Robust Importance Sampling for Rare Events via Constrained Gaussian Mixtures

Paweł Lorek[1,2*] **Rafał Nowak**[1,2] **Rafał Topolnicki**[1,2,3]
Tomasz Trzciński[2,4,5] **Maciej Zięba**[2,6]

[1]University of Wrocław [2]Tooploox [3]TRAILS University of Warsaw
[4]Warsaw University of Technology [5]IDEAS Research Institute
[6]Wrocław University of Science and Technology

Abstract

We study estimating rare-event probabilities $I = \mathbb{P}(g(\mathbf{X}) > \gamma)$ with $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and general $g : \mathbb{R}^d \to \mathbb{R}$. We address this problem through importance sampling, and propose a framework that substantially improves efficiency and robustness over baselines such as crude Monte Carlo, adaptive cross-entropy, variational-inference-based methods (including reverse- and forward-KL approaches), as well as Safe-ICE, Subset Simulation, and Sequential Monte Carlo, *drawing* on ideas from both rare-event estimation and cross-entropy optimization. The key contribution has two parts: first, we separate the problem into *coverage*, to overcome the cold-start barrier, and *fitting*, to refine proposals once a meaningful signal is available; second, we constrain the final GMM proposal so that it has finite importance-sampling variance (since coverage alone is not sufficient – without safeguards, importance sampling may still suffer from infinite variance). Together, these ingredients yield expressive proposals; finite variance does not by itself guarantee practical stability at a fixed sampling budget. Extensive experiments demonstrate substantial variance reduction, strong robustness across diverse benchmarks, and favorable cost–efficiency trade-offs, with the proposed approach often outperforming these baselines, particularly in high-dimensional and multimodal settings where competing methods frequently become unstable or fail. Our code is available at https://github.com/lorek/robust-cfi-is.

1 Introduction

We consider events $\{g(\mathbf{X}) > \gamma\}$ with $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ and possibly nonconvex $g : \mathbb{R}^d \to \mathbb{R}$; let f denote the density of $\mathbf{X}$. Our target is:

$$I \;=\; \mathbb{P}\big(g(\mathbf{X}) > \gamma\big), \tag{1}$$

in regimes where I is tiny (e.g., 10^{-6}–10^{-3}), rendering crude Monte Carlo impractical.

A long line of work has sought to overcome this difficulty with importance sampling, cross-entropy (CE) methods, and variational inference (VI). While in some cases these approaches can reduce variance dramatically, in others they fail completely, since they face two recurring obstacles in rare-event regimes. First, *cold start*: when the event probability is extremely small, the learning signal obtained from naive sampling is nearly absent, leading to unstable or stalled optimization. Second, *lack of variance control*: minimizing KL divergence alone does not prevent proposals that yield infinite-variance estimators, even when optimization appears successful. These issues are well documented in reliability engineering, rare-event simulation, and, more recently, in machine

*Corresponding author: pawel.lorek@math.uni.wroc.pl

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

learning. For example Owen [19] noted: "*Importance sampling can bring enormous gains, making an otherwise infeasible problem amenable to Monte Carlo. It can also backfire, yielding an estimate with infinite variance when simple Monte Carlo would have had a finite variance. It is the hardest variance reduction method to use well.*"

We address these challenges by introducing a pipeline that cleanly separates coverage from fitting and builds robustness into the proposal family. The key idea is to first construct a proposal that is designed to provide overlap with the rare set, ensuring that subsequent training receives a meaningful signal. We then refine this proposal within a constrained Gaussian mixture family whose covariance structure precludes the infinite-variance pathology often encountered in importance sampling. Finally, we optionally add a normalizing-flow layer to capture additional complexity. The finite-variance guarantee applies when the final proposal remains a constrained GMM and does not automatically extend to the flow-refined variant (although our experiments show that flow refinement often improves variance reduction while remaining stable across the tested benchmarks).

