A Proximal Newton Adaptive Importance Sampler

Víctor Elvira, *Senior, IEEE*, Émilie Chouzenoux *Senior, IEEE*, O. Deniz Akyildiz

***Abstract*—Adaptive importance sampling (AIS) algorithms are a rising methodology in signal processing, statistics, and machine learning. An effective adaptation of the proposals is key for the success of AIS. Recent works have shown that gradient information about the involved target density can greatly boost performance, but its applicability is restricted to differentiable targets. In this paper, we propose a proximal Newton adaptive importance sampler for the estimation of expectations with respect to non-smooth target distributions. We implement a scaled Newton proximal gradient method to adapt the proposal distributions, enabling efficient and optimized moves even when the target distribution lacks differentiability. We show the good performance of the algorithm in two scenarios: one with convex constraints and another with non-smooth sparse priors.**

***Index Terms*—Adaptive importance sampling, proximal methods, Newton algorithm, preconditioning.**

I. Introduction

Statistical signal processing applications often require Monte Carlo methods to approximate distributions and integrals, e.g., inference in state-space models [19], rare event estimation [43], or in more general Bayesian inference tasks [10]. Importance sampling (IS) is a well-known technique to tackle such problems [28], particularly when direct Monte Carlo from the target distribution is unfeasible (e.g. in Bayesian inference) or when it is inefficient. The performance of IS depends highly on the proposal quality [3], [2], which has to be chosen appropriately w.r.t. the target density. One popular way to automatize IS is to adapt one or several proposals over an iterative process. In particular, this class of algorithms is called *adaptive importance sampling* (AIS), e.g., see [9] for a review. AIS methods are importance samplers with time-changing (improving) proposal distributions. More precisely, AIS methods specify a sequence of proposal densities, which are adapted over time, and perform importance sampling to estimate expectations with respect to a targeted density. These methods are widely used in Bayesian inference, where the target density is the posterior distribution, making AIS particularly suited for Bayesian signal processing and machine learning [45], [41].

The biggest challenge of AIS methods is to design proposals which move probability mass to the regions of interest under the target density. For certain scenarios, proposing samples which have high probability under target is of certain interest. Since proposal design includes a certain flexibility, using information from the target density, has become a popular route to take. In particular, one can use gradient or Hessian information [35], [52], [25], [27]. Stochastic gradient approximations have also been used, computed with weighted samples [23], [21] or via reinforcement learning [22]. When applied to Bayesian inference, a main difficulty arises, however, in models where the log-likelihood or the log-prior is nonsmooth (not differentiable), e.g., in models where the prior promotes sparsity, e.g., in sparse linear regression [46] or in other Bayesian tasks with non-differentiable priors (see for instance [38]). In particular, many interesting test functions, such as indicator functions of sets, are not differentiable. In these cases, gradient adaptive methods are not effectively applicable. Variable transformation or projection approaches have been used to deal with simple constraints. The former however often yields an artificial deformation of the initial target, often detrimental to sampler efficiency. The latter, initially dedicated to constrained problems, can be naturally extended to any convex non-smooth function, by using its Moreau-Yosida's envelope, through a proximal step [16]. In [53], a framework also leveraging Moreau-Yosida envelopes to approximate non-differentiable target distributions has been recently proposed. This proximal approach enables the use of gradient-based MCMC methods in conjunction with importance sampling for improved efficiency of the estimators. Subgradient updates could also be employed, but these are often suboptimal in terms of convergence speed, difficult to compute for complex targets, making proximal step a more suitable and versatile approach (see discussions in [39], [20], [36], in the context of Langevin MCMC).

In this paper, we propose a class of adaptive importance samplers, called *proximal Newton adaptive importance sampler* (PNAIS), which can use available information of the target efficiently. PNAIS can handle a large class of non-smooth targets, and still efficiently use first and second order information of some terms in the target to propose samples. Our algorithm can be thought of as the importance sampling counterpart of the proximal MCMC method proposed in [18] (itself improving upon [49]), where we integrate several safe rules to cope with non log-concave targets. Without loss of generality, we limit ourselves here to differentiable likelihoods and nonsmooth priors, although the reverse construction would be straightforward to handle. We discuss the algorithmic choices based on theoretical justifications. We provide a numerical analysis in two challenging examples where subset and sparsity constraints turn off-the-shelf methods unfeasible.

The rest of the paper is organized as follows. In Section II, we provide background about convex optimization and adaptive importance sampling. In Section III, we described the proposed algorithm. In Section IV we provide numerical results, and we conclude in Section V.

E.C. is with the team-project OPIS, Inria Saclay, University Paris-Saclay. E.C. acknowledges support from the European Research Council Starting Grant MAJORIS ERC-2019-STG-85092. V. E. is with School of Mathematics, University of Edinburgh. The work of V. E. is supported by ARL/ARO under grant W911NF-22-1-0235. O. D. A is with Department of Mathematics, Imperial College London.

II. Background

A. Problem statement

We aim at estimating the integrals of the following form:

$$\int_{\mathsf{X}} \varphi(\mathbf{x})\widetilde{\pi}(\mathbf{x})\mathrm{d}\mathbf{x}, \tag{1}$$

where $\widetilde{\pi}(\mathbf{x})$ is the target density defined on $\mathsf{X} \subseteq \mathbb{R}^{d_x}$ and $\varphi(\mathbf{x})$ is an integrable function. We focus on targets of the form

$$\widetilde{\pi}(\mathbf{x}) \propto \exp(-f(\mathbf{x}) - g(\mathbf{x})) \equiv \pi(\mathbf{x}), \tag{2}$$

where f is differentiable, and g is not differentiable and convex. A motivation can come from Bayesian statistics where, given some data $\mathbf{y} \in \mathbb{R}^{d_y}$, the target is the posterior

$$\widetilde{\pi}(\mathbf{x}) \propto p(\mathbf{y}|\mathbf{x})p(\mathbf{x}). \tag{3}$$

By comparing Eq. (2) and Eq. (3), we can choose $p(\mathbf{y}|\mathbf{x}) \propto \exp(-f(\mathbf{x}))$, and $p(\mathbf{x}) \propto \exp(-g(\mathbf{x}))$ as a log-concave prior distribution. For instance, $g(\mathbf{x})$ can encode a sparsity inducing prior $g(\mathbf{x}) = \alpha\|\mathbf{x}\|_1$. Our proposed algorithm can tackle more generic problems where $\pi(\mathbf{x})$ is decomposed into a product of smooth and nonsmooth parts.

