bar{x}) \in \mathcal{O} \times \mathcal{O}$;

Denote $\widehat{Z}(x; \bar{x}) \triangleq \displaystyle\int_{\Xi} \exp(\widehat{H}(x, z; \bar{x}))\, dz$, then under the assumptions stated above, we can deduce the following properties of $\widehat{c}$ and $\widehat{Z}$ by Fatou's Lemma and Theorem 7.44 of [43]:

- $\widehat{Z}$ is upper semicontinuous on $X \times X$ and $\widehat{Z}(x; \bar{x}) < \infty$ for all $(x, \bar{x}) \in X \times X$.
- The following inequalities hold true for any $dx \in \mathbb{R}^n$:

$$\widehat{c}(\bullet; \bar{x})'(\bar{x}; dx) \;\geq\; c'(\bar{x}; dx) \quad \text{and} \quad (\,\log \widehat{Z}(\bullet; \bar{x})\,)'(\bar{x}; dx) \;\geq\; (\,\log Z\,)'(\bar{x}; dx), \tag{20}$$

The inequalities (20) pertain to **directional derivative majorization** of the surrogation functions. If equalities hold in (20), then the majoring functions $\widehat{c}(\bullet; \bar{x})$ and $\widehat{H}(\bullet, z; \bar{x})$ are said to be *directional derivative consistent* at $\bar{x}$.

With the convex surrogation functions $\widehat{c}(\bullet; \bar{x})$ and $\widehat{H}(\bullet, z; \bar{x})$, we may consider the surrogate optimization problem, for a given $\bar{x} \in X$ and $\rho > 0$:

$$\widehat{\mathbf{P}}_\rho(\bar{x}): \quad \underset{x \in X}{\textbf{minimize}}\ \widehat{c}(x; \bar{x}) + \log \int_{\Xi} \exp(\widehat{H}(x, z; \bar{x}))dz + \frac{1}{2\rho}\|x - \bar{x}\|_2^2 \tag{21}$$

The lemma below asserts that this is a convex program. For later purposes, we also include the same convexity of an empirical-average-version of the integral function. We omit the proof as it is a simple consequence of the renowned Jensen's inequality.

Lemma 2. Suppose that the bivariate function $e : X \times \Xi \to \mathbb{R}$ is such that $e(\bullet, z)$ is convex for all $z \in \Xi$ and $e(x, \bullet)$ is integrable on Ξ. Then the two functions

$$\log \int_{\Xi} \exp(e(\bullet, z))dz \quad \text{and} \quad \log \left(\frac{1}{N} \sum_{s=1}^{N} \exp(e(\bullet, \zeta^s)) \right)$$

are convex for all $\boldsymbol{\zeta}^N \triangleq \{\zeta^s\}_{s=1}^N$. □

The problem (21) allows us to define an important solution concept for the problem (19). Let $\widehat{\mathcal{M}}_\rho(\bar{x})$ denote the optimal solution set of problem $\widehat{\mathbf{P}}_\rho(\bar{x})$,

Definition 3. For given bivariate surrogation functions $\widehat{c}(\bullet; \bar{x})$ and $\widehat{H}(\bullet, z; \bar{x})$ of $c(\bullet)$ and $H(\bullet, z)$, respectively, satisfying the majorizaiton and touching conditions, a vector $\bar{x} \in X$ is a $(\widehat{c}, \widehat{H})$*-surrogation stationary point* of (19) if $\bar{x} \in \widehat{\mathcal{M}}_\rho(\bar{x})$ for some $\rho > 0$. □

Equivalently, the inclusion $\bar{x} \in \widehat{\mathcal{M}}_\rho(\bar{x})$ states that $\bar{x}$ is a fixed point of the "surrogation stationarity map":

$$\widehat{\mathcal{M}}_\rho : \bar{x} \in X \ \rightarrow \ \mathbf{argmin} \ \widehat{\mathbf{P}}_\rho(\bar{x}) \ \subseteq \ X.$$

The role of the surrogation stationarity concept is that it is a necessary condition for a local minimizer of (19) as asserted by the following simple result; the condition is sufficient for directional stationarity if the surrogation is directional derivative consistent. Being an immediate consequence of the directional derivative majorization (20), the result does not require a proof.

Proposition 4. Let $\bar{x}$ be a local minimizer of (19), then $\bar{x}$ is a $(\widehat{c}, \widehat{H})$-surrogation stationary point of (19) for any pair of bivariate functions $(\widehat{c}(\bullet; \bar{x}), \widehat{H}(\bullet, z; \bar{x}))$ that majorizes and touches $(c, H(\bullet, z))$ at $\bar{x}$ for all $z \in \Xi$. Conversely, if the pair of surrogation functions $(\widehat{c}(\bullet; \bar{x}), \widehat{H}(\bullet, z; \bar{x}))$ are directional derivative consistent at $\bar{x}$ for all $z \in \Xi$, and if $\bar{x}$ is a $(\widehat{c}, \widehat{H})$-surrogation stationary point of (19), then $\bar{x}$ is a directional stationary point of (19). □

While (21) is a convex program, its practical solution remains daunting, if not intractable, because of the challenging task of evaluating the multi-dimensional integral function. Thus it is necessary to approximate the latter function; as mentioned before, this is accomplished by the statistical technique of importance sampling to discretize the integration.

4.2 AIS-based surrogation method

For any bivariate function $e : \mathcal{O} \times \mathcal{Z} \rightarrow \mathbb{R}$ that is (Lebesgue) integrable in $z \in \mathcal{Z}$ for all $x \in \mathcal{O}$, and d-dimensional random vector $\widetilde{\zeta}$ with support Ξ and positive probability density function π, we have

$$Z_e(x) \triangleq \int_\Xi \exp(e(x,z))\, dz \ = \ \int_\Xi \pi(z) \frac{\exp(e(x,z))}{\pi(z)} dz \ = \ \mathbb{E}_{\widetilde{\zeta} \sim \pi} \left[\frac{\exp(e(x, \widetilde{\zeta}))}{\pi(\widetilde{\zeta})} \right]. \tag{22}$$

For a given x and a batch $\boldsymbol{\zeta}^N \triangleq \{\zeta^s\}_{s=1}^N$ of size N of iid samples drawn from distribution π, written as $\boldsymbol{\zeta}^N \overset{\text{iid}}{\sim} \pi$, the Sample Average Approximation (SAA) of $Z_e(x)$ is given by

$$Z_e(x) \ \approx \ Z_e^\pi(x, \boldsymbol{\zeta}^N) \ \triangleq \ \frac{1}{N} \sum_{s=1}^N \frac{\exp(e(x, \zeta^s))}{\pi(x, \zeta^s)}\,. \tag{23}$$

Applying the *importance sampling* (IS) reformulation (22) to function Z in (19) turns it into a stochastic program albeit of nonconvex and nondifferentiable type, which can either be naïvely approximated by SAA as shown in (23) or more rigorously handled by algorithms that combine surrogation and incremental SAA (see for instance [30] and Section 10.2 of [11]). It is worth noting that such treatments will achieve their respective convergence results for any arbitrary positive density π. A natural question arises as to what density π to choose if we desire a more efficient approximation from sampling. It turns out that this question can be answered based on the principle of importance sampling, which suggests to minimize the variance of the SAA. This minimizer is given by the following lemma.

Lemma 5. [41, Theorem 3.12] Let $f : \Xi \rightarrow \mathbb{R}$ be a (Lebesgue) integrable function with $|f| > 0$. Let

$$\int_\Xi f(z)\, dz \overset{\text{SAA}}{\approx\approx} I(\pi, \boldsymbol{\zeta}^N) \ \triangleq \ \frac{1}{N} \sum_{s=1}^N \frac{f(\zeta^s)}{\pi(\zeta^s)}$$

where $\boldsymbol{\zeta}^N \triangleq \{\zeta^s\}_{s=1}^N \overset{\text{iid}}{\sim} \pi$ and π is a positive density function. Then for all N

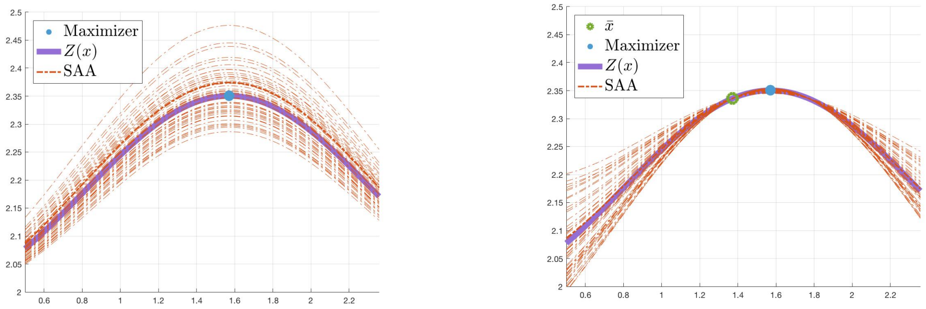
$$\pi_{\text{IS}}^f \;\triangleq\; \frac{|\,f\,|}{\displaystyle\int_\Xi |\,f(z)\,|\,dz} \;\in\; \underset{\pi}{\textbf{argmin}}\ \text{Var}_{\widetilde{\zeta}\sim\pi}\left[\frac{f(\widetilde{\zeta})}{\pi(\widetilde{\zeta})}\right] \;=\; \underset{\pi}{\textbf{argmin}}\ \text{Var}_{\boldsymbol{\zeta}^N\sim\pi}\left[\,I(\pi;\boldsymbol{\zeta}^N)\,\right],$$

where "$\text{Var}_{\widetilde{\zeta}\sim\pi}$" and "$\text{Var}_{\boldsymbol{\zeta}^N\sim\pi}$" denote the variance when the random variables follow π. Moreover, if f is positive, then $\text{Var}_{\widetilde{\zeta}\sim\pi_{\text{IS}}^f}\left[\dfrac{f(\widetilde{\zeta})}{\pi_{\text{IS}}^f(\widetilde{\zeta})}\right] = 0$; thus $\displaystyle\int_\Xi f(z)\,dz = I(\pi_{\text{IS}}^f, \boldsymbol{\zeta}^N)$ almost surely. □

The lemma above, when applied to function $f(\bullet) = \exp H(\bar{x}, \bullet)$ at a given $\bar{x} \in X$, indicates that the density achieving the minimal variance of estimating Z at $\bar{x}$ actually depends on the reference point $\bar{x}$. Thus, as far as variance reduction is concerned, the probability distribution in the stochastic programming reformulation (22) of (19) is implicitly decision dependent, as there exists the following density family parametrized by $\bar{x} \in X$

$$\pi_{\text{IS}}^{H(\bar{x},\bullet)}(z) \;\triangleq\; \frac{\exp(H(\bar{x},z))}{\displaystyle\int_\Xi \exp(H(\bar{x},z'))dz'} \;=\; \frac{\exp(H(\bar{x},z))}{Z(\bar{x})}. \tag{24}$$

whose associated SAA of type (23) equals to Z at $\bar{x}$. In other words, by explicitly controlling the variance, sampling from (24) and constructing (23) will yield a good approximation of the function Z locally around $\bar{x}$ with moderate sample size. To demonstrate this, suppose we approximate the function $Z(x) = \displaystyle\int_{-1}^{1} \exp\{-z\sin(x)\}dz$ with SAA of type (23), and we compare the variance of SAA from a π that is uniform on $[-1, 1]$ with $\pi = \pi_{\text{IS}}^{H(\bar{x},\bullet)}$ under a particular choice of $\bar{x}$. The variance is exemplified by 50 independent replications of SAA under the aforementioned two choices of π. As indicated by Figure 1a, when we apply the uniform π, the variance around a local maximizer of Z (the blue dot) is large even when the sample size N is $1,000$. On the contrary, from Figure 1b, we can obtain a better recovery of the landscape around the local maximizer (the blue dot in Figure 1b) with smaller variance if we adopt $\pi = \pi_{\text{IS}}^{H(\bar{x},\bullet)}$ with a $\bar{x}$ (e.g., the green dot in Figure 1b) that is near the local maximizer, and we are able to achieve this with fewer samples, e.g., $N = 100$.