In summary, our contributions are as follows: (i) We propose a two-stage, coverage-first framework for rare-event estimation that overcomes cold-start issues. (ii) We extend a sufficient finite-variance condition from Gaussian to Gaussian mixture proposals and make it enforceable during proposal learning. (iii) We demonstrate strong variance reduction and robustness through extensive experiments across diverse benchmarks.

2 Background

We consider the problem of estimating rare probabilities, given by eq. (1). To tackle the problem, we utilize the importance sampling (IS) technique:

$$I = \mathbb{E}_{\mathbf{X}\sim h}[k_1(\mathbf{X})], \quad k_1(\mathbf{x}) = \mathbb{1}(g(\mathbf{x}) > \gamma)\,\frac{f(\mathbf{x})}{h(\mathbf{x})}, \tag{2}$$

where $h(\mathbf{x})$ is the proposal distribution. Please note that zero-variance choice is $h^{\mathrm{opt}}(\mathbf{x}) = \mathbb{1}(g(\mathbf{x}) > \gamma)f(\mathbf{x})/I$, but to sample from h^{opt}, knowledge[2] of I itself is required. We therefore restrict to a parametric family $h \equiv p_{\boldsymbol{\theta}}$ and fit $p_{\boldsymbol{\theta}}$ to approximate h^{opt}.

Baseline approaches tackle the problem by minimizing a KL divergence, either $\mathrm{KL}(p_{\boldsymbol{\theta}}\|h^{\mathrm{opt}})$ (the so-called Reverse KL (**RKL**)), or $\mathrm{KL}(h^{\mathrm{opt}}\|p_{\boldsymbol{\theta}})$ (the so-called forward KL (**FKL**)). More explicitly, minimizing the FKL is equivalent to maximizing

$$\max_{\boldsymbol{\theta}}\ \mathbb{E}_{\mathbf{X}\sim f}[\mathbb{1}(g(\mathbf{X}) > \gamma)\,\log p_{\boldsymbol{\theta}}(\mathbf{X})]\,, \tag{3}$$

which corresponds to the classical cross-entropy (CE) objective used in rare-event simulation. In contrast, directly minimizing the reverse KL divergence $\mathrm{KL}(p_{\boldsymbol{\theta}}\|h^{\mathrm{opt}})$ is ill-posed due to indicator – which in practice we replace by a smooth gate $s_\tau(g(\mathbf{x}),\gamma) = \sigma((g(\mathbf{x})-\gamma)/\tau) \in (0,1)$, where $\sigma(z) = (1+e^{-z})^{-1}$. This leads to the surrogate reverse-KL objective

$$\max_{\boldsymbol{\theta}}\ \mathbb{E}_{\mathbf{X}\sim p_{\boldsymbol{\theta}}}\left[\log s_\tau\big(g(\mathbf{X}),\gamma\big) - \log\frac{p_{\boldsymbol{\theta}}(\mathbf{X})}{f(\mathbf{X})}\right], \tag{4}$$

which is commonly used in variational formulations of rare-event estimation (see Appendix A)

Although these techniques are widely used to find surrogates for h^{opt} , they are prone to two major issues while estimating rare probability regions. First, they are sensitive to cold start issues, where a significant number of samples is required to satisfy $(g(\mathbf{x}) > \gamma)$. Second, even when the reverse or forward KL problem is solved exactly, the resulting $p_{\boldsymbol{\theta}}$ may yield an IS estimator with high – even *infinite* – variance. Consequently, KL minimization alone (reverse/forward) does not ensure a safe proposal. In order to eliminate those limitations, we propose a novel, two-stage (optionally refined further by a normalizing flow model) technique that learns a proposal $p_{\boldsymbol{\theta}}$ robust to these issues.