B. Multiple importance sampling

Multiple importance sampling (MIS) is a Monte Carlo method for approximating distributions and integrals by simulating samples from a set of proposals $\{q_n(\mathbf{x})\}_{n=1}^N$ [51], [44], [47], [34]. A common setup simulates N samples, one from each proposal, i.e., for $n = 1, \ldots, N$:

1) **Sampling.** Generate samples $\mathbf{x}_n \sim q_n(\mathbf{x})$.
2) **Weighting.** Two common schemes are:
   - Standard MIS (s-MIS): $w_n = \frac{\pi(\mathbf{x}_n)}{q_n(\mathbf{x}_n)}$.
   - Deterministic Mixture MIS (DM-MIS): $w_n = \frac{\pi(\mathbf{x}_n)}{\psi(\mathbf{x}_n)}$, where $\psi(\mathbf{x}) = \frac{1}{N}\sum_{j=1}^N q_j(\mathbf{x})$ is the mixture PDF.

Both schemes allow the construction of the classical IS estimators: (a) the unnormalized IS (UIS) estimator, $\widehat{I} = \frac{1}{NZ}\sum_{n=1}^N w_n h(\mathbf{x}_n)$, which requires the normalization constant Z to be known; and (b) the self-normalized IS (SNIS) estimator, $\widetilde{I} = \sum_{n=1}^N \bar{w}_n h(\mathbf{x}_n)$, where $\bar{w}_n = w_n / \sum_{j=1}^N w_j$. However, the UIS estimator with DM-MIS always outperforms s-MIS in terms of variance reduction as shown in [34]. Different efficient weighting schemes that also reduce variance have been proposed in [29], [30], [31].

C. Adaptive importance sampling

Adaptive importance sampling (AIS) iteratively improves the proposal distributions to reduce estimator variance (see a review in [9]). It adds a third step to MIS, after sampling and weighting in which, typically, the parameters of a mixture proposal are adapted. A key family of AIS algorithms is *population Monte Carlo* (PMC), where this third adaptation step updates the location parameters of the proposals by performing resampling from the current weighted particles. The standard PMC algorithm is described in [12], and further extensions are provided in [11], [32], [26].

D. The proximal operator and proximal methods

Proximal methods are a powerful set of techniques that can be used to optimize general cost functions, possibly involving nonsmooth terms [16], [48]. These algorithms utilize *proximity operators* in order to move towards a fixed-point solution. The proximity operator of a function $g \in \Gamma_0(\mathbb{R}^{d_x})$ (i.e., the set of proper, lower semicontinuous, convex functions from $\mathbb{R}^{d_x}$ to $\mathbb{R} \cup \{+\infty\}$), at a point $x \in \mathbb{R}^{d_x}$, is defined as

$$\mathrm{prox}_g(\mathbf{x}) = \underset{\mathbf{z}\in\mathbb{R}^{d_x}}{\mathrm{argmin}}\, g(\mathbf{z}) + \frac{1}{2}\|\mathbf{z} - \mathbf{x}\|_2^2. \tag{4}$$

Our proposed adaptation scheme builds upon the class of proximal gradient methods, whose core principles are now reviewed.

1) The proximal gradient algorithm: Consider the minimization of $f + g$ over $\mathbb{R}^{d_x}$, where f is differentiable, and $g \in \Gamma_0(\mathbb{R}^{d_x})$. Given some $\mathbf{x}^{(0)} \in \mathbb{R}^{d_x}$, the proximal gradient algorithm iterates as

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t)} = \mathrm{prox}_{\gamma g}\left(\mathbf{x}^{(t-1)} - \gamma\nabla f(\mathbf{x}^{(t-1)})\right), \tag{5}$$

where $\gamma > 0$ is a step-size of the algorithm set to obtain convergence of the sequence (5) to a solution to the problem. For instance, if $f \in \Gamma_0(\mathbb{R}^{d_x})$ and its gradient is L-Lipschitz continuous, $(\mathbf{x}^{(t)})_{t\in\mathbb{N}}$ converges to a minimizer of $f + g$, for $\gamma \in (0, 2/L)$ [13], [54].

2) Scaled forms for proximal gradient method: As a first-order method, the algorithm (5) can display slow convergence. Various accelerated forms of proximal gradient algorithm have been investigated. In our adaptation strategy, we will rely on scaled proximal gradient [17], [8], modifying the underlying metric in (4), as follows. Let $\mathbf{A} \in \mathbb{R}^{d_x \times d_x}$, symmetric definite positive (SDP). The proximal operator computed at $\mathbf{x} \in \mathbb{R}^{d_x}$, of $g \in \Gamma_0(\mathbb{R}^{d_x})$, relative to the metric induced by $\mathbf{A}$, is

$$\mathrm{prox}_{\mathbf{A},g}(\mathbf{x}) = \underset{\mathbf{z}\in\mathbb{R}^{d_x}}{\mathrm{argmin}}\, g(\mathbf{z}) + \frac{1}{2}(\mathbf{z} - \mathbf{x})^\top \mathbf{A}(\mathbf{z} - \mathbf{x}). \tag{6}$$

This new definition yields the following algorithm (again with $x^{(0)} \in \mathbb{R}^{d_x}$), whose convergence has been explored for instance in [6], [7], for various classes of f and $(\mathbf{A}^{(t)})_{t\in\mathbb{N}}$:

$$(\forall t \in \mathbb{N}) \quad \mathbf{x}^{(t)} = \mathrm{prox}_{\mathbf{A}^{(t)},g}\left(\mathbf{x}^{(t-1)} - \mathbf{A}^{(t)}\nabla f(\mathbf{x}^{(t-1)})\right). \tag{7}$$

3) Extension to non-convex settings: The algorithms above have been extended this last decade, to the non-convex settings, under specific assumptions on (f, g) to guarantee the well-posedness of the iterates, in particular, to ensure that the proximal map of g is well-defined (uniqueness/existence). Convergence guarantees to critical points, can be obtained under the Kurdyka-Łojasewicz framework [4], [14], [5].