(a) uniform π, $N = 1,000$

(b) $\pi = \pi_{\text{IS}}^{H(\bar{x},\bullet)}$, $N = 100$

Figure 1: Effectiveness of sampling from $\pi_{\text{IS}}^{H(\bar{x},\bullet)}$ with a fixed $\bar{x} \in \mathcal{X}$

This observation provides the following hints for designing an algorithm for (19) that iteratively solves subproblems which depends on IS-based SAA (23) of Z. Namely, instead of restricting to one

IS distribution for SAA throughout, in each iteration t we should make such distribution adapted to our current iterate x^t; more specifically, sampling from $\pi_{\text{IS}}^{H(x^t,\bullet)}$ so that we can benefit from its reduced variance. Combining such an adaptive IS scheme with the previously defined surrogation $\widehat{H}$ results in the following surrogated sampling approximation of $\log Z$ at a reference point $\bar{x} \in X$:

$$
\begin{array}{rcll}
\log Z(x) & \overset{\text{IS}}{=} & \log\left(\mathbb{E}_{\widetilde{\zeta}\sim\pi_{\text{IS}}^{H(\bar{x},\bullet)}} \left[\dfrac{\exp(H(x,\widetilde{\zeta}))}{\pi_{\text{IS}}^{H(\bar{x},\bullet)}(\widetilde{\zeta})} \right] \right) & \\[2ex]
& \leq & \log\left(\mathbb{E}_{\widetilde{\zeta}\sim\pi_{\text{IS}}^{H(\bar{x},\bullet)}} \left[\dfrac{\exp(\widehat{H}(x,\widetilde{\zeta};\bar{x}))}{\pi_{\text{IS}}^{H(\bar{x},\bullet)}(\widetilde{\zeta})} \right] \right), & \text{surrogation of } H(\bullet,\widetilde{\zeta}) \text{ by } \widehat{H}(\bullet,\widetilde{\zeta};\bar{x}) \\[2ex]
& \overset{\text{SAA}}{\approx\!\approx} & \log\left(\dfrac{1}{N} \displaystyle\sum_{s=1}^{N} \dfrac{\exp(\widehat{H}(x,\zeta^s;\bar{x}))}{\pi_{\text{IS}}^{H(\bar{x},\bullet)}(\zeta^s)} \right), & \text{draw } \boldsymbol{\zeta}^N \triangleq \{\zeta^s\}_{s=1}^{N} \overset{\text{iid}}{\sim} \pi_{\text{IS}}^{H(\bar{x},\bullet)} \\[2ex]
& = & \log\left(\dfrac{1}{N} \displaystyle\sum_{s=1}^{N} \dfrac{\exp(\widehat{H}(x,\zeta^s;\bar{x}))}{\exp(H(\bar{x},\zeta^s))} \right) + \underbrace{\log Z(\bar{x})}_{\text{constant given } \bar{x}}, & \begin{array}{l}\text{substituting out } \pi_{\text{IS}}^{H(\bar{x},\bullet)}(\widetilde{\zeta}) \\ \text{in the denominator.}\end{array}
\end{array}
$$

With the above constructions, we obtain the following approximation of problem (19), regularized by a proximal term with coefficient $\rho > 0$ to ensure its strict convexity hence uniqueness of solution

$$
\widehat{\mathbf{P}}_{\rho}^{N}(\bar{x};\boldsymbol{\zeta}^N): \quad \underset{x\in X}{\textbf{minimize}}\ \widehat{c}(x;\bar{x}) + \log\left(\frac{1}{N} \sum_{s=1}^{N} \frac{\exp(\widehat{H}(x,\zeta^s;\bar{x}))}{\pi_{\text{IS}}^{H(\bar{x},\bullet)}(\zeta^s)} \right) + \frac{1}{2\rho} \parallel x - \bar{x} \parallel_2^2. \tag{25}
$$

Problem (25) is the computational workhorse in the iterative algorithm described below. At each iteration, we employ the most recent iterate to define the reference vector $\bar{x}$; thus both the IS density $\pi_{\text{IS}}^{H(\bar{x},\bullet)}$ and surrogation are iterate dependent. The overall procedure is the promised AIS-based surrogation scheme that aims to compute a surrogation stationary solution of the original logarithmic integral optimization problem (19) corresponding to a given pair $(\widehat{c},\widehat{H})$ of surrogation functions; such a solution is a $x^\infty \in X$ such that $x^\infty \in \widehat{\mathcal{M}}_\gamma(x^\infty)$.

Algorithm for (19): AIS-based Surrogation Method

Initialization. Let $x^0 \in X$, $\rho > 0$, and a sequence of (positive) increasing integers $\{N_t\}_{t=0}^{\infty}$ be given. Set $t = 0$.

General Step. Given $x^t \in X$ and sample batch $\boldsymbol{\zeta}^{\,t} \triangleq \{\zeta^{st}\}_{s=1}^{N_t}$ with $\zeta^{st} \,|\, x^t \overset{\text{iid}}{\sim} \pi_{\text{IS}}^{H(x^t,\bullet)}$. Let x^{t+1} be the unique optimal solution of the problem $\widehat{\mathbf{P}}_{\rho}^{N_t}(x^t;\boldsymbol{\zeta}^{\,t})$, which is equivalent to

$$
\underset{x\in X}{\textbf{minimize}} \quad \underbrace{\widehat{c}(x;x^t) + \log\left(\frac{1}{N_t} \sum_{s=1}^{N_t} \frac{\exp(\widehat{H}(x,\zeta^{st};x^t))}{\exp(H(x^t,\zeta^{st}))} \right)}_{\text{convex in } x \text{ given } x^t} + \frac{1}{2\rho} \parallel x - x^t \parallel_2^2. \tag{26}
$$

Stop if a prescribed termination criterion is satisfied; otherwise, return to the general step with t replaced by $t + 1$. □

5 Analysis of the AIS-based Surrogation Algorithm

In this section, we present an analysis of the AIS-based surrogation algorithm which eventually establishes its almost sure subsequential convergence to a surrogation stationary point. While our adaptive choice of the IS density has the advantage of minimizing the (conditional) SAA variance at x^t for each iteration, it jeopardizes much of the typical analysis of SAA-based surrogation method (cf. [11, Theorem 10.2.1]), which relies critically on a uniform law of large numbers (ULLN) for iid samples. Since our iterate-dependent sampling process easily leads to a special *triangular array* of non-iid samples, our first order of business is to extend such laws, which are then applied to prove the convergence of our algorithm via two key steps, namely showing *sufficient descent* and *asymptotic fixed point stationarity*. Before further details, we summarize the notations that will be used in the analysis. For a given distribution with density π and samples $\zeta^s \overset{\text{iid}}{\sim} \pi$ for all $s = 1, \cdots, N$, define

$$\textbf{SAA of } Z\textbf{:}\quad \bar{Z}^N_\pi(x) \triangleq \frac{1}{N}\sum_{s=1}^{N} \frac{\exp(H(x,\zeta^s))}{\pi(\zeta^s)}$$

$$\textbf{Surrogation of } Z \textbf{ given } \bar{x}\textbf{:}\quad \widehat{Z}(x;\bar{x}) \triangleq \int_\Xi \exp(\widehat{H}(x,z;\bar{x}))dz \;=\; \mathbb{E}_{\zeta\sim\pi}\frac{\exp(\widehat{H}(x,\zeta;\bar{x}))}{\pi(\zeta)}$$

$$\textbf{Sampled surrogation of } Z \textbf{ given } \bar{x}\textbf{:}\quad \widehat{Z}^N_\pi(x;\bar{x}) \triangleq \frac{1}{N}\sum_{s=1}^{N} \frac{\exp(\widehat{H}(x,\zeta^s;\bar{x}))}{\pi(\zeta^s)}.$$

Additionally we write $\pi^*(\bar{x})$ as a shorthand for $\pi_{\text{IS}}^{H(\bar{x},\bullet)}$ when it appears in the subscripts.

5.1 A Digression: ULLN for triangular arrays

In this section, we extend ULLN to a special triangular array that generalizes our AIS scheme, which can be understood as the following data generating mechanism written in generic notations.

Definition 6. (AIS triangular array) For a subset $\mathcal{Y} \subseteq \mathbb{R}^n$ and a compact set $\mathcal{K} \subset \mathbb{R}^m$, let $\{\eta(\bullet, y) : \mathcal{K} \to \mathbb{R}_{++}\}_{y\in\mathcal{Y}}$ be a parametric family of probability density functions. We generate an array $\{\boldsymbol{z}^t\}_{t\geq 1}$ of samples according to the following process:

$$\begin{aligned} &\text{Step } 1 && \Longrightarrow \quad \text{fix an arbitrary } y^0 \in \mathcal{Y} \\ &\text{Step } t \geq 1 && \Longrightarrow \quad \text{given } y^t \in \mathcal{Y} \text{ sample } \boldsymbol{z}^t \triangleq \{z^{st}\}_{s=1}^{N_t} \text{ with } z^{st}\,|\,y^t \overset{\text{iid}}{\sim} \eta(\bullet, y^t), \\ & && \qquad\;\; \text{and set } y^{t+1} = \xi^t(\boldsymbol{z}^t, y^t) \in \mathcal{Y} \end{aligned} \tag{27}$$

where $\xi^t : \mathcal{K}^{N_t} \times \mathcal{Y} \to \mathcal{Y}$ is measurable and $\{N_t\}_{t\geq 1}$ is a prescribed sequence of positive integers. □

Note that samples $\{\boldsymbol{z}^t\}_{t\geq 1}$ are obviously not iid, instead $\{z^{st}\}_{s=1}^{N_t}$ are conditionally independent given y^t with conditional distribution $\eta(\bullet, y^t)$. In what follows, we establish several asymptotic results for such a triangular array. The first result is a strong LLN for the array $\{\varphi(z^{st}) : s \in [N_t], t \geq 1\}$ when φ is some function with a bounded range. We then generalize this result to a strong ULLN for triangular array of random functions. Finally, this strong ULLN is applied to prove a customized strong "one-sided" ULLN for a triangular array of random functions defined by our surrogate $\widehat{H}$.

Lemma 7. Let the following be given:

- an array $\{\boldsymbol{z}^t\}_{t\geq 1}$ and sequence $\{y^t\}_{t\geq 1}$ as generated in Definition 6;
- a measurable function $\varphi : \mathcal{K} \to \mathbb{R}$ such that $\sup\limits_{z\in\mathcal{K}} |\,\varphi(z)\,| < \infty$; and

• a sequence of positive integers $\{N_t\}_{t\geq 1}$ satisfying for some scalar $\kappa > 0$ and integer T_κ the condition that $N_t \geq \kappa t$ for all $t \geq T_\kappa$.

It then holds that

$$\lim_{t\to\infty} \left| \frac{1}{N_t} \sum_{s=1}^{N_t} \varphi(z^{st}) - \mathbb{E}_{\zeta\sim\eta(\bullet, y^t)}\, \varphi(\zeta) \right| \;=\; 0, \quad \text{almost surely.} \tag{28}$$

Proof. Denote $M_\varphi \triangleq \sup_{z\in\mathcal{K}} \varphi(z) - \inf_{z\in\mathcal{K}} \varphi(z) < \infty$, then for arbitrary $\bar{y} \in \mathcal{Y}$ and $\varepsilon > 0$ we have

$$\mathbb{P}\left(\left| \frac{1}{N_t} \sum_{s=1}^{N_t} \varphi(z^{st}) - \mathbb{E}_{\zeta\sim\eta(\bullet, y^t)}\varphi(\zeta) \right| \geq \varepsilon \;\middle|\; y^t = \bar{y} \right) \;\leq\; 2\exp\left\{ -\frac{2N_t\varepsilon^2}{M_\varphi^2} \right\} \tag{29}$$

holds due to the conditional independence among $\{z^{st}\}_{s=1}^{N_t}$, the fact that for all $s = 1, \ldots, N_t$ we have $\mathbb{E}(\varphi(z^{st}) - \mathbb{E}_{\zeta\sim\eta(\bullet,y^t)}\varphi(\zeta) \,|\, y^t = \bar{y}) = 0$, and virtually the same argument of proving the conventional Hoeffding's inequality. For all $t \geq 1$, denote $S^t \triangleq \{\boldsymbol{z}^1, \ldots, \boldsymbol{z}^t, y^1, \ldots, y^t\}$ and $\mathcal{F}_t$ as the σ-algebra generated by S^t with $\mathcal{F}_0 \triangleq \{\emptyset, \Omega\}$ where Ω is the underlying set of possible outcomes. By our data generating process and the definition of conditional probability distribution, from (29) the following holds true almost surely

$$\begin{aligned} &\mathbb{P}\left(\left| \frac{1}{N_t} \sum_{s=1}^{N_t} \varphi(z^{st}) - \mathbb{E}_{\zeta\sim\eta(\bullet, y^t)}\varphi(\zeta) \right| \geq \varepsilon \;\middle|\; \mathcal{F}_{t-1} \right) \\ =\;& \mathbb{P}\left(\left| \frac{1}{N_t} \sum_{s=1}^{N_t} \varphi(z^{st}) - \mathbb{E}_{\zeta\sim\eta(\bullet, y^t)}\varphi(\zeta) \right| \geq \varepsilon \;\middle|\; y^t \right) \quad \leq\; 2\exp\left\{ -\frac{2N_t\varepsilon^2}{M_\varphi^2} \right\} \end{aligned} \tag{30}$$

By the second Borel-Cantelli Theorem [14, Theorem 4.3.4], and our assumptions on N_t

$$\begin{aligned} 0 \;&=\; \mathbb{P}\left(\sum_{t=1}^{\infty} \mathbb{P}\left(\left| \frac{1}{N_t} \sum_{s=1}^{N_t} \varphi(z^{st}) - \mathbb{E}_{\zeta\sim\eta(\bullet, y^t)}\varphi(\zeta) \right| \geq \varepsilon \;\middle|\; \mathcal{F}_{t-1} \right) = \infty \right) \\ &=\; \mathbb{P}\left(\left| \frac{1}{N_t} \sum_{s=1}^{N_t} \varphi(z^{st}) - \mathbb{E}_{\zeta\sim\eta(\bullet, y^t)}\varphi(\zeta) \right| \geq \varepsilon \;\; \text{i.o.} \right) \end{aligned}$$

where "i.o." stands for infinitely often. Note that this is sufficient to show that

$$\left| \frac{1}{N_t} \sum_{s=1}^{N_t} \varphi(z^{st}) - \mathbb{E}_{\zeta\sim\eta(\bullet, y^t)}\varphi(\zeta) \right| \longrightarrow 0$$

almost surely. □

With Lemma 7, we can follow the same line of proof of [43, Theorem 7.48], i.e., ULLN of SAA, and obtain the following result for the $\{\boldsymbol{z}^t\}_{t\geq 1}$ of our interests. Details of the proof are omitted.