3 Related Work

Estimating rare-event probabilities has a long history across simulation, reliability engineering, and machine learning. Au and Beck introduced Subset Simulation, which uses a sequence of intermediate

[2]MCMC may (in some cases) sample from unnormalized targets, but breaks i.i.d. draws and standard unbiased IS.

events to estimate a rare probability progressively [3]. Sequential importance sampling methods adapt proposals over stages to concentrate samples in the rare-event region [21]; related particle methods include adaptive multilevel splitting and SMC-based rare-event estimators [4]. Rubinstein formalized the cross-entropy (CE) method [27], which Rubinstein and Kroese developed further [28]. Subsequent work introduced adaptive CE updates and mixture proposals for rare-event simulation [15, 6, 34].

Safe and defensive importance sampling control unstable likelihood ratios by mixing an adaptive proposal with a fixed component. Owen and Zhou developed safe mixture sampling [20], while Hesterberg's defensive mixture includes the nominal density as a fixed component and thus bounds the corresponding likelihood ratio [14]. Delyon and Portier studied adaptive mixtures with a heavy-tailed safe component and established asymptotic guarantees [7]. Our safeguard uses a different mechanism: we constrain every covariance matrix in a learned Gaussian mixture so that all components can move toward the rare-event region while retaining sufficiently heavy Gaussian tails. The resulting guarantee is a finite-variance statement for the final constrained GMM proposal. The half-variance threshold has classical Gaussian precedents: Owen's Examples 9.1 and 9.3 derive it for a scalar Gaussian moment problem and a d-dimensional standard Gaussian with isotropic proposal covariance $\sigma^2\mathbf{I}$, respectively [19]. The corresponding full-covariance condition for a single Gaussian proposal was established in [18]. Here, we extend this condition to Gaussian mixture proposals and incorporate it as an enforceable covariance safeguard during proposal learning. Safe-ICE instead modifies CE through a problem-dependent safe construction [11]; we include it in the empirical comparison.

Adaptive mixture importance sampling also updates the weights and parameters of a proposal mixture online. Douc et al. established convergence results for such schemes and studied Rao–Blackwellized updates [10]. These methods address the validity and efficiency of repeated adaptation. Our covariance constraint addresses a complementary failure mode: a learned Gaussian proposal can fit its target while producing an infinite-variance importance-sampling estimator.

Variational inference (VI) offers a flexible approach to approximate inference. Classical black-box VI uses REINFORCE-style gradient estimators [35, 23], which are unbiased but often have high variance. Work on variance reduction combines VI with overdispersed proposals [29], adaptive schemes [17], and importance-weighted objectives [1]. Jerfel et al. [16] and Su and Chen [30] studied forward-KL proposal refinement for importance sampling, linking variational objectives to sampling efficiency. We focus on KL objectives because they yield direct Monte Carlo estimators for the rare-event target; Appendix H explains why Wasserstein and KSD objectives are less convenient here.

Generative models provide more flexible proposal families. Rezende and Mohamed introduced normalizing flows [25], followed by tractable architectures including NICE [8], RealNVP [9], and MAF [22]. Gao et al. [12] and Gibson et al. [13] applied flows to rare-event simulation. Large-deviation methods offer a different route: Tong and Stadler combine large-deviation theory with adaptive importance sampling, using it to initialize a Gaussian biasing distribution and identify an informative low-dimensional subspace in high-dimensional problems [31]. Our method instead uses progressive CE-style discovery, without requiring large-deviation rate-function optimization or a problem-specific dominating-point calculation, and then fits an explicit constrained mixture with optional flow refinement. The next section describes this procedure. Appendix I discusses further large-deviation, neural SMC, and adjoint-based samplers.

4 Our Approach

We first construct a coverage-first proposal, then refine it by constrained KL fitting, optionally followed by normalizing-flow refinement. The finite-variance guarantee applies when the final proposal is a constrained GMM satisfying Theorem 1.