III. Proximal Newton adaptive importance sampler

We display our proposed PNAIS algorithm in Table I to adapt the set of N proposals for $t = 1, ..., T$ iterations. We denote them as $\{q_t(\mathbf{x}; \boldsymbol{\mu}_n^{(t)}, \boldsymbol{\Sigma}_n^{(t)}, \boldsymbol{\nu}_n)\}_{n=1}^N$, where $\boldsymbol{\mu}_n^{(t)}$ and $\boldsymbol{\Sigma}_n^{(t)}$ are the adapted location and scale parameters, respectively, and $\boldsymbol{\nu}_n$ are the static parameters (e.g., degrees of freedom in

TABLE I
PNAIS ALGORITHM.

1) **[Initialization]**: Set $\sigma > 0$, $(N, K, T) \in \mathbb{N}^+$, $\{\boldsymbol{\nu}_n\}_{n=1}^N$. For $n = 1, \dots, N$, select the initial adaptive parameters $\boldsymbol{\mu}_n^{(1)} \in \mathbb{R}^{d_x}$ and $\boldsymbol{\Sigma}_n^{(1)} = \sigma^2 \mathbf{I}_{d_x}$.
2) **[For $t = 1$ to T]**:
   a) **[Sampling]**: Simulate NK samples as
   $$\mathbf{x}_{n,k}^{(t)} \sim q_n^{(t)}(\mathbf{x}; \boldsymbol{\mu}_n^{(t)}, \boldsymbol{\Sigma}_n^{(t)}, \boldsymbol{\nu}_n) \tag{8}$$
   with $n = 1, \dots, N$, and $k = 1, \dots, K$.
   b) **[Weighting]**: Calculate the normalized IS weights as
   $$w_{n,k}^{(t)} = \frac{\pi(\mathbf{x}_{n,k}^{(t)})}{\frac{1}{N}\sum_{i=1}^N q_i^{(t)}(\mathbf{x}_{n,k}^{(t)})}. \tag{9}$$
   c) **[Adaptation]**: Adapt the location and scale parameters of the proposal
      i) **[Resampling step]** Resample N proposals densities from the pool of NK weighted samples at the iteration t. The means and scales of the resampled proposals are denoted as $\widetilde{\boldsymbol{\mu}}_n^{(t)}$ and $\widetilde{\boldsymbol{\Sigma}}_n^{(t)}$, respectively. See Section III-A for explicit definitions of the notation.
      ii) **[Optimization step]** Adapt the proposal parameters $\{(\boldsymbol{\mu}_n^{(t+1)}, \boldsymbol{\Sigma}_n^{(t+1)})\}_{n=1}^N$ according to (10)-(13).
3) **[Output, $t = T$]**: Return the pairs $\{\mathbf{x}_{n,k}^{(t)}, w_{n,k}^{(t)}\}$, for $n = 1, \dots, N$, $k = 1, \dots, K$ and $t = 1, \dots, T$.

Student's t-distributions). The algorithm runs over T iterations to finally produce a set of KNT weighted samples, enabling the construction of IS estimators. Three steps are performed at each iteration t, namely sampling K samples per proposal (Step 2a), weighting the samples with the DM-MIS scheme (Step 2b), and adapting the location and scale parameters (Step 2c). The last step involves a resampling procedure, the mean adaptation, and the covariance adaptation; each of these mechanisms is detailed in the next three subsections.

A. Resampling procedure

As in most PMC algorithms, the resampling step simulates the location of the N next proposals. The global resampling (GR) creates a pool of NK samples, with weights proportional to (9) and normalized over the NK samples; GR resamples N location parameters with replacement from the pool. The local resampling (LR) performs instead exactly one drawing from N different pools, each of them composed of the K samples simulated from each proposal, and with weights proportional to (9) and normalized over the K samples. GR generally converges faster but loses diversity, while LR keeps diversity (exactly one sample per proposal survives) at the cost of keeping proposals that may not cover significant probability mass of the target (see more details in [32], [33]). Here, we follow a hybrid approach, called *"glocal"* resampling (GLR) [26], which is well tailored to be followed by the optimization step described in the next section. In GLR, most iterations perform an LR step except that every $\Delta \in \mathbb{N}^+$ iterations, a GR step is performed instead.

B. Mean and covariance adaptation

The mean adaptation follows the optimization strategy in (7). Specifically, at each iteration t, we compute the proposal mean $\boldsymbol{\mu}_n^{(t+1)}$, for every $n \in \{1, \dots, N\}$ by performing one step of the proximal Newton algorithm from [6] initialized at $\widetilde{\boldsymbol{\mu}}_n^{(t)}$. The adapted mean, for a given (n, t), is

$$\boldsymbol{\mu}_n^{(t+1)} = \operatorname{prox}_{\mathbf{A}(\widetilde{\boldsymbol{\mu}}_n^{(t)})^{-1}, g}\left(\widetilde{\boldsymbol{\mu}}_n^{(t)} - \mathbf{A}(\widetilde{\boldsymbol{\mu}}_n^{(t)})\nabla f(\widetilde{\boldsymbol{\mu}}_n^{(t)})\right). \tag{10}$$