Proposition 8. In addition to the settings of Lemma 7, let $\mathcal{X}$ be a compact set in an Euclidean space and function $\psi : \mathcal{X} \times \mathcal{K} \to \mathbb{R}$ so that $\psi(\chi, \bullet)$ is measurable for all $\chi \in \mathcal{X}$ and there exists $M_\psi < \infty$ with $|\psi(\chi, z)| < M_\psi$ for all $(\chi, z) \in \mathcal{X} \times \mathcal{K}$. Then

$$\lim_{t\to\infty} \sup_{\chi\in\mathcal{X}} \left| \frac{1}{N_t} \sum_{s=1}^{N_t} \psi(\chi, z^{st}) - \mathbb{E}_{\zeta\sim\eta(\bullet,y^t)}\psi(\chi, \zeta) \right| = 0, \quad \text{almost surely.} \qquad \square$$

By [43, Theorem 7.48], for a given x^t, we can easily obtain error $\sup_{x\in X} \left| \widehat{Z}^{N_t}_{\pi^*(x^t)}(x, x^t) - \widehat{Z}(x, x^t) \right| \to 0$ almost surely as $N_t \to \infty$. However, as we will see in the analysis of AIS method in the subsequent sections, we need the rate for such convergence to be uniform in x^t, and the following proposition provides a "one-sided" version of this type of result.

Proposition 9. Let $\{x^t\}_{t=0}^{\infty}$ be as generated in the AIS-based surrogation algorithm and further assume that $\{N_t\}_{t\geq 0}$ is a sequence of positive integers satisfying for some scalar $\kappa > 0$ and integer T_κ the condition that $N_t \geq \kappa t$ for all $t \geq T_\kappa$. Then almost surely we can find a subsequence $\{x^t\}_{t\in\mathcal{T}}$ so that $x^t(t \in \mathcal{T}) \to x^\infty \in X$ and that for all $x \in X$

$$\limsup_{t(\in\mathcal{T})\to\infty} \widehat{Z}^{N_t}_{\pi^*(x^t)}(x; x^t) \;\leq\; \widehat{Z}(x; x^\infty)$$

Proof. See Appendix A.1 for details. $\square$

5.2 Convergence of deterministic surrogation algorithm

To make our discussion self-contained, we present the following convergence result for the deterministic surrogation method as a reference for the intuition behind the formal analysis of AIS-based surrogation. The main idea is that our method can be treated as a combination of deterministic surrogation and some stochastic error subjected to our sampling scheme. While the convergence of deterministic surrogation method can be straightforwardly established as below, what remains to be shown is that the accumulated stochastic error will diminish asymptotically.

Proposition 10. Let $x^0 \in X, \rho > 0$ be arbitrary and for all $t \geq 0$ let $x^{t+1} \in \widehat{\mathcal{M}}_\rho(x^t)$. Then the sequence $\{x^t\}_{t\geq 0}$ has an accumulation point $x^\infty \in X$ such that $x^\infty \in \widehat{\mathcal{M}}_\rho(x^\infty)$.

Proof. **Step 1 (sufficient descent)** By the majorization and touching properties of $(\widehat{c}, \widehat{H})$, and the optimality of x^{t+1} to the problem associated with $\widehat{\mathcal{M}}_\rho(x^t)$, we obtain

$$c(x^{t+1}) + \log Z(x^{t+1}) + \frac{1}{2\rho}\|x^{t+1} - x^t\|_2^2 \;\;\leq\;\; c(x^t) + \log Z(x^t) \tag{31}$$

By c and $\log Z$ being bounded from below, we can easily deduce that $\lim_{t\to\infty} \|x^{t+1} - x^t\|_2 = 0$.

Step 2 (asymptotic fixed-point stationarity) Fix an arbitrary $x \in X$, we have

$$c(x^{t+1}) + \log Z(x^{t+1}) + \frac{1}{2\rho}\|x^{t+1} - x^t\|_2^2 \;\;\leq\;\; \widehat{c}(x;\, x^t) + \log \widehat{Z}(x;\, x^t) + \frac{1}{2\rho}\|x - x^t\|_2^2 \tag{32}$$

from majorization and optimality of x^{t+1}. By compactness of X, we can restrict to a subsequence indexed by $\mathcal{T}$ so that $x^t \to x^\infty \in X$ when $t(\in \mathcal{T}) \to \infty$ and

$$\begin{aligned}
& \widehat{c}(x^\infty;\, x^\infty) + \log \widehat{Z}(x^\infty;\, x^\infty) + \frac{1}{2\rho}\|x^\infty - x^\infty\|_2^2 \\
= \quad & c(x^\infty) + \log Z(x^\infty) && \longleftarrow \text{ by touching} \\
\leq \quad & \widehat{c}(x;\, x^\infty) + \log \widehat{Z}(x;\, x^\infty) + \frac{1}{2\rho}\|x - x^\infty\|_2^2 && \longleftarrow \text{ by upper semicontinuity}
\end{aligned}$$

if we take "limsup" for $t(\in \mathcal{T}) \to \infty$ on both sides of (32), which gives us $x^\infty \in \widehat{\mathcal{M}}_\rho(x^\infty)$. □

5.3 Sufficient descent with stochastic error

Similar to the analysis of the deterministic surrogation method, the detailed convergence analysis of the AIS-based surrogation method consists of two main steps, the first of which is the sufficient descent property that aims to show $\|x^{t+1} - x^t\|_2 \to 0$ almost surely. Note that we can obtain the following inequality similar to (31), but with the SAA of the function Z instead:

$$c(x^{t+1}) + \log \bar{Z}^{N_t}_{\pi^*(x^t)}(x^{t+1}) + \frac{1}{2\rho}\|x^{t+1} - x^t\|_2^2 \quad \leq \quad c(x^t) + \log \bar{Z}^{N_t}_{\pi^*(x^t)}(x^t) \tag{33}$$

Define $e_t \triangleq \log \bar{Z}^{N_t}_{\pi^*(x^t)}(x^t) - \log \bar{Z}^{N_{t-1}}_{\pi^*(x^{t-1})}(x^t)$ and note that $\bar{Z}^{N_t}_{\pi^*(x^t)}(x^t) = Z(x^t)$ by the AIS scheme and Lemma 5; thus $e_t \leq e'_{t-1} \triangleq \sup\limits_{x \in X} \left|\log Z(x) - \log \bar{Z}^{N_{t-1}}_{\pi^*(x^{t-1})}(x)\right|$. For all $t \geq 1$ we get:

$$c(x^{t+1}) + \log \bar{Z}^{N_t}_{\pi^*(x^t)}(x^{t+1}) + \frac{1}{2\rho}\|x^{t+1} - x^t\|_2^2 \quad \leq \quad c(x^t) + \log \bar{Z}^{N_{t-1}}_{\pi^*(x^{t-1})}(x^t) + e'_{t-1} \tag{34}$$

As for the case when $t = 0$, we can let $\log \bar{Z}^{N_{-1}}_{\pi^*(x^{-1})}(x^0)$ as $\log Z(x^0)$ and $e'_{-1} \triangleq 0$. In this way, the last inequality (34) holds for all t.

To obtain $\|x^{t+1} - x^t\|_2 \to 0$ from (34), we essentially need error $e'_{t-1} \to 0$ fast enough in some appropriate notion of convergence. This can be ensured by the following conditions on $\{N_t\}_{t \geq 0}$, which we assume to be valid for the remaining analysis: $\displaystyle\sum_{t=0}^{\infty} \frac{1}{N_t^\alpha} < \infty$ for a fixed $\alpha \in \left(0, \frac{1}{2}\right)$. To show this, the following result is useful in providing a uniform (in $\bar{x} \in X$) non-asymptotic rate for the convergence of $\log \bar{Z}^N_{\pi^*(\bar{x})}$ to $\log Z$ on X when $N \to \infty$.

Lemma 11. For any $\alpha \in \left(0, \frac{1}{2}\right)$ there exists $C_\alpha > 0$ such that for all integer $N \geq 0$,

$$\mathbb{E}\left(\sup_{x \in X}\left|\log \bar{Z}^N_{\pi^*(\bar{x})}(x) - \log Z(x)\right|\right) \quad \leq \quad \frac{2C_\alpha}{N^\alpha}, \quad \text{for all } \bar{x} \in X \tag{35}$$

where the expectation is understood as the data in $\bar{Z}^N_{\pi^*(\bar{x})}$ having density $\pi_{\text{IS}}^{H(\bar{x}, \bullet)}$.

Proof. See Appendix A.2 for details. □

Denote by $\mathcal{D}^t$ the σ-algebra generated by $\cup_{\tau=0}^{t} \cup_{s_\tau=1}^{N_\tau} \{\zeta^{s_\tau \tau}\}$ for all $t \geq 0$, namely all the samples generated until step t, and define $\mathcal{D}^{-1} \triangleq \{\emptyset, \Omega\}$. Then the following holds for all $t \geq 0$ from (34):

$$\begin{aligned} & \mathbb{E}\left(c(x^{t+1}) + \log \bar{Z}^{N_t}_{\pi^*(x^t)}(x^{t+1}) \;\Big|\; \mathcal{D}^{t-1}\right) + \frac{1}{2\gamma}\mathbb{E}\left(\|x^{t+1} - x^t\|_2^2 \;\Big|\; \mathcal{D}^{t-1}\right) \\ \leq \quad & c(x^t) + \log \bar{Z}^{N_{t-1}}_{\pi^*(x^{t-1})}(x^t) + e'_{t-1} \end{aligned} \tag{36}$$

Furthermore, $\sum\limits_{\tau=-1}^{\infty} e'_\tau < \infty$ almost surely, otherwise $\mathbb{P}\left(\sum\limits_{\tau=-1}^{\infty} e'_\tau = \infty\right) > 0$ hence $\mathbb{E}\left(\sum\limits_{\tau=-1}^{\infty} e'_\tau\right) = \infty$; this contradicts the followings under our previous assumptions on $\{N_t\}_{t\geq 0}$:

$$\mathbb{E}\left(\sum_{\tau=0}^{\infty} e'_\tau\right) \;=\; \sum_{\tau=0}^{\infty}\mathbb{E}(e'_\tau) \;=\; \sum_{\tau=0}^{\infty}\mathbb{E}\left(\mathbb{E}(e'_\tau|\mathcal{D}^{\tau-1})\right) \;\leq\; \sum_{\tau=0}^{\infty}\frac{2C_\alpha}{N_\tau^\alpha} \;<\; \infty \tag{37}$$

The first equality of (37) is by $e'_\tau \geq 0$ and applying Theorem 2.15 in [16]. The second equality of (37) is by the law of total expectation. The first inequality of (37) is by x^0 being fixed, the fact that $\mathbb{E}(e'_0|\mathcal{D}^{-1}) = \mathbb{E}(e'_0)$, $\mathbb{E}(e'_\tau|\mathcal{D}^{\tau-1}) = \mathbb{E}(e'_\tau|x^\tau)$ for all $\tau \geq 1$ and Lemma 11 whose applicability is established by our definition of the conditional distribution $\zeta^{s\tau} \,|\, x^\tau \overset{\text{iid}}{\sim} \pi^*(\bullet, x^\tau) = \pi_{\text{IS}}^{H(x^\tau,\bullet)}$ for all $s \in [N_\tau]$. The second inequality of (37) is by our assumptions on $\{N_\tau\}_{\tau\geq 0}$. In sum, we almost surely have $\sum\limits_{\tau=-1}^{\infty} e'_\tau < \infty$ holds.

Finally, by $c(x^t) + \log \bar{Z}^{N_{t-1}}_{\pi^*(x^{t-1})}(x^t)$ being uniformly bounded from below, we can without loss of generality combine the result that $\sum\limits_{\tau=-1}^{\infty} e'_\tau < \infty$ almost surely with (36) and apply the Robbins-Siegmund nonnegative almost supermartingale convergence theorem [40, Theorem 1] to conclude that the following holds almost surely

$$\sum_{t=0}^{\infty}\mathbb{E}\left(\|x^{t+1} - x^t\|_2^2 \;\Big|\; \mathcal{D}^{t-1}\right) \;<\; \infty$$

This in turns gives us $\|x^{t+1} - x^t\|_2 \to 0$ almost surely by a straightforward application of conditional Markov inequality and the second Borel-Cantelli lemma [14, Theorem 5.3.2].

5.4 Asymptotic fixed-point stationarity

Similar to Step 2 in the analysis of Proposition 10, we can fix an arbitrary $x \in X$ and get:

$$c(x^{t+1}) + \underbrace{\log \bar{Z}^{N_t}_{\pi^*(x^t)}(x^{t+1})}_{\text{Term I}} + \frac{1}{2\rho}\|x^{t+1} - x^t\|_2^2 \;\leq\; \widehat{c}(x; x^t) + \underbrace{\log \widehat{Z}^{N_t}_{\pi^*(x^t)}(x; x^t)}_{\text{Term II}} + \frac{1}{2\rho}\|x - x^t\|_2^2 \tag{38}$$

The key intuition is that restricted to a subsequence $\mathcal{T}$ such that $x^t \to x^\infty \in X$, when $t(\in \mathcal{T}) \to \infty$ (due to the compactness of X), Term I in (38) should converge to $\log Z(x^\infty) = \log \widehat{Z}(x^\infty; x^\infty)$ and the limit of Term II in (38) should be upper bounded by $\log \widehat{Z}(x; x^\infty)$. More precisely, since Term I and II can be understood as the SAA of their respective conditional expectation, when $N_t \to \infty$ as we let $t(\in \mathcal{T}) \to \infty$, the law of large number that dominates the convergence of these SAAs should be uniform in x^t. In what follows, Term I and II will be analyzed under such overarching goals.