4.1 Coverage-first initializer (CFI)

We first construct a proposal distribution p_{CFI} that allocates a prescribed fraction $1-\rho$ of its mass inside or near $\{g(\mathbf{x}) > \gamma\}$. This provides overlap with the rare-event region and a nontrivial training signal for subsequent RKL and FKL refinement, even in ultra-rare regimes. Like adaptive CE [28, 6, 34], CFI draws samples from the current proposal and progresses through a sequence of intermediate levels. However, CFI performs an unweighted fit to the elite samples and does not evaluate the likelihood-ratio weights $f/p_{\boldsymbol{\theta}_{t-1}}$ used by full adaptive CE.

CFI is therefore used as a coverage-first initialization rather than as the final importance-sampling adaptation procedure. The full procedure is given in Alg. 1.

At each iteration, CFI samples from the current proposal, updates the intermediate level using the empirical $(1-\rho)$-quantile, and refits to elite samples with parameter smoothing.

CFI vs gradient-based mode seeking. A natural question is whether the coverage stage of CFI can be replaced by gradient-based optimization of $g(\mathbf{x})$, e.g., via multiple gradient ascent restarts. In Appendix J we show that this approach is unreliable, as it can systematically miss relevant modes or diverge, failing to provide adequate coverage even in simple settings.

Algorithm 1: CFI: Coverage-First Initialization

Require: Family $\{p_{\boldsymbol{\theta}}\}$ with $\boldsymbol{\theta}_0$, function g, level γ, quantile ρ, batch M, smoothing $\alpha \in [0,1)$

$t = 0$

repeat

  $t \leftarrow t+1$; sample $\mathbf{X}_{1:M} \sim p_{\boldsymbol{\theta}_{t-1}}$

  Set $\eta_i = g(\mathbf{X}_i)$, sort: $\eta_{(1)} \le \ldots \le \eta_{(M)}$

  $(1-\rho)$-quantile: $\gamma_t \leftarrow \eta_{(\lceil (1-\rho)M \rceil)}$

  CE update:

$$\boldsymbol{\theta}'_t = \arg\max_{\boldsymbol{\theta}} \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}\big(g(\mathbf{X}_i) \ge \gamma_t\big) \log p_{\boldsymbol{\theta}}(\mathbf{X}_i)$$

  $\boldsymbol{\theta}_t \leftarrow \alpha\, \boldsymbol{\theta}'_t + (1-\alpha)\, \boldsymbol{\theta}_{t-1}$

until $\gamma_t \ge \gamma$

Return $p_{\mathrm{CFI}} \equiv p_{\boldsymbol{\theta}_t}$

4.2 Distribution proposals

A crucial aspect of CFI is the selection of the proposal distribution family. In our approach, we propose using Gaussian Mixture Models (GMMs). Since our method relies on gradient-based optimization, we first introduce an adjusted parameterization of the model, which we refer to as the unconstrained GMM. However, the unconstrained GMM can lead to infinite variance in importance sampling (IS), see e.g., a toy example in Section 4.5. To address this issue, we propose a parameterization of a constrained GMM variant.

Unconstrained GMM. The standard proposal is a Gaussian mixture $p_{\boldsymbol{\theta}}(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$, with parameters $\boldsymbol{\theta} = \{\boldsymbol{\mu}_k, \mathbf{L}_k, \ell_k\}_{k=1}^{K}$ optimized by gradient methods. For the sake of gradient updates, weights and covariances are parameterized as:

$$\pi_k = \frac{\exp(\ell_k)}{\sum_{j=1}^{K} \exp(\ell_j)}, \qquad \boldsymbol{\Sigma}_k = \mathbf{L}_k \mathbf{L}_k^{\top}. \tag{5}$$

Unconstrained Gaussian proposals can lead to unstable or even invalid importance-sampling estimators. The one-dimensional motivating example in Section 4.5 shows that, when both the mean and variance are fitted by KL minimization, the optimizer may choose a variance below the critical threshold required for finite IS variance. This is not merely a one-dimensional artifact, but reflects a more general covariance-collapse mechanism.