Hereabove, $\mathbf{A}(\widetilde{\boldsymbol{\mu}}_n^{(t)})$ is an SDP matrix of $\mathbb{R}^{d_x \times d_x}$ that is scaling the proximal gradient update. Following [6], we define

$$\mathbf{A}(\widetilde{\boldsymbol{\mu}}_n^{(t)}) = \theta_n^{(t)} \boldsymbol{\Gamma}(\widetilde{\boldsymbol{\mu}}_n^{(t)}), \tag{11}$$

with the Newton-like matrix

$$\boldsymbol{\Gamma}(\widetilde{\boldsymbol{\mu}}_n^{(t)}) = \begin{cases} \left(\nabla^2 f(\widetilde{\boldsymbol{\mu}}_n^{(t)})\right)^{-1}, & \text{if } \nabla^2 f(\widetilde{\boldsymbol{\mu}}_n^{(t)}) \succ 0, \\ \widetilde{\boldsymbol{\Sigma}}_n^{(t)}, & \text{otherwise.} \end{cases} \tag{12}$$

Since f might be non convex (i.e., smooth part of the target is not log-concave), we introduce two safe rules. First, if the Hessian of f at $\widetilde{\boldsymbol{\mu}}_n^{(t)}$ is non invertible, we instead set the scaling matrix to the covariance of the proposal that generated the sample. Second, we introduced in (11) a damped factor $\theta_n^{(t)} \in (0, 1]$, tuned according to a backtracking scheme [7] to avoid the degeneracy of the proximal Newton iteration, and thus of our adaptation scheme. Initialized with unit stepsize value, we apply a reduction factor $\tau = 1/2$ until the target increases.

In order to be consistent with the mean adaptation, the covariance matrix of the proposal is also adapted as

$$\boldsymbol{\Sigma}_n^{(t+1)} = \mathbf{A}(\widetilde{\boldsymbol{\mu}}_n^{(t)}). \tag{13}$$

We emphasize that our proposed proposal adaptation method, in (10)-(13), is reminiscent from the preconditioned proximal unadjusted Langevin algorithm (PP-ULA) introduced in [18] in the context of an MCMC sampler for ultrasound imaging. In particular, as the authors of [18] show in their appendix, the considered adaptation can be viewed as the Euler discretization of the Langevin diffusion equation applied to the target π with preconditioning matrix (13). The convergence of the scheme (10), without resampling (i.e., $\widetilde{\boldsymbol{\mu}}_n^{(t)} \equiv \boldsymbol{\mu}_n^{(t)}$), to a minimizer of $f + g$ has been established in the convex setting in [42]. The non-convex case has been studied, for instance in [14], [50], under extra assumptions on the scaling matrix and stepsize, later extended to the stochastic setting in [37].

C. Practical calculation of the proximal step

In most useful applications of PNAIS, the scaling metric given by (11)-(12) might be non trivial (e.g., not diagonal) and/or function g might be complicated (e.g., non separable), hence the evaluation of the proximity operator in (10) might not take a closed-form, and an inner solver is necessary. Several approaches are possible. Here, we opted for the dual forward-backward algorithm [15], provided in the supplementary material. This algorithm only requires the expression for the proximity operator of g, which is simple for a wide class of examples.[1] Note that parallelized/distributed implementations for the dual forward backward algorithm are also available [1].

[1] See https://proximity-operator.net/

IV. Numerical examples

We now present two numerical examples concerning a distribution constrained in the simplex (Example A) and a non-differentiable target (Example B), both with $d_x = 2$. We refer the reader to our supplementary file, for an Example C, on a high-dimensional setting. All our results are averaged over 100 independent runs. We use Gaussian proposals, with the location parameters initialized with random elements from $[0, 1]^{d_x}$. Except otherwise stated, we set $\sigma = 1$ (i.e., initial standard deviation of isotropic covariances). In particular, the initial samples do not necessarily belong to the domain of g. The GLR of Section III-A has a parameter $\Delta = 5$, and we set $(N, K, T) = (50, 20, 20)$. We refer the reader to our supplementary material, for numerical results using the LR strategy.

Ablation study. Our experimental analysis takes the form of an ablation study. Namely, we compare the performance of PNAIS, to modified versions of it, where some features have been changed/discarded. Our competitors are as follows:

- DM-PMC: We discard step c) of PNAIS algorithm, that is we do not perform any proposal adaptation.
- PNAIS-nocov: We simplify (13), and set, for every (n, t), $\mathbf{\Sigma}_n^{(t)} = \sigma^2 \mathbf{I}_{d_x}$.
- PNAIS-rcov: Instead of (13), we use the robust covariance adaptation from [24].
- PNAIS-grad: We modify the mean adaptation (10) using the standard proximal gradient step (i.e., no Newton scaling is performed),

$$\boldsymbol{\mu}_n^{(t+1)} = \mathrm{prox}_{\theta_n^{(t)} g}\left(\widetilde{\boldsymbol{\mu}}_n^{(t)} - \theta_n^{(t)} \nabla f(\widetilde{\boldsymbol{\mu}}_n^{(t)})\right), \qquad (14)$$

with stepsize $\theta_n^{(t)}$ computed following a similar backtracking procedure than PNAIS.

A. Example A: Gaussian mixture over the simplex

We consider a truncated version of an equally weighted mixture of bivariate Gaussian distributions, with means $[0.1, 0.3]^\top$ and $[0.7, 0.4]^\top$ and both covariances equal to $[0.01, 0; 0, 0.01]$. The mixture is truncated to be defined only in the unit simplex i.e., $\mathbf{x} \geq 0$, $x_1 + x_2 \leq 1$, as displayed in Fig. 1 (left). We define f as the neg-logarithm of the Gaussian mixture distribution, and g as the indicator function of the simplex set. The proximity operator of g is the projection over this set. The ground truth values, for the mean, second-order moment, and normalization constants, determined by numerical integration with a rough grid, are, respectively, $E_{\tilde{\pi}}[X] = [0.2369, 0.3023]^\top$, $E_{\tilde{\pi}}[X^2] = [0.1024, 0.1005]^\top$ and $Z = 0.5398$. The results for estimating these values, are summarized in Table II (left), in terms of relative mean squared error (MSE). The proposed PNAIS clearly outperforms its competitors. In particular, DM-PMC is largely behind in terms of performance. Moreover, the ablated versions of PNAIS present limitations.