Term I Denote $\psi(\bar{x}, y, z) \triangleq Z(y)\exp\{H(\bar{x}, z) - H(y, z)\}$ so that

$$\bar{Z}^{N_t}_{\pi^*(x^t)}(x^{t+1}) = \frac{1}{N_t}\sum_{s=1}^{N_t} \psi(x^{t+1}, x^t, \zeta^{st}), \quad Z(x^{t+1}) = \mathbb{E}_{\zeta\sim\pi^*(x^t)}\left(\psi(x^{t+1}, x^t, \zeta)\right)$$

Note that by our assumptions on $\{N_t\}_{t\geq 0}$, Proposition 8 is applicable to ψ hence for almost every $\omega \in \Omega$ (Ω being the set of possible outcomes) we can restrict to a subsequence indexed by $\mathcal{T}(\omega)$ so that $x^\infty(\omega) \triangleq \lim_{t(\in\mathcal{T}(\omega))\to\infty} x^t(\omega) \in X$ and the followings hold:

$$\lim_{t(\in\mathcal{T}(\omega))\to\infty} \sup_{(\bar{x},y)\in X\times X} \left| \frac{1}{N_t}\sum_{s=1}^{N_t} \psi(\bar{x}, y, \zeta^{st}(\omega)) - E_{\zeta\sim\pi^*(x^t(\omega))}\left(\psi(\bar{x}, y, \zeta)\right) \right| = 0 \tag{39}$$

Also note that after some simple operations:

$$\begin{aligned} & \left| \bar{Z}^{N_t}_{\pi^*(x^t(\omega))}(x^{t+1}(\omega)) - Z(x^\infty(\omega)) \right| \\ \leq\ & \underbrace{\sup_{(\bar{x},y)\in X\times X} \left| \frac{1}{N_t}\sum_{s=1}^{N_t} \psi(\bar{x}, y, \zeta^{st}(\omega)) - E_{\zeta\sim\pi^*(x^t(\omega))}\left(\psi(\bar{x}, y, \zeta)\right) \right|}_{\text{Term I.1}} + \underbrace{\left| Z(x^{t+1}(\omega)) - Z(x^\infty(\omega)) \right|}_{\text{Term I.2}} \end{aligned}$$

When we take $t(\in\mathcal{T}(\omega)) \to \infty$, Term I.1 of this inequality goes to zero by (39) and Term I.2 also goes to zero by an application of Dominated Convergence Theorem. Thus

$$\lim_{t(\in\mathcal{T}(\omega))\to\infty} \bar{Z}^{N_t}_{\pi^*(x^t(\omega))}(x^{t+1}(\omega)) = Z(x^\infty(\omega))$$

In sum if we take $t(\in\mathcal{T}(\omega)) \to \infty$, the left hand side of (38) goes to

$$c(x^\infty(\omega)) + \log Z(x^\infty(\omega)) = \widehat{c}(x^\infty(\omega); x^\infty(\omega)) + \log \widehat{Z}(x^\infty(\omega); x^\infty(\omega))$$

since from the sufficient descent analysis $\|x^{t+1}(\omega) - x^t(\omega)\|_2 \to 0$ as $t(\in\mathcal{T}(\omega)) \to \infty$.

Term II By our assumptions we can treat logarithm as Lipschitz continuous with constant $\widetilde{L}$ hence

$$\log \widehat{Z}^{N_t}_{\pi^*(x^t(\omega))}(x; x^t(\omega)) \leq \widetilde{L}\left(\widehat{Z}^{N_t}_{\pi^*(x^t(\omega))}(x; x^t(\omega)) - \widehat{Z}(x; x^\infty(\omega))\right) + \log \widehat{Z}(x; x^\infty(\omega)) \tag{40}$$

Given the assumptions on $\{N_t\}_{t\geq 0}$, Proposition 9 is applicable so that without loss of generality

$$\limsup_{t(\in\mathcal{T}(\omega))\to\infty} \widehat{Z}^{N_t}_{\pi^*(x^t(\omega))}(x; x^t(\omega)) - \widehat{Z}(x; x^\infty(\omega)) \leq 0$$

If we apply this to (40) and subsequntly to the right hand side of (38) we can obtain:

$$\begin{aligned} & \widehat{c}(x^\infty(\omega); x^\infty(\omega)) + \log \widehat{Z}(x^\infty(\omega); x^\infty(\omega)) + \frac{1}{2\rho}\|x^\infty(\omega) - x^\infty(\omega)\|_2^2 \\ \leq\ & \widehat{c}(x; x^\infty(\omega)) + \log \widehat{Z}(x; x^\infty(\omega)) + \frac{1}{2\rho}\|x - x^\infty(\omega)\|_2^2 \end{aligned}$$

namely, $x^\infty(\omega) \in \mathop{\mathbf{argmin}}_{\bar{x}\in X}\ \widehat{c}(\bar{x}; x^\infty(\omega)) + \log \widehat{Z}(\bar{x}; x^\infty(\omega)) + \frac{1}{2\rho}\|\bar{x} - x^\infty(\omega)\|_2^2 = \widehat{M_\rho}(x^\infty(\omega))$.

5.5 The main theorem with discussion

Combining the analysis presented in Section 5.3 and 5.4, we formalize the following conclusion which guarantees an almost sure subsequential convergence of the AIS-based surrogation method to a surrogation stationary point of the logarithmic integral optimization problem (19).

Theorem 12. Let the following be given:

- A pair $(\widehat{c}, \widehat{H})$ of surrogate functions satisfying the conditions in Section 4.1;
- A sequence of positive integers $\{N_t\}_{t\geq 0}$ such that $\displaystyle\sum_{t=0}^{\infty} \frac{1}{N_t^{\alpha}} < \infty$ for some $\alpha \in \left(0, \frac{1}{2}\right)$.

Then with probability one, the sequence $\{x^t\}_{t\geq 0}$ produced by the AIS-based surrogation method has an accumulation point $x^{\infty} \in X$ such that x^{∞} is a $(\widehat{c}, \widehat{H})$-surrogation stationary point of (19). □

Remark 13. We have the following comments pertaining to Theorem 12:

1. By Proposition 4, the accumulation point x^{∞} will be a directional stationary point of (19) if the selected surrogation $(\widehat{c}, \widehat{H})$ satisfies the condition that $(\widehat{c}(\bullet; x), \widehat{H}(\bullet, z; x))$ is directional derivative consistent at x for all $(x, z) \in X \times \Xi$, namely, $\widehat{c}(\bullet; x)'(x; dx) = c'(x; dx)$ and $\widehat{H}(\bullet, z; x)'(x; dx) = H'(x, z; dx)$, which ensures $(\log \widehat{Z}(\bullet; \bar{x}))'(\bar{x}; dx) = (\log Z)'(\bar{x}; dx)$, for any $dx \in \mathbb{R}^n$.

2. From Theorem 12, a sequence of N_t that grows faster than t^2 should be sufficient to guarantee our convergence result. However, as indicated by our numerical results, such estimation of sample complexity can be too conservative in practice. In Appendix B, we present a modified version of our scheme which suggests a N_t that increases linearly in t. This alternative scheme requires ρ to be small enough instead of any arbitrary positive value, and uses $\|x - x^t\|_2$ for the regularization term instead of its square in subproblem (26). However, this modification should be regarded as a theoretical insight into improving sample complexity at the cost of a sacrificed practical utility, due to the restricted "step-size" ρ and the use of nondifferentiable subproblems which are more challenging to solve. Therefore, in the subsequent experiments, we opt to implement only the original AIS-based surrogation scheme, as its practical effectiveness requires a more manageable sample size than what the theoretical results suggest, as we just mentioned. □

6 Numerical Experiments

In this section, we test the numerical performance of our AIS-based surrogation method (abbreviated as "AIS" from now on) with two driving goals. First, we aim to demonstrate the advantage of AIS when compared with the following two approaches for (19):

- **Sample Average Approximation (SAA):** We substitute the integral term in (19) by the SAA of type (23) under a prescribed density π_{SAA} and sample size S, and the remaining problem

$$\operatorname*{\mathbf{minimize}}_{x\in X} \quad c(x) + \log\left(\frac{1}{S}\sum_{s=1}^{S} \frac{\exp\left(H(x, \zeta^s)\right)}{\pi_{\text{SAA}}(\zeta^s)}\right) \quad = \quad c(x) + \log\left(\bar{Z}^S_{\pi_{\text{SAA}}}(x)\right) \tag{41}$$

is handled by a surrogation method which iteratively solves (with a fixed $x^0 \in X, \rho > 0$)

$$x^{t+1} \;\in\; \widehat{\mathcal{M}}^{\text{SAA}}_{\rho}(x^t; S) \;\triangleq\; \operatorname*{\mathbf{argmin}}_{x\in X} \;\; \widehat{c}(x; x^t) \,+\, \log\left(\widehat{Z}^S_{\pi_{\text{SAA}}}(x; x^t)\right) \,+\, \frac{1}{2\rho}\|x - x^t\|_2^2$$

and produces a sequence $\{x^t\}_{t\geq 0}$ with an accumulation point x^{SAA} satisfying $x^{\text{SAA}} \in \widehat{\mathcal{M}}_{\rho}^{\text{SAA}}(x^{\text{SAA}}; S)$. The proof is virtually the same as Proposition 10 hence omitted here.

• **Stochastic Majorization Minimization (SMM):** With a prescribed continuous density function π_{SMM} throughout the procedure, we have the following two implementations of SMM:

— **Non-incremental SMM:** this is essentially the same as AIS method except for now we sample non-adaptively from π_{SMM} in each step t with sample size N_t identical to that of AIS.

— **Incremental SMM:** in each step we still draw N_t iid samples from π_{SMM} but now we apply all the samples up to step t to construct subproblem (26). Convergence result of SMM under this setting can be found in Theorem 10.2.1. of [11], where the same subsequential convergence as in Theorem 12 can be established if sequence N_t satisfies the followings

$$\sum_{t=1}^{\infty} \frac{N_t - N_{t-1}}{N_t N_{t-1}^{\alpha}} < \infty, \quad \text{for some } \alpha \in \left(0, \frac{1}{2}\right).$$

Through the applications of AIS, SMM and SAA on the same set of generic problems under various configurations, we verify that AIS outperforms the other methods in terms of computing time, objective value at termination, and stability, namely the variance in solutions, near convergence.

Our second goal is to show that AIS, as a suitable method for MAP inference of BHMs with intractable normalizers, enables the practical benefits of such models that are otherwise handicapped by naïve treatments of the intractable normalizers, e.g., model (9) as a simplificaiton of (11). Finally, all the numerical experiments are conducted on a Mac OS X personal computer with 2.3 GHz Intel Core i7 and 8 GB RAM. The reported times are in seconds on this computer.

6.1 Comparisons with SMM and SAA

This section compares the performance of AIS, SMM and SAA by applying them to randomly generated instances of the o-BHOP (4) for model (13) under the following specifications:

• Both y_i and $\widetilde{\theta}_i$ are scalar valued, i.e., $m = 1$, hence vectors y and $\widetilde{\theta}$ are both $|\mathcal{V}| = M$-dimensional, and $p_i^y(y_i \,|\, \widetilde{\theta})$ in (13) is Gaussian with mean $\widetilde{\theta}_i$ and a known variance $\sigma_i^2 > 0$.

• Density $q(\theta \,|\, \widetilde{\gamma})$ in (13) is specified with $h_{ij}(\theta_i, \theta_j) = |\theta_i - \theta_j|$.

The o-BHOP (4) under these settings is formulated as:

$$\begin{aligned} \underset{\theta,\gamma}{\textbf{minimize}} \quad & \sum_{i=1}^{M} \frac{1}{2\sigma_i^2}(y_i - \theta_i)^2 + \sum_{i<j} \gamma_{ij}|\theta_i - \theta_j| + \log \underbrace{\left(\int_{\Theta} \exp\left\{ -\sum_{i<j} \gamma_{ij} |\theta_i' - \theta_j'| \right\} d\theta' \right)}_{\triangleq\, Z(\gamma)} \qquad (42) \\ \textbf{subject to} \quad & \theta \in \Theta = [0,1]^M, \quad \underline{\gamma}_{ij} \leq \gamma_{ij} \leq \overline{\gamma}_{ij}, \, \forall i < j. \end{aligned}$$

We set M to be 10 and 20, resulting in $D = 55$ and $D = 210$ total variables in (42), respectively. Data y is generated from model (13), with $\sigma_i = 0.1$ for all $i \in \mathcal{V}$, and $\underline{\gamma}_{ij} = 0$ and $\bar{\gamma}_{ij} = 0.01$ for all $i < j$, except for two randomly chosen pairs whose upper bound $\bar{\gamma}_{ij}$ is set to be 100.

When we implement AIS, we adopt sample size $N_t = \min\{t^{1.2}, \bar{N}\}$ where $\bar{N}$ is a predetermined positive integer. Given the current γ^t, we apply Metropolis-Hastings method [41] to draw conditionally independent samples $\{z^{st}\}_{s=1}^{N_t}$ from density function (in z): $\dfrac{1}{Z(\gamma^t)} \exp\left\{ -\sum_{i<j} \gamma_{ij}^t |z_i - z_j| \right\}$, and then construct AIS subproblem (26) whose formulation is specified as (46) in Appendix C.1.

On the other hand, the sampling distribution π_{SMM} for both incremental and non-incremental SMM are fixed as uniform over set Θ. In each step t, given iid samples from π_{SMM}, we solve subproblems whose formulations under the two SMM schemes can be found in (46) of Appendix C.1. Finally, for SAA we set π_{SAA} as uniform over Θ and for a fixed sample size S we solve (41) with the surrogation method we introduced in the beginning of this section.

Throughout our experiments, we test AIS, SMM and SAA with $\rho = 100$, $\bar{N} \in \{10, 20, 50, 100\}$ and $S \in \{100, 1000, 10000\}$, and the detailed configurations are summarized in Table 1.