Proposition 1 (KL-induced covariance collapse)**.** *Let $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$ and consider the* FKL *objective (3) with an unconstrained Gaussian proposal $p_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Then the optimizer satisfies $\boldsymbol{\mu}_{\mathrm{opt}} = \mathbb{E}[\mathbf{X} \mid g(\mathbf{X}) > \gamma]$ and $\boldsymbol{\Sigma}_{\mathrm{opt}} = \mathrm{Cov}(\mathbf{X} \mid g(\mathbf{X}) > \gamma)$. Moreover, if $\lambda_{\min}(\mathrm{Cov}(\mathbf{X} \mid g(\mathbf{X}) > \gamma)) \to 0$ as $\gamma \to \infty$, then $\lambda_{\min}(\boldsymbol{\Sigma}_{\mathrm{opt}}) \to 0$.*

The proof is provided in Appendix B. Proposition 1 shows that forward-KL fitting may inherit the concentration of the rare-event conditional distribution, causing the proposal covariance to collapse. This is problematic for importance sampling, since the second moment contains the term $\int_{\{g(\mathbf{x})>\gamma\}} f(\mathbf{x})^2/p(\mathbf{x})\, d\mathbf{x}$, which may diverge if the proposal decays too quickly in relevant tail directions. Thus, a well-optimized KL objective does not necessarily yield a safe proposal.

Fig. 1 illustrates the same variance pathology for $\mathbb{P}(|X| > 4)$ with $X \sim \mathcal{N}(0,1)$; see also Section 4.5. This motivates imposing an explicit covariance floor. Theorem 1 shows that requiring $\boldsymbol{\Sigma}_k \succ \frac{1}{2}\boldsymbol{\Sigma}$ for every mixture component guarantees finite variance of the resulting importance-sampling estimator.

Theorem 1. *[Finite Variance for Constrained GMMs] Let $f(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ be the nominal density, and let $p(\mathbf{x}) = \sum_k \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ be a GMM proposal. For i.i.d. $\mathbf{X}_i \sim p$ and the IS*

estimator

$$\hat{I}_R^{\mathrm{IS}} = \frac{1}{R}\sum_{i=1}^{R} \mathbb{1}(g(\mathbf{X}_i) > \gamma)\, \frac{f(\mathbf{X}_i)}{p(\mathbf{X}_i)},$$

if $\boldsymbol{\Sigma}_k \succ \frac{1}{2}\,\boldsymbol{\Sigma}$ *for all* k*, then* $\mathrm{Var}\,(\hat{I}_R^{\mathrm{IS}})$ *is finite.*

The proof is provided in Appendix B. The theorem guarantees only finiteness of the variance; it does not imply that the variance is practically small, or guarantee good contribution ESS, representative finite-budget estimates, or accurate and stable behavior at a fixed budget R. Accordingly, the suffix -**c** denotes the covariance constraint, not a promise of good finite-budget performance. The factor $1/2$ matches the Gaussian thresholds in Owen's Examples 9.1 and 9.3 [19]. These examples treat a scalar moment problem and a d-dimensional standard Gaussian with an isotropic proposal, respectively. The full-covariance version for a single Gaussian proposal was established in [18]. Theorem 1 extends this condition to full-covariance Gaussian mixture proposals.

The componentwise condition in Theorem 1 is sufficient but not necessary for a general GMM. In particular, one component with positive mixture weight and sufficiently broad covariance can control the tails of the full mixture even when other components have smaller covariances. Necessity also depends on the geometry of the rare-event set, specifically on whether it is unbounded in directions in which the proposal is too light-tailed. We impose the stronger componentwise condition because it is independent of the geometry of g, directly checkable, and enforceable during optimization.

As a consequence of this theorem, we propose a constrained variant of GMM.