B. Example B: Gaussian likelihood with sparse prior

We consider a target of the form (3), with $p(\mathbf{y}|\mathbf{x})$ a Gaussian distribution with mean $[0.5, 0.5]^\top$ and covariance $[0.25, 0; 0, 0.25]$, and $p(\mathbf{x}) \propto \exp(-\alpha\|\mathbf{x}\|_1)$, with $\alpha = 2$. The target function is displayed in Figure 1 (right). We define f as the neg-logarithm of the Gaussian distribution, and $g = \alpha\|\cdot\|_1$. The proximity operator of g is the soft-thresholding operator with scale α. We aim at estimating the ground truth integrals $E_{\tilde{\pi}}[X] = [0.2025; 0.2025]^\top$, $E_{\tilde{\pi}}[X^2] = [0.252, 0.252]^\top$, and $Z = 0.1641$, using PNAIS or its variants. The results are summarized in Table II (right). In this example, the first and second order moments are well estimated by all methods. PNAIS shows superior performance with the Hessian-based covariance adaptation and without metric acceleration in the proximal gradient mean adaptation. The normalization constant estimation is difficult in this example, due to an asymmetric target shape with highly non Gaussian shape. Here, the proposed PNAIS is largely superior to its competitors, showing the interest of both adaptation schemes to accurately explore the target. In both examples, we noted a similar time complexity of PNAIS, when compared to its competitors.

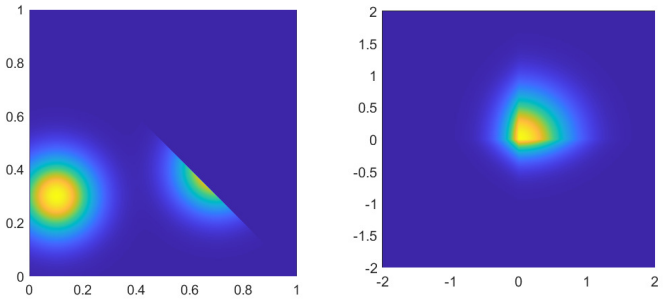

Fig. 1. (left) Target function for Example A. The Target equals 0 outside the unit simplex set. (right) Target function for Example B.

		Example A			Example B		
		$E_{\tilde{\pi}}[X]$	$E_{\tilde{\pi}}[X^2]$	Z	$E_{\tilde{\pi}}[X]$	$E_{\tilde{\pi}}[X^2]$	Z
DM-PMC	$\sigma = 1$	1.12×10^{-4}	6.02×10^{-5}	1.80×10^{-3}	2.96×10^{-5}	1.79×10^{-5}	1.09×10^{-5}
	$\sigma = 3$	2.16×10^{-3}	1.07×10^{-3}	3.58×10^{-2}	2.75×10^{-4}	1.23×10^{-4}	1.16×10^{-4}
	$\sigma = 5$	2.96×10^{-1}	6.70×10^{3}	1.75×10^{-1}	7.50×10^{-4}	4.09×10^{-4}	3.64×10^{-4}
PNAIS-grad	nocov	1.39×10^{-4}	7.31×10^{-5}	1.40×10^{-3}	2.68×10^{-5}	1.56×10^{-5}	7.22×10^{-6}
	rcov	1.96×10^{-5}	7.32×10^{-6}	1.12×10^{-4}	6.68×10^{-5}	7.45×10^{-5}	3.75×10^{-6}
	(13)	5.69×10^{-3}	3.65×10^{-3}	8.87×10^{-2}	$\mathbf{1.10 \times 10^{-5}}$	$\mathbf{1.13 \times 10^{-5}}$	1.07×10^{-6}
PNAIS	nocov	1.30×10^{-4}	5.83×10^{-5}	1.18×10^{-3}	3.78×10^{-5}	2.28×10^{-5}	1.06×10^{-5}
	rcov	8.98×10^{-6}	4.37×10^{-6}	4.77×10^{-5}	2.32×10^{-3}	2.08×10^{-3}	4.63×10^{-4}
	(13)	$\mathbf{5.02 \times 10^{-6}}$	$\mathbf{2.45 \times 10^{-6}}$	$\mathbf{1.63 \times 10^{-5}}$	1.56×10^{-5}	1.81×10^{-5}	$\mathbf{5.64 \times 10^{-7}}$

TABLE II
Relative MSE for Examples A (left) and B (right).

V. Conclusion

The success of AIS heavily depends on the effective adaptation of proposal distributions. In this paper, we proposed a proximal Newton adaptive importance sampler, an algorithm that exploits geometric information for estimating expectations with respect to non-smooth target distributions. By leveraging a scaled Newton proximal gradient, the algorithm adapts multiple proposals to approximate targets that are partially non-differentiable. We have shown its good performance in two challenging numerical experiments. Future work will explore reduced complexity extensions of PNAIS in higher dimensions, e.g., for inference in Bayesian neural networks [40].

REFERENCES

[1] F. Abboud, M. Stamm, E. Chouzenoux, J.-C. Pesquet, and H. Talbot. Distributed Algorithms for Scalable Proximity Operator Computation and Application to Video Denoising. *Digital Signal Processing*, 128:103610, Aug. 2022.

[2] O. D. Akyildiz. Global convergence of optimized adaptive importance samplers. *Foundations of Data Science*, 2024.

[3] Ö. D. Akyildiz and J. Míguez. Convergence rates for optimised adaptive importance samplers. *Statistics and Computing*, 31:1–17, 2021.

[4] H. Attouch and J. Bolte. On the convergence of the proximal algorithm for nonsmooth functions involving analytic features. *Mathematical Programming*, 116:5–16, 2009.

[5] H. Attouch, J. Bolte, and B. F. Svaiter. Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward–backward splitting, and regularized gauss–seidel methods. *Mathematical Programming*, 137(1):91–129, 2013.

[6] S. Becker and J. Fadili. A quasi-newton proximal splitting method. In *Advances in Neural Information Processing Systems*, volume 25, 2012.