Methods	Configurations
AIS under a $(M, \bar{N})$	3 instances × 5 initializations
SMM under a $(M, \bar{N})$	3 instances × 5 initializations
SAA under a (M, S)	3 instances × 5 initializations × 3 replications

Table 1: Experiment configurations

For SAA, we refer to distinct batches of S samples from π_{SAA} as "replications". For each M, we use "3 instances" to denote three randomly generated problems that are shared by different methods under varying arrangements, including different choices of $\bar{N}$ and S, as well as different initializations and replications. All the statistics shown in Table 2 are averaged over the total number of runs under a particular pair of $(M, \bar{N})$ or (M, S): 15 for AIS and SMM, 45 for SAA. Finally, all the methods are terminated when two stopping criteria are met, namely when the relative difference between two consecutive solutions are small enough (we use notation x to represent (θ, γ) in what follows):

$$\frac{\|x^t - x^{t-1}\|}{\|x^{t-1}\|} < \varepsilon_{\text{sol}}, \text{ values of } \varepsilon_{\text{sol}} : \left[\begin{array}{c|cccc} & \bar{N}=10 & \bar{N}=20 & \bar{N}=50 & \bar{N}=100 \\ \hline M=10 & \text{7e-5} & \text{2e-5} & \text{2e-5} & \text{2e-5} \\ M=20 & \text{1e-4} & \text{1e-4} & \text{1e-4} & \text{1e-4} \end{array}\right]$$

and the changes in objective are relatively small among the three most recent steps:

$$\frac{1}{3}\sum_{\tau=t-2}^{t} \frac{\left|\widehat{f}(x^{\tau}) - \widehat{f}(x^{\tau-1})\right|}{\left|\widehat{f}(x^{\tau-1})\right|} < \varepsilon_{\text{obj}}, \text{ values of } \varepsilon_{\text{obj}} : \left[\begin{array}{c|cccc} & \bar{N}=10 & \bar{N}=20 & \bar{N}=50 & \bar{N}=100 \\ \hline M=10 & \text{2e-5} & \text{1e-5} & \text{1e-5} & \text{1e-5} \\ M=20 & \text{2e-5} & \text{1e-5} & \text{1e-5} & \text{1e-5} \end{array}\right]$$

where $\widehat{f}$ is the objective of (42) with $Z(\gamma)$ approximated by 10^5 predetermined iid uniform samples on Θ. Additionally, we terminate if these two stopping rules are not satisfied at step 100.

All the test results are summarized in Table 2, in which "Tot. Time" stands for how long it takes for the algorithms to terminate, "Samp. Time" and "Sol. Time" represents the time that is purely spent on sampling and solving the subproblems, "Obj." is the approximated objective $\widehat{f}$ value at termination, and "Steps" records the number of subproblems we solve before we stop. Incremental and non-incremental SMM are referred to as "SMM (inc.)" and "SMM (non-inc.)" in the table.

First, from Table 2, the final objective attained by AIS is generally better than the two SMM schemes. While a larger $\bar{N}$ does not significantly affect the objective value at termination, it does increase computing time. Additionally, both AIS and incremental SMM can be stopped within 100 steps but the former requires less steps to terminate and it is apparently faster than the latter, e.g., twice as fast when $M = 20, \bar{N} = 10$. On the other hand, for all sizes of $\bar{N}$, non-incremental SMM is not capable of meeting the stopping criteria within 100 iterations. In fact, as demonstrated by the curves of approximated objective $\widehat{f}$ vs. steps from AIS and the two SMM schemes when they are applied to a typical case with $M = 20, \bar{N} = 10$ for 100 steps (see Figure 2), non-incremental SMM oscillates too significantly for us to conclude its convergence while AIS has the most stable performance. It is worth noting that AIS achieves all the aforementioned properties despite spending a substantially longer time on sampling, as the distribution we sampled from are more complex than simple uniform law. However, a further decrease in AIS sampling time can be expected if we adopt a more adequate sampling method than the current naïve implementation of Metropolis Hastings.

$M = 10, D = 55$						
Methods		Tot. Time	Samp. Time	Sol. Time	Obj.	Steps
SAA	$S = 100$	12.99	0.01	12.98	-4.627	26
	$S = 1000$	12.33	0.02	12.31	-4.633	15
	$S = 10000$	75.57	0.12	75.45	-4.635	15
AIS	$\bar{N} = 10$	7.87	2.50	5.37	-4.635	12
	$\bar{N} = 20$	11.76	4.59	7.17	-4.636	16
	$\bar{N} = 50$	13.04	5.79	7.25	-4.636	17
	$\bar{N} = 100$	16.31	7.48	8.83	-4.636	20
SMM (inc.)	$\bar{N} = 10$	11.60	0.01	11.59	-4.632	23
	$\bar{N} = 20$	18.36	0.01	18.35	-4.633	34
	$\bar{N} = 50$	20.03	0.02	20.02	-4.634	33
	$\bar{N} = 100$	28.15	0.03	28.12	-4.635	41
SMM (non-inc.)	$\bar{N} = 10$	41.39	0.02	41.37	-4.637	100
	$\bar{N} = 20$	43.22	0.03	43.19	-4.637	100
	$\bar{N} = 50$	44.39	0.06	44.33	-4.636	100
	$\bar{N} = 100$	45.97	0.10	45.68	-4.636	100

Table 2: Comparison of AIS, SMM and SAA

$M = 20, D = 210$						
Methods		Tot. Time	Samp. Time	Sol. Time	Obj.	Steps
SAA	$S = 100$	31.11	0.01	31.11	-6.228	29
	$S = 1000$	68.97	0.02	68.95	-6.251	26
	$S = 10000$	635.19	0.11	635.08	-6.262	30
AIS	$\bar{N} = 10$	24.51	5.23	19.28	-6.263	23
	$\bar{N} = 20$	33.10	9.41	23.69	-6.264	27
	$\bar{N} = 50$	39.45	13.84	25.62	-6.264	30
	$\bar{N} = 100$	43.91	15.98	27.94	-6.264	31
SMM (inc.)	$\bar{N} = 10$	41.37	0.01	41.36	-6.241	34
	$\bar{N} = 20$	120.51	0.02	120.49	-6.253	63
	$\bar{N} = 50$	165.91	0.03	165.88	-6.258	58
	$\bar{N} = 100$	316.44	0.06	316.38	-6.258	68
SMM (non-inc.)	$\bar{N} = 10$	89.03	0.02	89.00	-6.256	100
	$\bar{N} = 20$	87.19	0.04	87.16	-6.257	100
	$\bar{N} = 50$	92.55	0.06	92.49	-6.257	100
	$\bar{N} = 100$	100.05	0.11	99.94	-6.258	100

Table 2: Comparison of AIS, SMM and SAA (continued)

The advantage of AIS over both SMM schemes is endowed by its adaptive sampling scheme, which provides accurate approximations of the objective with smaller variance (at least locally around the current (θ^t, γ^t)), even with a small sample size such as $\bar{N} = 10$. This is particularly crucial as most computations in all methods are spent on solving their respective subproblems, whose expenses heavily depend on the sample size. Thus, AIS achieves a lighter computation than incremental SMM, by requiring fewer samples to retain an effective approximation in each step. On the other hand, non-incremental SMM, which applies the same sample size as AIS without accumulation, suffers from

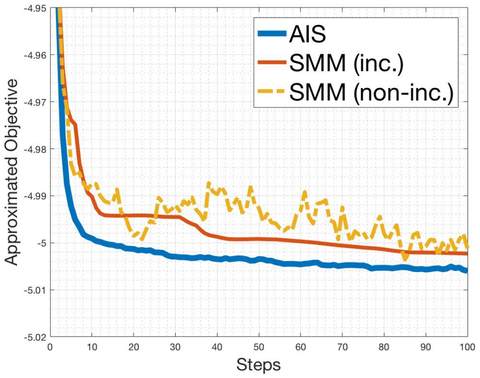


Figure 2: AIS vs. SMM

Methods	$\bar{N} = 10$		$\bar{N} = 20$	
	Var. Obj.	Var. Sol.	Var. Obj.	Var. Sol.
AIS	2.2e-4	9.3e-4	1.5e-4	1.1e-3
SMM (inc.)	2.4e-3	4.2e-3	2.4e-4	5.7e-3
SMM (non-inc.)	1.0e-3	1.2e-2	1.2e-3	3.6e-3
Methods	$\bar{N} = 50$		$\bar{N} = 100$	
	Var. Obj.	Var. Sol.	Var. Obj.	Var. Sol.
AIS	1.3e-4	6.3e-4	1.2e-4	9.7e-4
SMM (inc.)	1.9e-3	4.8e-3	1.5e-3	4.9e-3
SMM (non-inc.)	7.6e-4	2.8e-3	5.0e-4	2.2e-3

Table 3: Stability of AIS over SMM

a larger variance at each step. This is the key reason behind the smoother and more stable behavior of AIS compared to non-incremental SMM, as seen in Figure 2. Another interesting observation from Figure 2 is that the approximated objective curve of AIS is nearly non-increasing. It is known that the deterministic surrogation method as described in Proposition 10, has the property of non-increasing objective along the iterations. Therefore, if we view AIS as an inexact surrogation method, the near monotonicity of the AIS curve further supports its approximation accuracy.

Table 3 presents some additional quantifications for the stability comparison between AIS and SMM, where 10 repetitions are conducted for each pair of initialization and random instance ($M = 10$). Standard deviation of the final objectives and the norm of final solutions among the 10 repetitions are averaged over all the instances and initializations, and reported as "Std. Obj." and "Std. Sol." respectively. Based on the table, AIS can effectively control the variance from the intermediate sampling hence the variability in the solutions computed, which is favorable in practice. For instance, when $\bar{N} = 10$, AIS achieves one-fifth (resp. one-tenth) standard deviation in objective (resp. norm of solution) compared to the non-incremetal SMM.

The advantage of AIS over SAA is also significant from Table 2. Overall, AIS outperforms SAA in terms of efficiency and final objective value. The performance of SAA primarily relies on the sample size S, meaning that an inaccurate SAA from a small sample size like $S = 100$ will lead to an unreliable solution despite being easier to compute. As S grows larger, the quality of SAA solutions approaches those obtained by AIS, as indicated by the objective value at termination. Interestingly, even when $\bar{N} = 10$, AIS can still achieve a comparable objective as SAA under $S = 10000$, but with a computing speed that is at least five times faster.

6.2 Applications in BHM inference

In this part, AIS is applied to the following approximation of MRF with unknown edges (12)

$$\begin{aligned}
&\text{level 2} \quad \widetilde{y}_i \,|\, \widetilde{\theta} && \overset{\text{ind.}}{\sim} \quad p_i^y(y_i \,|\, \widetilde{\theta}), \quad \text{for all } i \in \mathcal{V}, \\
&\text{level 1} \quad \widetilde{\theta} \,|\, \widetilde{u} && \sim \quad \frac{1}{Z_{\text{edge}}(\widetilde{u})} \exp\left\{ -\sum_{i<j} \varphi_{ij}(\widetilde{u}_{ij}) h_{ij}(\theta_i, \theta_j) \right\}, \qquad (43) \\
&\text{level 0} \quad \widetilde{u}_{ij} && \overset{\text{iid}}{\sim} \quad \text{Uniform}([0,1]), \quad \text{for all } i < j.
\end{aligned}$$

where $Z_{\text{edge}}(\widetilde{u}) \triangleq \int_{\Theta} \exp\left\{-\sum_{i<j} \varphi_{ij}(\widetilde{u}_{ij}) h_{ij}(z_i, z_j)\right\} dz$. This model is attained by first applying the uniform transformation as in (5), i.e., substituting $\widetilde{\gamma}_{ij}$ in (12) by $F^{-1}_{\widetilde{\gamma}_{ij}}(\widetilde{u}_{ij})$ with $\widetilde{u}_{ij}$ being uniformly distributed on $[0, 1]$ and $F^{-1}_{\widetilde{\gamma}_{ij}}(s) = \mathbf{1}_{(0,\infty)}(s - p_{ij})$ being the generalized inverse of Bernoulli cdf. Then we employ the treatment in Section 3 to approximate $F^{-1}_{\widetilde{\gamma}_{ij}}$ by a nonconvex piecewise affine φ_{ij}, derived from φ_{ub} in (17), whose formulation is listed in Appendix C.2. Additionally, $p^y_i(y^i \mid \widetilde{\theta})$ is univariate Gaussian with mean $\widetilde{\theta}_i$ and fixed variance $\sigma_i^2 > 0$ for all $i \in \mathcal{V}$.