Constrained GMM. The constrained proposal has the same mixture form, with covariance parameterization $\boldsymbol{\Sigma}_k = \mathbf{L}_k\mathbf{L}_k^\top + \beta\,\boldsymbol{\Sigma}$. Theorem 1 applies to the actual final proposal whenever $\boldsymbol{\Sigma}_k \succ \frac{1}{2}\,\boldsymbol{\Sigma}$ for every component. For $\beta > \frac{1}{2}$, this strict inequality follows from the parameterization; for $\beta = \frac{1}{2}$, it additionally requires $\mathbf{L}_k\mathbf{L}_k^\top \succ 0$ for every component. We use $\beta = 0.5$ except during the CFI-c annealing phase described below. At $\beta = 0.5$, the strict condition must be verified for the final proposal; a numerically nonzero $\mathbf{L}_k\mathbf{L}_k^\top$ is insufficient.

Throughout, suffixes -**c** and -**u** denote **c**onstrained and **u**nconstrained models, respectively.

Adaptive β annealing in CFI-c. During the CE iterations, the proposal must remain sufficiently flexible to localize rare-event regions. A fixed $\beta = 0.5$ can be overly conservative at this stage, leading to premature saturation of the level sequence. Therefore, in CFI-c only, we adapt β during training: when progress of the level sequence stagnates, β is temporarily reduced (from initial 0.5) to relax the covariance floor; when progress resumes, β is gradually increased. This adaptive annealing improves coverage and prevents early plateaus. Once the target level γ is reached, we set $\beta = 0.5$ and terminate the CE iterations. A final refitting step at the target level is performed with $\beta = 0.5$; the finite-variance guarantee applies when the resulting final proposal satisfies $\boldsymbol{\Sigma}_k \succ \frac{1}{2}\,\boldsymbol{\Sigma}$ for all k.

Parameter smoothing. In line 7 of Alg. 1, we exponentially smooth the means, mixture weights, and covariance matrices using α. For constrained GMMs, we smooth only the free covariance component $\mathbf{S}_k = \mathbf{L}_k\mathbf{L}_k^\top$, preserving the floor $\beta\boldsymbol{\Sigma}$. Details are given in Appendix F.

4.3 Constrained-KL fitting

Starting from the coverage-first proposal – denote it by p_{CFI} – we refine the proposal within the constrained Gaussian mixture family, either by forward–KL (FKL-c) or reverse–KL (RKL-c). Both variants enforce the covariance floor $\boldsymbol{\Sigma}_k \succ \frac{1}{2}\,\boldsymbol{\Sigma}$, which guarantees finite variance of the resulting importance sampling estimator.

FKL-c (constrained forward KL). As shown in Section 2, minimizing $\mathrm{KL}(h^{\mathrm{opt}} \| p_{\boldsymbol{\theta}})$ is equivalent to maximizing the objective in (3). Since the corresponding expectation under f cannot be evaluated directly in the rare-event regime, we estimate it by importance sampling using the CFI proposal p_{CFI} as the sampling distribution. This yields the constrained forward-KL fitting problem

$$\max_{\boldsymbol{\theta}}\ \mathbb{E}_{\mathbf{X}\sim p_{\mathrm{CFI}}}\left[\mathbb{1}\big(g(\mathbf{X}) > \gamma\big)\,\frac{f(\mathbf{X})}{p_{\mathrm{CFI}}(\mathbf{X})}\,\log p_{\boldsymbol{\theta}}(\mathbf{X})\right], \tag{6}$$

which is optimized over $\boldsymbol{\theta}$ within the constrained family.