[7] S. Bonettini, I. Loris, F. Porta, and M. Prato. Variable metric inexact line-search-based methods for nonsmooth optimization. *SIAM Journal on Optimization*, 26(2):891–921, 2016.

[8] J. F. Bonnans, J. C. Gilbert, C. Lemaréchal, and C. A. Sagastizábal. A family of variable metric proximal methods. *Mathematical Programming*, 68:15–47, 1995.

[9] M. F. Bugallo, V. Elvira, L. Martino, D. Luengo, J. Miguez, and P. M. Djuric. Adaptive Importance Sampling: The past, the present, and the future. *IEEE Signal Processing Magazine*, 34(4):60–79, 2017.

[10] J. V. Candy. *Bayesian Signal Processing: Classical, Modern, and Particle Filtering Methods*, volume 54. John Wiley & Sons, 2016.

[11] O. Cappé, R. Douc, A. Guillin, J.-M. Marin, and C. P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18:447–459, 2008.

[12] O. Cappé, A. Guillin, J.-M. Marin, and C. P. Robert. Population Monte Carlo. *Journal of Computational and Graphical Statistics*, 13(4):907–929, 2004.

[13] G. H. Chen and R. T. Rockafellar. Convergence rates in forward–backward splitting. *SIAM Journal on Optimization*, 7(2):421–444, 1997.

[14] E. Chouzenoux, J.-C. Pesquet, and A. Repetti. Variable metric forward-backward algorithm for minimizing the sum of a differentiable function and a convex function. *Journal of Optimization Theory and Applications*, 162(1):107–132, July 2014.

[15] P. L. Combettes, D. Dung, and B. C. Vu. Proximity for sums of composite functions. *Journal of Mathematical Analysis and Applications*, 380(2):680–688, Aug. 2011.

[16] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point Algorithms for Inverse Problems in Science and Engineering*, pages 185–212. Springer, 2011.

[17] P. L. Combettes and B. C. Vũ. Variable metric forward–backward splitting with applications to monotone inclusions in duality. *Optimization*, 63(9):1289–1318, 2014.

[18] M.-C. Corbineau, D. Kouamé, E. Chouzenoux, J.-Y. Tourneret, and J.-C. Pesquet. Preconditioned P-ULA for Joint Deconvolution-Segmentation of Ultrasound Images. *IEEE Signal Processing Letters*, 26(10):1456–1460, Oct. 2019.

[19] P. M. Djuric, J. H. Kotecha, J. Zhang, Y. Huang, T. Ghirmai, M. F. Bugallo, and J. Miguez. Particle filtering. *IEEE Signal Processing Magazine*, 20(5):19–38, 2003.

[20] A. Durmus, S. Majewski, and B. Miasojedow. Analysis of langevin monte carlo via convex optimization. *Journal of Machine Learning Research*, 20(73):1–46, 2019.

[21] Y. El-Laham and M. F. Bugallo. Stochastic gradient population monte carlo. *IEEE Signal Processing Letters*, 27:46–50, 2019.

[22] Y. El-Laham and M. F. Bugallo. Policy gradient importance sampling for bayesian inference. *IEEE Trans. on Sig. Proc.*, 69:4245–4256, 2021.

[23] Y. El-Laham, P. M. Djurić, and M. F. Bugallo. A variational adaptive population importance sampler. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5052–5056. IEEE, 2019.

[24] Y. El-Laham, V. Elvira, and M. Bugallo. Robust covariance adaptation in adaptive importance sampling. *IEEE Signal Processing Letters*, 25(8):1049–1053, 2018.

[25] V. Elvira and É. Chouzenoux. Langevin-based strategy for efficient proposal adaptation in population monte carlo. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Proc. (ICASSP 2019)*, pages 5077–5081, 2019.

[26] V. Elvira and E. Chouzenoux. Optimized population Monte Carlo. *IEEE Transactions on Signal Processing*, 70:2489–2501, 2022.

[27] V. Elvira, E. Chouzenoux, Ö. D. Akyildiz, and L. Martino. Gradient-based adaptive importance samplers. *Journal of the Franklin Institute*, 360(13):9490–9514, 2023.

[28] V. Elvira and L. Martino. Advances in importance sampling. *arXiv preprint arXiv:2102.05407*, 2021.

[29] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Efficient multiple importance sampling estimators. *IEEE Signal Processing Letters*, 22(10):1757–1761, 2015.

[30] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Heretical multiple importance sampling. *IEEE Sig. Proc. Letters*, 23(10):1474–1478, 2016.

[31] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Multiple importance sampling with overlapping sets of proposals. In *Proceedings of the IEEE Workshop on Statistical Signal Processing (SSP 2016)*, 2016.

[32] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Improving population monte carlo: Alternative weighting and resampling schemes. *Signal Processing*, 131:77–91, 2017.

[33] V. Elvira, L. Martino, D. Luengo, and M. F. Bugallo. Population Monte Carlo schemes with reduced path degeneracy. In *Proceedings of the 7th IEEE International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP 2017)*, pages 1–5, 2017.

[34] V. Elvira, L. Martino, D. Luengo, M. F. Bugallo, et al. Generalized multiple importance sampling. *Statistical Science*, 34(1):129–155, 2019.

[35] V. Elvira, L. Martino, D. Luengo, and J. Corander. A gradient adaptive population importance sampler. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2015)*, pages 4075–4079, 2015.

[36] P. C. Encinar, F. R. Crucinio, and O. D. Akyildiz. Proximal interacting particle langevin algorithms. *arXiv preprint arXiv:2406.14292*, 2024.

[37] G. Fort and E. Moulines. Stochastic variable metric proximal gradient with variance reduction for non-convex composite optimization. *Statistics and Computing*, 33(3), Apr. 2023.

[38] J. V. Goldman, T. Sell, and S. S. Singh. Gradient-based markov chain Monte Carlo for Bayesian inference with non-differentiable priors. *Journal of the American Stat. Association*, 117(540):2182–2193, 2022.

[39] A. Habring, M. Holler, and T. Pock. Subgradient langevin methods for sampling from nonsmooth potentials. *SIAM Journal on Mathematics of Data Science*, 6(4):897–925, 2024.