The MAP associated with (43) is typically nonconvex and nondifferentiable with intractable integral term Z_{edge}, making it computationally challenging. For a simplification, we can resort to the naïve benchmark modeling (10), by setting $\widetilde{u}_{ij}$ in (43) to be deterministic so that the intractable Z_{edge} will vanish. For example, if we let $\widetilde{u}_{ij}$ to be some known $\bar{u}_{ij} \in [0, 1]$ in (43) when $h_{ij}(\theta_i, \theta_j) = A_{ij}(\theta_i - \theta_j)^2$ with fixed $A_{ij} > 0$, then the simplified MAP becomes

$$\underset{\theta \in \Theta}{\textbf{minimize}} \quad \sum_{i=1}^{M} \frac{1}{2\sigma_i^2}(y_i - \theta_i)^2 + \sum_{i<j} \varphi_{ij}(\bar{u}_{ij}) A_{ij}(\theta_i - \theta_j)^2 \tag{44}$$

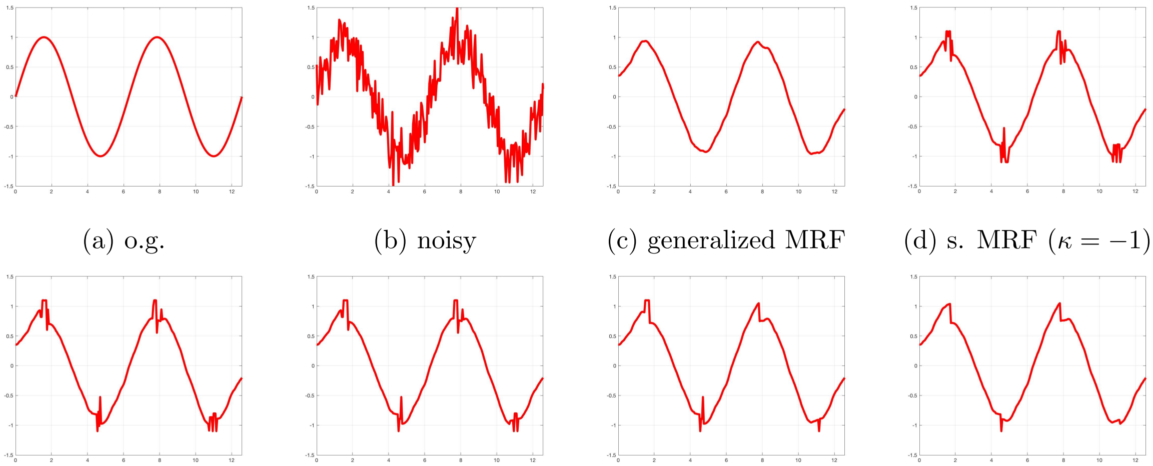
whose objective is quadratic with a special Stieltjes structure, making its global optimal solution computable in strongly polynomial time when Θ is defined by box constraint; [37]. For more advanced discussion when θ is additionally sparse in [22, 24]. However, as discussed in Section 2.2, such simplifications are not always reliable, as the prescribed $\bar{u}_{ij}$ might be misspecified if naïvely estimated from noisy data y. In this sense, model (43) can benefit us with its flexibility in allowing both network topology $\widetilde{u}_{ij}$ and ground truth $\widetilde{\theta}$ to be recovered simultaneously.

To emphasize such advantages, comparisons between the more generalized model (43) and its simplification are made within the context of signal and image recovery tasks. Through the following experiments, we highlight the value of AIS as an effective method to enable the utility of model (43) which summarizes the underlying data generating process more faithfully.

Smooth signal recovery

Suppose we discretize function $\sin(t)$ for $t \in [0, 4\pi]$ by sampling it at equidistant points with a spacing of 0.05 (except for $t = 4\pi$), resulting in $M \triangleq |\mathcal{V}| = 253$. We define $\Theta = [-1, 1]^M$ and construct noisy signal y by adding iid Gaussian noises with mean 0 and variance σ^2 to each point sampled, where σ^2 is 0.09 (resp. 0.25) to represent small (resp. large) noises. Model (43) is applied with $h_{ij}(\theta_i, \theta_j) = A_{ij}(\theta_i - \theta_j)^2$ where $A_{ij} = 50$. Intuitively, the ideal value $\widetilde{u}^*$ in (43) should reflect the smoothness of the underlying signal, namely $\varphi_{ij}(\widetilde{u}^*_{ij}) = 1$ if $i + 1 = j$ and $\varphi_{ij}(\widetilde{u}^*_{ij}) = 0$ otherwise. To simplify our computation, we only estimate the similarities around the four peaks of sine signal between 0 and 4π. More specifically, with edge set $\mathcal{N} \triangleq \{(k\pi + l - 1,\, k\pi + l) : k = 1...4,\, l = -5...5\}$, we fix $\widetilde{u}_{ij}$ in (43) to be the ideal values $\widetilde{u}^*_{ij}$ given that $(i, j) \notin \mathcal{N}$. When our AIS method is applied to the MAP of (43) whose formulation can be found as (49) in Appendix C.2, subproblem (26) is formulated as (50) and solved by MOSEK [2]. Finally, AIS is terminated when the relative difference of solutions (in ℓ_2 norm) between the two most recent steps is less than 5×10^{-5}.

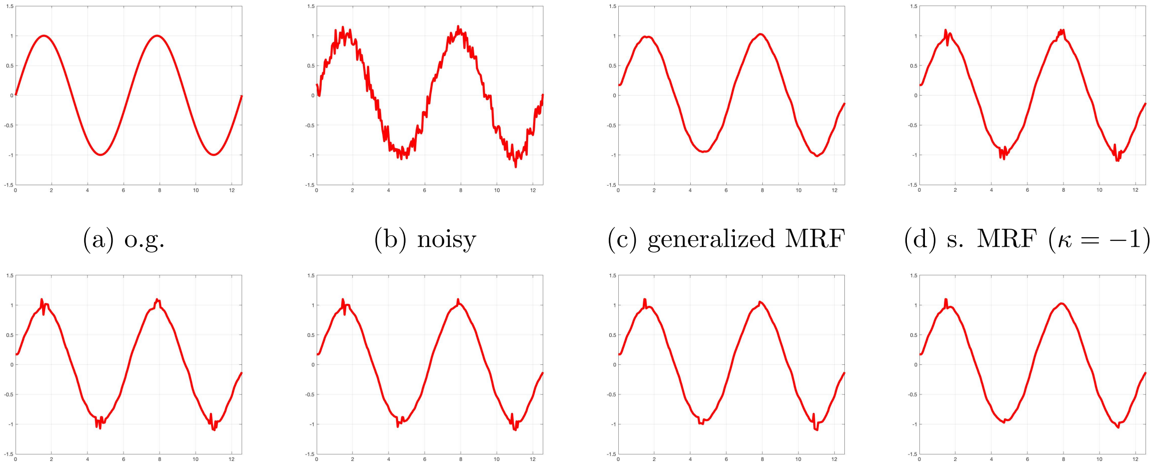
To facilitate comparison, we simplify model (43) for $(i, j) \in \mathcal{N}$ by setting $\widetilde{u}_{ij}$ to be $\bar{u}_{ij} = 1$ heuristically if $|y_i - y_j| < \bar{y}$ and $\widetilde{u}_{ij}$ to be $\bar{u}_{ij} = 0$ otherwise, where the threshold is $\bar{y} \triangleq \frac{1}{m-1}\sum_{i=1}^{m-1} |y_{i+1} - y_i| + \kappa\nu$ with $\kappa \in \{-1, -0.5, 0, 0.5, 1\}$ and ν being the standard deviation among $\{|y_{i+1} - y_i| : i = 1...M - 1\}$. The resulting simplified MAP is (44) and solved by method introduced in [37].



(a) o.g. (b) noisy (c) generalized MRF (d) s. MRF ($\kappa = -1$)

(e) s. MRF ($\kappa = -0.5$) (f) s. MRF ($\kappa = 0$) (g) s. MRF ($\kappa = 0.5$) (h) s. MRF ($\kappa = 1$)

Figure 3: Comparison of MRF models for signal recovery ($\sigma^2 = 0.25$)



(a) o.g. (b) noisy (c) generalized MRF (d) s. MRF ($\kappa = -1$)

(e) s. MRF ($\kappa = -0.5$) (f) s. MRF ($\kappa = 0$) (g) s. MRF ($\kappa = 0.5$) (h) s. MRF ($\kappa = 1$)

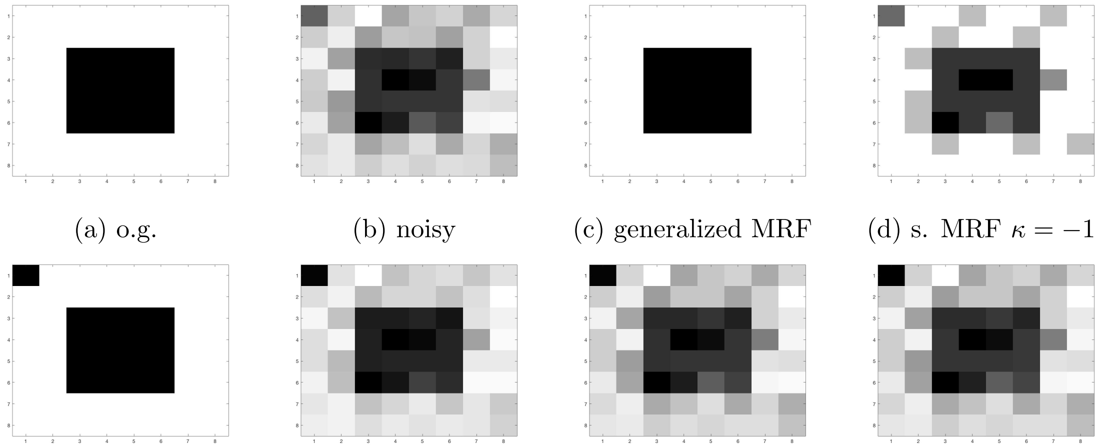
Figure 4: Comparison of MRF models for signal recovery ($\sigma^2 = 0.09$)

Figures 3 and 4 depict the signals recovered from model (43) (referred to as "generalized MRF") and its simplifications (referred to as "s. MRF") at various κ for threshold $\bar{y}$. For reference, we also plot the noisy signal (labelled as "noisy") and the ground truth (labelled as "o.g.").

As demonstrated by the figures, the signals computed using model (43) are overall smooth around the peaks and closely resemble the ground truth signal. In other words, the solutions obtained by AIS are capable of capturing the ideal similarity structure $\widetilde{u}^*$. Conversely, the signals recovered from the simplified model exhibit some undesired spikes around the peaks, even when the noise is relatively small ($\sigma^2 = 0.09$), see for instance Figure 4d. This indicates that while the simplified model benefits from faster inference, the quality of its solution may be compromised by an unreliable specification of hyperparameters $\bar{u}_{ij}$ from noisy data.

Image recovery

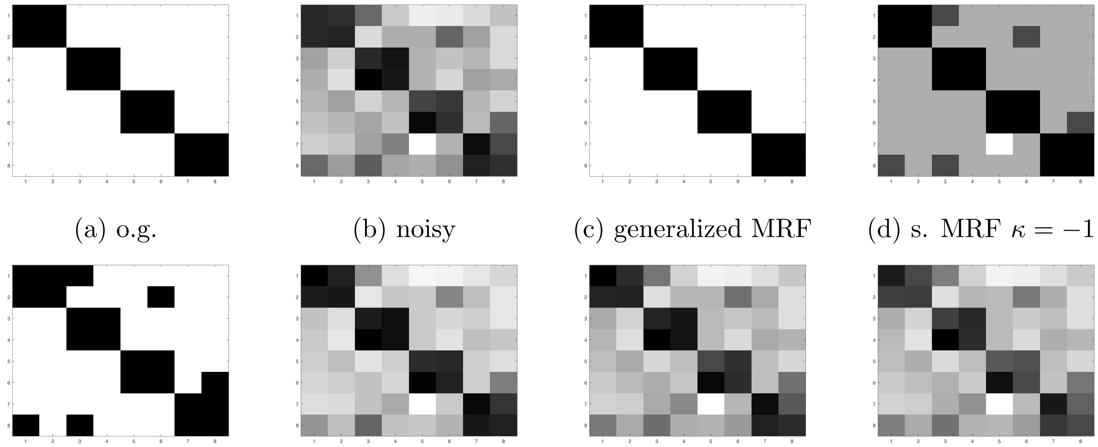
AIS is also tested on the two image recovery tasks. In case 1, the original image, denoted as "o.g." in Figure 5a, is an 8-by-8 black and white image ($M = 64$) with the 16 pixels in the middle taking value 25.5 while the background pixels being 153. We set $\Theta = [0, 255]^M$ and contaminate each pixel with iid Gaussian noise with zero mean and variance $\sigma^2 = 25$. The resulting observation y is further trimmed to take value in $[0, 255]$, see Figure 5b. The observation for case 2 is similarly generated except for the original image being block diagonal as shown in Figure 6a.



(a) o.g. (b) noisy (c) generalized MRF (d) s. MRF $\kappa = -1$

(e) s. MRF $\kappa = -0.5$ (f) s. MRF $\kappa = 0$ (g) s. MRF $\kappa = 0.5$ (h) s. MRF $\kappa = 1$

Figure 5: Comparison of MRF models for image recovery (case 1)



(a) o.g. (b) noisy (c) generalized MRF (d) s. MRF $\kappa = -1$

(e) s. MRF $\kappa = -0.5$ (f) s. MRF $\kappa = 0$ (g) s. MRF $\kappa = 0.5$ (h) s. MRF $\kappa = 1$

Figure 6: Comparison of MRF models for image recovery (case 2)

To recover the original image as well as the pairwise similarities between pixels, we consider (43) with $h_{ij}(\theta_i, \theta_j) = A_{ij}|\theta_i - \theta_j|$ and $A_{ij} = 50$. The associated MAP and its corresponding AIS subproblem (26) are similar to (49) and (50) in Appendix C.2 respectively, thus are omitted here. AIS is terminated when we reach a 5×10^{-4} relative difference in the norm of solutions between the most recent two steps. The image we recovered from (43) is shown in Figure 5c. Similar to signal recovery, we also implement the simplification of (43) by fixing $\widetilde{u}_{ij}$ to be $\bar{u}_{ij} = 1$ if $|y_i - y_j| < \bar{y}$ and $\widetilde{u}_{ij}$ to be $\bar{u}_{ij} = 0$ otherwise, where the threshold is $\bar{y} \triangleq \dfrac{2}{m(m-1)} \displaystyle\sum_{i<j} |y_i - y_j| + \kappa\nu$ with $\kappa \in \{-1, -0.5, 0, 0.5, 1\}$ and ν is the standard deviation among $\{|y_i - y_j| : i < j\}$. The figures showing the images recovered by solving the simplified MAPs under different κ are presented in Figure 5d to 5h and Figure 6d to 6h, labelled as "s. MRF".