RKL-c (constrained reverse KL). Similarly, the reverse-KL surrogate introduced in (4) defines an optimization problem over $\boldsymbol{\theta}$. To evaluate the corresponding expectation in a stable manner, we again rely on the CFI proposal p_{CFI} as the sampling distribution, yielding

$$\max_{\boldsymbol{\theta}} \ \mathbb{E}_{\mathbf{X}\sim p_{\mathrm{CFI}}}\left[\frac{p_{\boldsymbol{\theta}}(\mathbf{X})}{p_{\mathrm{CFI}}(\mathbf{X})}\Big(\log s_\tau\big(g(\mathbf{X}),\gamma\big) - \log\frac{p_{\boldsymbol{\theta}}(\mathbf{X})}{f(\mathbf{X})}\Big)\right]. \tag{7}$$

Both FKL-c and RKL-c operate within the same constrained (-**c**) family, are initialized from p_{CFI} to ensure sufficient coverage, and optimize the corresponding KL objectives by Monte Carlo approximation of the expectations and gradient-based updates under the covariance constraint.

Adaptive selection of the smoothing parameter τ. RKL-based approaches rely on the smoothed gate $s_\tau(g(\mathbf{x}),\gamma)$ and are therefore sensitive to the choice of the smoothing parameter τ. To avoid manual tuning, we set $\tau_t = c\,\mathrm{Std}(g(\mathbf{X})-\gamma)$, where the standard deviation is estimated from samples drawn from the current proposal and $c>0$ is shared across experiments. We clip τ_t to a fixed interval $[\tau_{\min},\tau_{\max}]$ to avoid excessively sharp or flat gates; see Appendix F for details.

4.4 Refinement with normalizing flows

As a final stage, we refine constrained model p^{KL} obtained via RKL or FKL, using a more expressive proposal $p^{\mathrm{flow}}_{\boldsymbol{\theta}}$. To be more specific: we train the flow model to minimize $\mathrm{KL}(p^{\mathrm{flow}}_{\boldsymbol{\theta}}||h^{\mathrm{opt}})$ or $\mathrm{KL}(h^{\mathrm{opt}}||p^{\mathrm{flow}}_{\boldsymbol{\theta}})$, which yields optimizing (4) or (3) with $p_{\boldsymbol{\theta}} \equiv p^{\mathrm{flow}}_{\boldsymbol{\theta}}$. We estimate the expectations by importance sampling using p^{KL} proposals, yielding maximization problem for FKL-c (cf. (6))

$$\mathbb{E}_{\mathbf{X}\sim p^{\mathrm{KL}}}\left[\mathbb{1}\big(g(\mathbf{X})>\gamma\big)\,\frac{f(\mathbf{X})}{p^{\mathrm{KL}}(\mathbf{X})}\,\log p^{\mathrm{flow}}_{\boldsymbol{\theta}}(\mathbf{X})\right].$$

and maximization problem for RKL-c (cf. (7))

$$\mathbb{E}_{\mathbf{X}\sim p^{\mathrm{KL}}}\left[\frac{p^{\mathrm{flow}}_{\boldsymbol{\theta}}(\mathbf{X})}{p^{\mathrm{KL}}(\mathbf{X})}\Big(\log s_\tau\big(g(\mathbf{X}),\gamma\big) - \log\frac{p^{\mathrm{flow}}_{\boldsymbol{\theta}}(\mathbf{X})}{f(\mathbf{X})}\Big)\right].$$

In other words, the flow is trained to better capture the geometry of the rare-event set while remaining anchored to the stable constrained-GMM backbone p^{KL}. This improves robustness empirically (note, however, that because the final proposal is a flow, the finite-variance guarantee of Theorem 1 does not formally apply).

A normalizing flow [25] defines $p^{\mathrm{flow}}_{\boldsymbol{\theta}}$ via an invertible transformation $T_{\boldsymbol{\theta}}$: $p^{\mathrm{flow}}_{\boldsymbol{\theta}}(\mathbf{X}) = p_{\mathbf{Z}}(T_{\boldsymbol{\theta}}(\mathbf{X}))\cdot|\det J_{\boldsymbol{\theta}}(\mathbf{X})|$, where $p_{\mathbf{Z}}$ is a simple base density (p^{KL}, i.e., the GMM, in our case) and $J_{\boldsymbol{\theta}}$ is the Jacobian of $h_{\boldsymbol{\theta}}$. To ensure tractability of $\det J_{\boldsymbol{\theta}}$ we use RealNVP [9], which provides sufficient expressiveness at low computational cost.