[40] Y. Huang, E. Chouzenoux, V. Elvira, and J.-C. Pesquet. Efficient Bayes Inference in Neural Networks through Adaptive Importance Sampling. *Journal of The Franklin Institute*, 360(16):12125–12149, Sept. 2023.

[41] O. Kviman, H. Melin, H. Koptagel, V. Elvira, and J. Lagergren. Multiple importance sampling elbo and deep ensembles of variational approximations. In *International Conference on Artificial Intelligence and Statistics*, pages 10687–10702. PMLR, 2022.

[42] J. D. Lee, Y. Sun, and M. A. Saunders. Proximal newton-type methods for minimizing composite functions. *SIAM Journal on Optimization*, 24(3):1420–1443, 2014.

[43] H. Liao, X. Qian, J. Z. Huang, and P. Li. Rare event detection by acquisition-guided sampling. *IEEE Transactions on Automation Science and Engineering*, 2024.

[44] J. S. Liu and J. S. Liu. *Monte Carlo strategies in scientific computing*, volume 10. Springer, 2001.

[45] D. Luengo, L. Martino, M. Bugallo, V. Elvira, and S. Särkkä. A survey of Monte Carlo methods for parameter estimation. *EURASIP Journal on Advances in Signal Processing*, 2020:1–62, 2020.

[46] G. Mateos, J. A. Bazerque, and G. B. Giannakis. Distributed sparse linear regression. *IEEE Trans. on Sig. Proc.*, 58(10):5262–5276, 2010.

[47] A. B. Owen. *Monte Carlo theory, methods and examples*. https://artowen.su.domains/mc/, 2013.

[48] N. Parikh and S. Boyd. Proximal algorithms. *Foundations and Trends® in Optimization*, 1(3):127–239, 2014.

[49] M. Pereyra. Proximal Markov chain Monte Carlo algorithms. *Statistics and Computing*, 26(4):745–760, 2016.

[50] A. Repetti and Y. Wiaux. Variable metric forward-backward algorithm for composite minimization problems. *SIAM Journal on Optimization*, 31(2):1215–1241, 2021.

[51] C. P. Robert and G. Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer-Verlag New York, 1999.

[52] I. Schuster. Gradient importance sampling. *arXiv preprint arXiv:1507.05781*, 2015.

[53] A. Shukla, D. Vats, and E. C. Chi. Mcmc importance sampling via moreau-yosida envelopes. *arXiv preprint arXiv:2501.02228*, 2025.

[54] P. Tseng. A modified forward-backward splitting method for maximal monotone mappings. *SIAM Journal on Control and Optimization*, 38(2):431–446, 2000.

SUPLEMENTARY MATERIAL

Dual forward-backward algorithm to compute the proximal step

We summarize in Table S-I the iterations of the dual forward-backward algorithm [15], to compute the proximity operator of a function $g \in \Gamma_0(\mathbb{R}^{d_x})$, within the metric induced by an SDP matrix $\mathbf{A}^{-1}$, at some point $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^{d_x}$. Let $\|\mathbf{L}\|$ denote the spectral norm of matrix $\mathbf{L}$. Note that we use a particular form of the method in [16], with a single proximable term, a unit stepsize, and a specific choice for the initialization. The sequence $(\mathbf{L}\boldsymbol{\xi}_j)_{j\geq 1}$ converges to the solution of the sought problem (as a consequence of [15, Theorem 2.2], and continuity of the linear operator $\mathbf{L}$). In practice, a few iterations are needed to reach stability, and one can exit the loop as soon as the relative norm difference between two consecutive iterates of $(\boldsymbol{\xi}_j)_{j\geq 1}$ gets lower than some low value (typically, 10^{-7}).

TABLE S-I
DUAL FORWARD BACKWARD ALGORITHM TO COMPUTE $\text{PROX}_{\mathbf{A}^{-1},g}(\tilde{\boldsymbol{\xi}})$.

1) **[Initialization]**: Set $\mathbf{A} \in \mathbb{R}^{d_x \times d_x}$ SDP, $\tilde{\boldsymbol{\xi}} \in \mathbb{R}^{d_x}$, and $g \in \Gamma_0(\mathbb{R}^{d_x})$. Set iteration number $J > 0$.
   Set $\mathbf{L} = \mathbf{A}^{1/2}$, $\tilde{\boldsymbol{\zeta}} = \mathbf{L}^{-1}\tilde{\boldsymbol{\xi}}$, $\rho = \|\mathbf{L}\|^2$, $\boldsymbol{\zeta}_1 = \mathbf{L}\tilde{\boldsymbol{\xi}}$.
2) **[For** $j = 1, \ldots, J$**]**:
   $\boldsymbol{\xi}_j = \tilde{\boldsymbol{\zeta}} - \mathbf{L}^\top \boldsymbol{\zeta}_j$
   $\tilde{\boldsymbol{\zeta}}_j = \boldsymbol{\zeta}_j + \rho^{-1}\mathbf{L}\boldsymbol{\xi}_j$
   $\boldsymbol{\zeta}_{j+1} = \tilde{\boldsymbol{\zeta}}_j - \rho^{-1}\text{prox}_{\rho g}(\rho\tilde{\boldsymbol{\zeta}}_j)$
3) **[Output]**: Return $\mathbf{L}\boldsymbol{\xi}_J$

Additional results

In Tables S-II and S-III, we provide the results for Examples A and B, respectively, using LR resampling strategy, instead of the GLR one, in step c)i) of PNAIS algorithm. Again, the proposed method displays excellent performance on both examples. In Example A, it reaches the best results among the competitors. In Example B, the first-order version of PNAIS is slightly better (i.e., lower MSE) for the three estimated quantities. Let us now compare the results to those obtained with the GLR strategy, on the same tasks by inspecting Table II, in the main file. In Example A, the LR strategy is slightly superior to GLR, in terms of relative MSE, in most cases. In contrast, on Example B, higher performance was obtained when using the GLR strategy, in particular for estimating the mean and the normalization constant of the target.