The solutions obtained by AIS from model (43), as shown in Figure 5 and 6, are able to accurately capture the dark pixels in the original image. This indicates that the model has effectively identified the relevant similarities between pixels that distinguish the dark blocks from the background. It is worth noting that such similarity structure is challenging to specify a priori, especially under the added noise. Consequently, the simplified MRF approaches considered with various κ values are unable to recover the original image, as they lack the generality of model (43).

Acknowledgment. The authors are grateful to Professor Enlu Zhou of Georgia Institute of Technology who suggested that the problem (4) is an instance of a Bayesian hierarchical model.

References

[1] G. Alain, A. Lamb, C. Sankar, A. Courville and Y. Bengio. Variance reduction in sgd by distributed importance sampling. *arXiv preprint arXiv:1511.06481* (2015).

[2] M. ApS. The MOSEK optimization toolbox for MATLAB manual. Version 9.3.21. "https://docs.mosek.com/latest/toolbox/index.html" (2022).

[3] J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)* 36(2): 192-225 (1974).

[4] J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society: Series D (The Statistician)* 24(3): 179-195 (1975).

[5] J. Besag and C. Kooperberg. On conditional and intrinsic autoregressions. *Biometrika* 82(4): 733-746 (1995).

[6] J. Besag, J. York and A. Mollié Bayesian image restoration, with two applications in spatial statistics. *Annals of the institute of statistical mathematics* 43: 1-20 (1991).

[7] O. Cappeé, A. Guillin, J.M. Martin and C.P. Robert. Population monte carlo. *Journal of Computational and Graphical Statistics* 13(4): 907-929 (2004).

[8] D. Csiba, Z. Qu and P. Richtárik. Stochastic dual coordinate ascent with adaptive probabilities. In *International Conference on Machine Learning* (pp. 674-683). PMLR (June 2015).

[9] Y. Cui, J. Liu, and J.S. Pang. Nonconvex and nonsmooth approaches for affine chance constrained stochastic programs. Preprint. *Journal of Set-valued and Variational Analysis* 30: 1149–1211 (2022).

[10] Y. Cui, J. Liu, and J.S. Pang. The minimization of piecewise functions: Pseudo stationarity. *Journal of Convex Analysis* (accepted August 2022).

[11] Y. Cui and J.S. Pang. *Modern Nonconvex Nondifferentiable Optimization*. MOS–SIAM Series on Optimization, SIAM Publications (December 2021).

[12] G.B. Dantzig and G. Infanger. Large-scale stochastic linear programs: Importance sampling and Benders decomposition. Technical report (1991).

[13] D. Drusvyatskiy and L. Xiao. Stochastic optimization with decision-dependent distribution. *Mathematics of Operations Research* 48(2): 954–998 (2023).

[14] R. Durrett *Probability: Theory and Examples*. Version 5 (January 2019).

[15] Y.M. Ermoliev AND V.I. Norkin. Sample average approximation method for compound stochastic optimization problems. *SIAM Journal on Optimization* 23(4): 2231–2263 (2013).

[16] G. Folland. *Real Analysis: Modern Techniques and Their Applications*. Volume 40. John Wiley & Sons (1999).

[17] A. Gelman and X. Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science* 13(2): 163–185 (1998).

[18] A. Gelfand and B. Carlin. Maximum likelihood estimation for constrained or missing data models. *Canadian Journal of Statistics* 21(3): 303-311 (1993).

[19] S. Geman and C. Graffigne. Markov Random Field image models and their applications to computer vision. In *Proceedings of the international congress of mathematicians* (Vol. 1, p. 2) (1986).

[20] C. Geyer and E. Thompson. Constrained Monte Carlo maximum likelihood for dependent data. *Journal of the Royal Statistical Society: Series B (Methodological)* 54(3): 657-683 (1992).

[21] C. Geyer. On the convergence of Monte Carlo maximum likelihood calculations. *Journal of the Royal Statistical Society: Series B (Methodological)* 56(1): 261-274 (1994).

[22] A. Gómez, Z. He and J.S. Pang. Linear-step solvability of some folded concave and singly-parametric sparse optimization problems. *Mathematical Programming* 198: 1339–1380 (2023).

[23] S. Goodreau, J. Kitts and M. Morris. Birds of a feather, or friend of a friend? Using exponential random graph models to investigate adolescent social networks. *Demography* 46(1): 103-125 (2009).

[24] Z. He, S. Han, A. Gómez, Y. Cui and J.S. Pang. Comparing solution paths of sparse quadratic minimization with a Stieltjes matrix. *Mathematical Programming* (2023). https://doi.org/10.1007/s10107-023-01966-0

[25] M. Ibáñez and A. Simó. Parameter estimation in Markov Random Field image modeling with imperfect observations. A comparative study. *Pattern recognition letters* 24(14): 2377-2389 (2003).

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

[27] T.B. Johnson and C. Guestrin. Training deep models faster with robust, approximate importance sampling. *Advances in Neural Information Processing Systems*, 31 (2018).

[28] A.B. Lawson. *Bayesian disease mapping: hierarchical modeling in spatial epidemiology*. Chapman and Hall/CRC (2018).

[29] W. Lee, H. Yu and H. Yang Reparameterization gradient for non-differentiable models. *Advances in Neural Information Processing Systems*, 31 (2018).

[30] J. Liu, Y. Cui, and J.S. Pang. Solving nonsmooth nonconvex compound stochastic programs with applications to risk measure minimization. *Mathematics of Operations Research* 47(4): 3051–3083 (2022).

[31] J. Liu, Y. Cui, J.S. Pang, and S. Sen. Two-stage stochastic programming with linearly biparameterized quadratic recourse. *SIAM Journal on Optimization* 30(3): 2530–2558 (2020).

[32] J. Liu and J.S. Pang. Risk-based robust statistical learning by stochastic difference-of-convex value-function optimization. *Operations Research* (accepted February 2022). https://doi.org/10.1287/opre.2021.2248.

[33] Y.C. MacNab. Bayesian disease mapping: Past, present, and future. *Spatial Statistics* 50: 100593 (2022).

[34] J. Møller, A. Pettitt, R. Reeves and K. Berthelsen. An efficient Markov chain Monte Carlo method for distributions with intractable normalising constants. *Biometrika* 93(2): 451-458 (2006).

[35] I. Murray, Z. Ghahramani and D. MacKay. MCMC for doubly-intractable distributions. *arXiv preprint arXiv:1206.6848* (2012).

[36] M.S. Oh and J.O. Berger. Adaptive importance sampling in Monte Carlo integration. *Journal of statistical computation and simulation* 41(3-4): 143-168 (1992).

[37] J.S. Pang and R. Chandrasekaran. Linear complementarity problems solvable by a polynomially bounded pivoting algorithm. *Mathematical Programming Essays in Honor of George B. Dantzig Part II*: 13–27 (1985).

[38] P. Parpas, B. Ustun, M. Webster and Q.K. Tran. Importance sampling in stochastic programming: A Markov chain Monte Carlo approach. *INFORMS Journal on Computing* 27(2): 358-377 (2015).

[39] J. Prithcard, M. Seielstad, A. Perez-Lezaun and M. Feldman. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. *Molecular biology and evolution* 16(12): 1791-1798 (1999).

[40] H. Robbins and D. Siegmund. A convergence theorem for non-negative almost supermartingales and some applications. *Optimizing Methods in Statistics*: 233–257 (1971).

[41] C.P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Second Editor. Springer New York (2004).

[42] R. Salakhutdinov and H. Larochelle. Efficient learning of deep Boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 693-700). JMLR Workshop and Conference Proceedings (2010).

[43] A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory*. MOS-SIAM Series on Optimization, Volume 9. SIAM, Philadelphia (2009).

[44] S.U. Stich, A. Raj and M. Jaggi. Safe adaptive importance sampling. *Advances in Neural Information Processing Systems*, 30 (2017).

[45] L. Younes. Estimation and annealing for Gibbsian fields. In *Annales de l'IHP Probabilités et statistiques* 24(2): 269-294 (1988).

A Proof of Some Intermediate Results

In what follows, we complete the proof of two results that are useful for our analysis, namely Proposition 9 and Lemma 11.

A.1 Proof of Proposition 9

Denote $\phi(\chi, \bar{x}, z) \triangleq \dfrac{\exp \widehat{H}(\chi, z; \bar{x})}{\exp H(\bar{x}, z)} Z(\bar{x})$, then by Proposition 8 and our assumptions, we have

$$\lim_{t\to\infty} \sup_{(\chi,\bar{x})\in X\times X} \left| \frac{1}{N_t} \sum_{s=1}^{N_t} \phi(\chi, \bar{x}, \zeta^{st}) - \mathbb{E}_{\zeta\sim\pi^*(x^t)} \phi(\chi, \bar{x}, \zeta) \right| = 0.$$

Note that for an arbitrary $x \in X$

$$\begin{aligned} \left| \widehat{Z}^{N_t}_{\pi^*(x^t)}(x; x^t) - \widehat{Z}(x; x^t) \right| &= \left| \frac{1}{N_t} \sum_{s=1}^{N_t} \phi(x, x^t, \zeta^{st}) - \mathbb{E}_{\zeta \sim \pi^*(x^t)} \phi(x, x^t, \zeta) \right| \\ &\leq \sup_{(\chi, \bar{x}) \in X \times X} \left| \frac{1}{N_t} \sum_{s=1}^{N_t} \phi(\chi, \bar{x}, \zeta^{st}) - \mathbb{E}_{\zeta \sim \pi^*(x^t)} \phi(\chi, \bar{x}, \zeta) \right|. \end{aligned}$$

Thus we can obtain

$$\lim_{t \to \infty} \left| \widehat{Z}^{N_t}_{\pi^*(x^t)}(x; x^t) - \widehat{Z}(x; x^t) \right| = 0,$$

which implies that $\limsup_{t \to \infty} \widehat{Z}^{N_t}_{\pi^*(x^t)}(x; x^t) - \widehat{Z}(x; x^t) \leq 0$. Then for any convergent subsequence $\{x^t\}_{t \in \mathcal{T}}$ with the limit point $x^\infty \in X$, we have

$$\limsup_{t(\in \mathcal{T}) \to \infty} \widehat{Z}^{N_t}_{\pi^*(x^t)}(x; x^t) \leq \limsup_{t(\in \mathcal{T}) \to \infty} \widehat{Z}(x; x^t) \leq \widehat{Z}(x; x^\infty)$$

where the last inequality is obtained by the upper semicontinuity of $\widehat{H}(x; \bullet)$ and Fatou's Lemma. □

A.2 Proof of Lemma 11

By our assumptions, we can without loss of generality treat logarithm as Lipschitz continuous with constant $\widetilde{L}$, and there exists $M_H > 0$ such that $\sup_{(x,z) \in X \times \Xi} \dfrac{\exp(H(x,z))}{\pi^*(z, \bar{x})} < M_H$ for all $\bar{x} \in X$ and there exists $L_H > 0$ such that $\dfrac{\exp(H(x,z))}{\pi^*(z, \bar{x})}$ is Lipschitz continuous in $x \in X$ with constant L_H for all $(\bar{x}, z) \in X \times \Xi$. By an application of Theorem 3.5 in [15], these conditions are enough to ensure (35) holds with $C_\alpha \triangleq \widetilde{L}\sqrt{n} \left(L_H D + \dfrac{M_H}{\sqrt{(1-2\alpha)e}} \right)$ where D is the diameter of compact X. □

B On an Alternative AIS Scheme with Better Sample Complexity

As we discussed in Section 5.5, we can obtain an improved N_t if we make a slight adjustment to our AIS-based surrogation method and require $\rho > 0$ to be small enough. More specifically, if $\|x - x^t\|_2$ is used for the regularizer instead of $\|x - x^t\|_2^2$ as in the original algorithm, then we can achieve the following result that is similar to Theorem 12 but only requires N_t to increases linearly in t.

Proposition 14. Let the followings be given:

- The regularization term in the AIS-based surrogation method is changed to $\dfrac{1}{2\rho}\|x - x^t\|_2$;
- A sequence of positive integers $\{N_t\}_{t \geq 0}$ satisfying for some $\kappa > 0$ and integer T_κ the condition that $N_t \geq \kappa t$ for all $t \geq T_\kappa$.

Then there exists a $\bar{\rho} > 0$ such that if we fix $\rho < \bar{\rho}$ then with probability one, the sequence $\{x^t\}_{t \geq 0}$ produced by the alternative algorithm will attain an accumulation point $x^\infty \in X$ that satisfies

$$x^\infty \in \mathop{\mathbf{argmin}}_{x \in X} \; \widehat{c}(x; x^\infty) + \log \widehat{Z}(x; x^\infty) + \frac{1}{2\rho}\|x - x^\infty\|_2$$

Proof. First note that if $\|x^{t+1} - x^t\|_2 \to 0$ (almost surely) is given, then the analysis of asymptotic fixed point stationarity, i.e., Section 5.4 after we substitute $\| \bullet \|_2^2$ with $\| \bullet \|_2$, only requires N_t to grow linearly in t. Thus it amounts to show that $\|x^{t+1} - x^t\|_2 \to 0$ almost surely if we additionally restrict $\rho > 0$ to be small enough.

Indeed, note that after some simple rearrangements, (33) gives:

$$\begin{aligned}
\frac{1}{2\rho}\|x^{t+1} - x^t\|_2 \quad &\leq \quad \left|c(x^t) - c(x^{t+1})\right| + 2\left|\log Z(x^t) - \log \bar{Z}^{N_{t-1}}_{\pi^*(x^{t-1})}(x^t)\right| \\
&\qquad + \left|\log Z(x^t) - \log Z(x^{t+1})\right| + \left|\log Z(x^{t+1}) - \log \bar{Z}^{N_t}_{\pi^*(x^t)}(x^{t+1})\right| \\
&\leq \quad \left|c(x^t) - c(x^{t+1})\right| + \widetilde{L}\left(\begin{array}{l} \left|Z(x^{t+1} - Z(x^t)\right| + 2\left|Z(x^t) - \bar{Z}^{N_{t-1}}_{\pi^*(x^{t-1})}(x^t)\right| \\ + \left|Z(x^{t+1}) - \bar{Z}^{N_t}_{\pi^*(x^t)}(x^{t+1})\right| \end{array}\right)
\end{aligned} \tag{45}$$

where the second inequality is by our assumptions so that logarithm can be treated as Lipschitz continuous with constant $\widetilde{L}$. Furthermore, by the identity that

$$Z(x^t) \quad = \quad \frac{1}{N_{t-1}} \sum_{s=1}^{N_{t-1}} \frac{\exp\left(H(x^t, \zeta^{s(t-1)})\right)}{\pi_{\text{IS}}^{H(x^t,\bullet)}(\zeta^{s(t-1)})}$$

there exists constant $L_H > 0$ so that the followings hold:

$$\begin{aligned}
\left|Z(x^t) - \bar{Z}^{N_{t-1}}_{\pi^*(x^{t-1})}(x^t)\right| \quad &\leq \quad \frac{1}{N_{t-1}} \sum_{s=1}^{N_{t-1}} \exp\left(H(x^t, \zeta^{s(t-1)})\right) \left| \begin{array}{l} \exp\left(-H(x^t, \zeta^{s(t-1)})\right) Z(x^t) \\ - \exp\left(-H(x^{t-1}, \zeta^{s(t-1)})\right) Z(x^{t-1}) \end{array} \right| \\
&\leq \quad L_H \|x^t - x^{t-1}\|_2
\end{aligned}$$

where the second inequality is by H, Z being continuous on compact set hence there exists $M_H < \infty$ such that $\exp(H(x, z)) < M_H$ for all $(x, z) \in X \times \Xi$, and $Z(x) \exp\left(-H(x, z)\right)$ can be treated as jointly Lipschitz continuous in x and z. For the same reason we have:

$$\left|Z(x^{t+1}) - \bar{Z}^{N_t}_{\pi^*(x^t)}(x^{t+1})\right| \quad \leq \quad L_H \|x^{t+1} - x^t\|_2$$

Furthermore, by c and Z being continuous on compact X hence Lipschitz continuous with constant L_c and L_Z respectively, we can eventually derive the followings from (45):

$$\frac{1}{2\rho}\|x^{t+1} - x^t\|_2 \quad \leq \quad (L_c + \widetilde{L}L_Z + \widetilde{L}L_H) \left\|x^{t+1} - x^t\right\|_2 + 2\widetilde{L}L_H \left\|x^t - x^{t-1}\right\|_2$$

If we take ρ to be small enough so that $\beta \triangleq \dfrac{2\widetilde{L}L_H}{\frac{1}{2\rho} - (L_c + \widetilde{L}L_Z + \widetilde{L}L_H)} \in (0, 1)$, then

$$\|x^{t+1} - x^t\|_2 \quad \leq \quad \beta \left\|x^t - x^{t-1}\right\|_2$$

for any $t \geq 1$, which means $\|x^{t+1} - x^t\|_2 \to 0$ by X being compact. □

Remark 15. Intuitively, restricting $\rho > 0$ to be small enough in exchange for a better complexity on N_t is necessary if we interpret ρ as the step size for our algorithm. The main idea of AIS-based surrogation is that by making both surrogation and SAA adaptive to our current progress x^t, we intend to obtain a more accurate approximation of function $Z(x)$ in (19) with fewer samples. However by Lemma 2 this is only effective when we are locally around x^t, hence in this sense we cannot afford a step size ρ that is too large if we intend to control N_t. □

C Some Details for Numerical Experiments

Below we provide the formulations for the subproblems in Section 6.1 and the piecwise affine approximation of indicator function which we employed in Section 6.2.

C.1 Subproblems for the experiments in Section 6.1

Given θ^t, γ^t from the most recent iteration, we can construct the following subproblem

$$\begin{aligned}
\underset{\theta,\gamma}{\textbf{minimize}} \quad & \sum_{i=1}^{M} \frac{1}{2\sigma_i^2}(y_i-\theta_i)^2 + \sum_{i<j} \left(\begin{array}{l} \dfrac{1}{2}\left(\gamma_{ij} + |\theta_i - \theta_j|\right)^2 \\ - \left(\begin{array}{l} \dfrac{1}{2}(\gamma_{ij}^t)^2 + \gamma_{ij}^t(\gamma_{ij}-\gamma_{ij}^t) + \dfrac{1}{2}(\theta_i^t - \theta_j^t)^2 \\ + (\theta_i^t - \theta_j^t)\left(\theta_i - \theta_i^t - (\theta_j - \theta_j^t)\right) \end{array} \right) \end{array} \right) \\
& + \log\left(Z^t(\gamma)\right) + \frac{1}{2\rho}\left(\|\gamma - \gamma^t\|_2^2 + \|\theta - \theta^t\|_2^2\right) \\
\text{subject to} \quad & \theta \in \Theta, \quad \underline{\gamma}_{ij} \le \gamma_{ij} \le \overline{\gamma}_{ij}, \; \forall i < j.
\end{aligned} \tag{46}$$

where $Z^t(\gamma)$ is specified as follows for AIS-based surrogation and SMM methods:

$$\text{AIS-based surrogation} \quad \left\{ \begin{array}{l} \text{iid } \{z^{st}\}_{s=1}^{N_t} \text{ drawn from density } \dfrac{1}{Z(\gamma^t)} \exp\left\{ -\displaystyle\sum_{i<j} \gamma_{ij}^t |z_i - z_j| \right\} \\ Z^t(\gamma) \; = \displaystyle\sum_{s=1}^{N_t} \exp\left\{ -\sum_{i<j} (\gamma_{ij} - \gamma_{ij}^t) \left| z_i^{st} - z_j^{st} \right| \right\} \end{array} \right.$$

$$\text{non-incremental SMM} \quad \left\{ \begin{array}{l} \text{iid } \{z^{st}\}_{s=1}^{N_t} \text{ drawn from uniform distribution over } \Theta \\ Z^t(\gamma) \; = \displaystyle\sum_{s=1}^{N_t} \exp\left\{ -\sum_{i<j} \gamma_{ij} \left| z_i^{st} - z_j^{st} \right| \right\} \end{array} \right.$$

$$\text{incremental SMM} \quad \left\{ \begin{array}{l} \text{iid } \{z^{st}\}_{s=1}^{N_t} \text{ drawn from uniform distribution over } \Theta \\ Z^t(\gamma) \; = \displaystyle\sum_{\tau=1}^{t} \sum_{s=1}^{N_\tau} \exp\left\{ -\sum_{i<j} \gamma_{ij} \left| z_i^{s\tau} - z_j^{s\tau} \right| \right\} \end{array} \right.$$

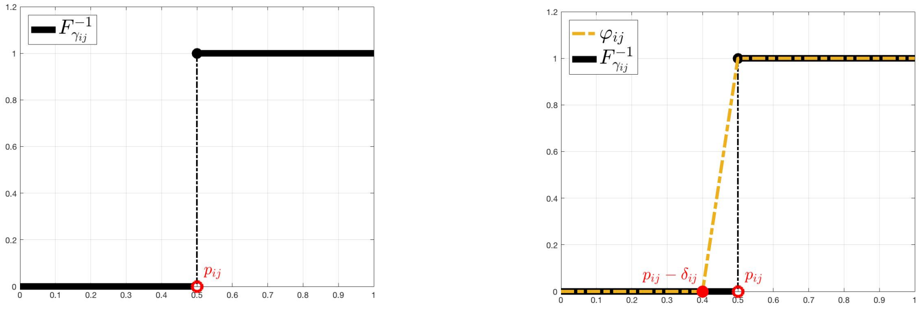
Note that by Lemma 2 problem (46) is convex and handled by MOSEK [2] in our experiments.

C.2 Piecewise affine approximation of indicator function

Suppose $\widetilde{\gamma}_{ij}$ is Bernoulli with parameter $p_{ij} > 0$, then the generalized inverse of its cumulative distribution function is $F_{\widetilde{\gamma}_{ij}}^{-1}(s) = \mathbf{1}_{(0,\infty)}(s - p_{ij})$ for $s \in [0,1]$. We can apply the $\widehat{\varphi}_{\text{ub}}$ treatment in (17) with $\widetilde{\varphi}_{\text{cvx}}(s) = s$ and result in the following nonconvex piecewise affine approximation of $F_{\widetilde{\gamma}_{ij}}^{-1}$

$$\varphi_{ij}(s) \;\triangleq\; \min\left(\max\left(1 + \frac{s - p_{ij}}{\delta_{ij}}, 0 \right), 1 \right) \tag{47}$$

where $\delta_{ij} > 0$ is a fixed hyperparameter controlling the approximation and as $\delta_{ij} \downarrow 0$ we have $\varphi_{ij} \to F^{-1}_{\widetilde{\gamma}_{ij}}$ pointwise. Visualization of such approximation can be found in Figure 7



(a) indicator function $F^{-1}_{\widetilde{\gamma}_{ij}}$

(b) piecewise affine φ_{ij}

Figure 7: Piecewise affine φ_{ij} approximation of indicator $F^{-1}_{\widetilde{\gamma}_{ij}}$

Note that φ_{ij} in (47) can be reformulated as the following difference-of-convex (DC) function

$$\varphi_{ij}(s) \quad = \quad \underbrace{\max\left(\alpha_{ij}s + \beta_{ij},\, 0\right)}_{\triangleq\, \varphi^{+}_{ij}(s)} - \underbrace{\max\left(\alpha_{ij}s + \beta_{ij} - 1,\, 0\right)}_{\triangleq\, \varphi^{-}_{ij}(s)} \tag{48}$$

where $\alpha_{ij} \triangleq \dfrac{1}{\delta_{ij}}$ and $\beta_{ij} \triangleq 1 - \dfrac{p_{ij}}{\delta_{ij}}$. The MAP associated with model (43) in Section 6.2 is

$$\begin{aligned} \underset{\theta,\, u}{\textbf{minimize}} \quad & \sum_{i=1}^{M} \frac{1}{2\sigma_i^2}(y_i - \theta_i)^2 + \sum_{i<j} \varphi_{ij}(u_{ij}) A_{ij}(\theta_i - \theta_j)^2 + \log\left(Z_{\text{edge}}(u)\right) \\ \text{subject to} \quad & \theta \in \Theta, \quad u_{ij} \in [0,1] \text{ for } i < j. \end{aligned} \tag{49}$$

where $h_{ij}(\theta_i, \theta_j)$ is substituted by $A_{ij}(\theta_i - \theta_j)^2$ with fixed $A_{ij} > 0$. In the context of AIS method, given u^t and θ^t from the previous step, we draw conditionally independent samples $\{z^{st}\}_{s=1}^{N_t}$ from density $\dfrac{1}{Z_{\text{edge}}(u^t)} \exp\left\{ -\displaystyle\sum_{i<j} \varphi_{ij}(u^t_{ij}) A_{ij}(z_i - z_j)^2 \right\}$ and construct the following formulation for AIS subproblem (26)

$$\begin{aligned} \underset{\theta,\, u}{\textbf{minimize}} \quad & \sum_{i=1}^{M} \frac{1}{2\sigma_i^2}(y_i - \theta_i)^2 + \sum_{i<j} \widehat{G}^t_{ij}(u,\theta) + \frac{1}{2\rho}\left(\|\theta - \theta^t\|_2^2 + \|u - u^t\|_2^2\right) \\ & + \log\left(\frac{1}{N_t} \sum_{s=1}^{N_t} \exp\left\{ -\sum_{i<j} A_{ij}\left(\widehat{\varphi}^t_{ij}\left(u_{ij}\right) - \varphi_{ij}(u^t_{ij})\right)\left(z_i^{st} - z_j^{st}\right)^2 \right\} \right) \\ \text{subject to} \quad & \theta \in \Theta, \quad u_{ij} \in [0,1] \text{ for } i < j. \end{aligned} \tag{50}$$

where

$$
\begin{aligned}
\widehat{G}_{ij}^{\,t}(u,\theta) \quad &\triangleq \quad \begin{cases} A_{ij}(\theta_i-\theta_j)^2 & \text{if } \alpha_{ij}u_{ij}^t+\beta_{ij}\geq 1 \\[2ex] \begin{array}{l} \dfrac{1}{2}\left(\varphi_{ij}^{+}(u_{ij})+A_{ij}(\theta_i-\theta_j)^2\right)^2 - \xi_{ij}^{\,t}(u_{ij}-u_{ij}^t) \\ -\,2A_{ij}^2(\theta_i^t-\theta_j^t)^3\left((\theta_i-\theta_i^t)-(\theta_j-\theta_j^t)\right) \\ -\,\dfrac{1}{2}A_{ij}^2(\theta_i^t-\theta_j^t)^4 - \dfrac{1}{2}\left(\varphi_{ij}^{+}(u_{ij}^t)\right)^2 \end{array} & \text{otherwise} \end{cases} \\[3ex]
\xi_{ij}^{\,t} \quad &\triangleq \quad \max\left\{\alpha_{ij}\left(\alpha_{ij}u_{ij}^t+\beta_{ij}\right),\, 0\right\} \\[3ex]
\widehat{\varphi}_{ij}^{t}(u_{ij}) \quad &\triangleq \quad \begin{cases} \min\left\{\alpha_{ij}u_{ij}+\beta_{ij},\, 1\right\} & \text{if } \alpha_{ij}u_{ij}^t+\beta_{ij}\geq 0 \\ 0 & \text{otherwise} \end{cases}
\end{aligned}
$$