	DM-PMC			PNAIS-grad			PNAIS		
	$\sigma = 1$	$\sigma = 3$	$\sigma = 5$	nocov	rcov	(13)	nocov	rcov	(13)
$E_{\tilde{\pi}}[X]$	3.3956e-04	1.1882e-02	1.5891e+00	2.1561e-04	2.1532e-05	3.7829e-03	1.0816e-04	1.4407e-05	**4.2893e-06**
$E_{\tilde{\pi}}[X^2]$	1.5761e-04	5.2274e-03	1.9780e+04	1.0776e-04	8.9325e-06	2.1195e-03	4.7324e-05	6.6360e-06	**2.1160e-06**
Z	1.1815e-03	1.1186e-01	5.5812e-01	2.2864e-03	2.0005e-04	1.3271e-01	1.9897e-03	5.6512e-05	**1.0255e-05**

TABLE S-II
EXAMPLE A. RELATIVE MSE USING LR APPROACH.

	DM-PMC			PNAIS-grad			PNAIS		
	$\sigma = 1$	$\sigma = 3$	$\sigma = 5$	nocov	rcov	(13)	nocov	rcov	(13)
$E_{\tilde{\pi}}[X]$	2.6789e-05	2.2452e-04	8.8604e-04	3.6815e-05	5.4597e-03	**1.4897e-05**	2.4109e-05	7.0972e-03	2.3900e-05
$E_{\tilde{\pi}}[X^2]$	1.9983e-05	1.6666e-04	4.3758e-04	2.0217e-05	1.1799e-02	**1.0674e-05**	1.7564e-05	1.1835e-02	2.8333e-05
Z	7.5993e-06	1.8578e-04	3.9083e-04	8.2244e-06	5.3964e-04	**1.2680e-06**	5.2137e-06	1.0151e-03	2.2691e-06

TABLE S-III
EXAMPLE B. RELATIVE MSE USING LR APPROACH.

In Table S-IV, we consider the problem of estimating the mean of a truncated banana-shaped distribution in up to $d_x = 100$ dimensions, constrained to the domain $\|x\|_2 \leq 4$ (see Fig. S-V(left)). The target distribution is defined as in [26, Sec. V-B]. The projection onto the ℓ_2-ball constraint has a closed form (see for instance, https://proximity-operator.net/). The same settings as in Examples A and B, were used, for (N, K, T, σ), and all methods implement GLR resampling. The Table S-IV displays the results of the different methods, in terms of relative MSE, for the estimation of the mean of the truncated distribution, in the cases $d_x \in \{2, 10, 50\}$ ($\times$ means that the method does not manage to stabilize). Note that the first coordinate of the mean is discarded in the MSE computation, as the ground truth is not known for this example. The table shows the clear superiority of our approach. In Fig. S-V(right), we show the computational time when running the methods all implemented with Matlab R2023a code, on a Intel(R) Core(TM) i7-8650U CPU @ 1.90GHz 2.11 GHz, 16Go RAM, for different problem sizes, $d_x \in \{2, 5, 10, 20, 50, 100\}$. We observe that the complexity increases with dimension in all methods. Our method does

not significantly increase the computation time compared to competitors, probably because the complexity is dominated by the sampling and resampling steps.

	DM-PMC			PNAIS-grad			PNAIS		
	$\sigma = 1$	$\sigma = 3$	$\sigma = 5$	nocov	rcov	(13)	nocov	rcov	(13)
$d_x = 2$	9.9112e-05	3.5901e-05	3.8034e-04	1.4155e-04	3.4049e-05	9.4337e-03	1.4751e-04	7.1918e-05	**6.7664e-06**
$d_x = 10$	1.7353e-03	5.6920e+00	1.3211e+01	4.0847e+00	4.3783e-02	6.9631e-02	7.8396e-01	3.7315e-02	**1.0493e-04**
$d_x = 50$	8.7256e-01	4.7679e+00	1.5457e+01	3.4186e+00	×	3.0231e+00	2.0223e+00	×	**4.0775e-01**

TABLE S-IV

EXAMPLE C. RELATIVE MSE ON THE MEAN OVER ALL COORDINATES BUT THE FIRST (FOR WHICH NO GROUND TRUTH IS AVAILABLE).

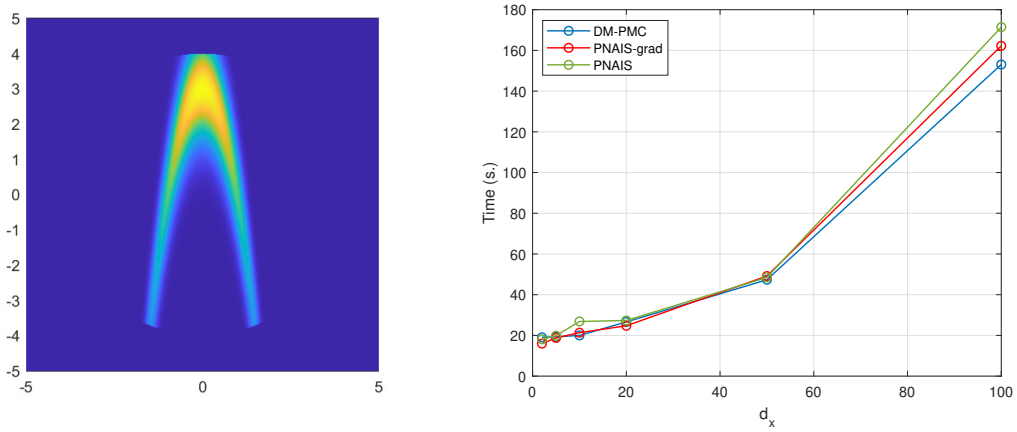

Fig. S-V. **Example C.** Left: two first dimensions of the target distribution with ℓ_2 ball constraint (with radius 4). Right: Computational time per run as a function of d_x for DM-PMC ($\sigma = 1$), PNAIS-grad (covariance adaptation as in (13)), and the proposed PNAIS method.