4.5 Motivating toy example: $\mathbb{P}(X>4), X\sim\mathcal{N}(0,1)$

We use the one-dimensional problem $I=\mathbb{P}(X>4)$, $X\sim\mathcal{N}(0,1)$, to illustrate why unconstrained KL fitting can be unsafe for importance sampling; full derivations are given in Appendix C.

For fixed-variance Gaussian proposals $p_{\boldsymbol{\theta}}=\mathcal{N}(\mu,1)$, both FKL and RKL yield stable proposals, with optima $\mu^*_{\mathrm{FKL}}=4.2256$ and $\mu^*_{\mathrm{RKL}}=4.2036$ (for $\tau=0.1$). However, once the variance is also optimized, $p_{\boldsymbol{\theta}}=\mathcal{N}(\mu,\sigma^2)$, both objectives choose overly concentrated proposals: $\sigma^2_{\mathrm{FKL}}=0.0466$ and $\sigma^2_{\mathrm{RKL}}=0.4306$, below the critical threshold $\sigma^2=\frac{1}{2}$ required for finite IS variance. Thus, optimizing the KL objective can still produce an infinite-variance estimator.

This pathology extends to Gaussian mixtures: unconstrained KL fitting may collapse individual covariance matrices. This is the covariance-collapse mechanism formalized in Proposition 1. To prevent it, Theorem 1 imposes the multivariate analogue $\boldsymbol{\Sigma}_k \succ \frac{1}{2}\boldsymbol{\Sigma}$, which is sufficient for finite variance. On the related problem $\mathbb{P}(|X|>4)$ with a $K=20$ GMM proposal, Fig. 1 and Table 1 show that CFI-u suffers severe variance pathologies (✗ means it failed), whereas the constrained CFI-c proposal remains stable and accurate. Among the evaluated methods, CFI-c+FKL-c+F achieves the best empirical variance reduction, log-error, and CoV, while the theorem-covered CFI-c+FKL-c variant ranks second on all three metrics. See the definition of evaluation metrics in Section 5.1.

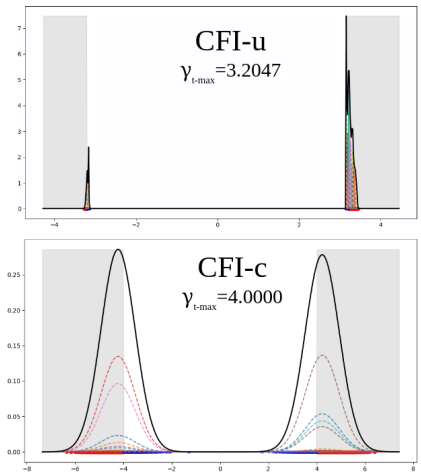

Figure 1: CFI-u vs CFI-c for estimating $\mathbb{P}(|X| > 4)$, $X \sim \mathcal{N}(0,1)$. GMM with $K = 20$ components. Red: samples with $g(x_i) > \gamma_{\mathrm{t_{max}}}$; blue: samples with $g(x_i) \leq \gamma_{\mathrm{t_{max}}}$; gray area: $|x| > 4$.

Table 1: Toy: estimating $\mathbb{P}(|X| > 4)$, $X \sim \mathcal{N}(0,1)$, $K = 20$. FG marks methods whose final proposal is a constrained GMM covered by Theorem 1. **Bold**, ***bold italic***, and *italic* mark the first, second, and third result, respectively.

	$I_{\mathrm{true}} =$	6.33425e-05			
Method	**FG**	$\hat{I}_R$	Var red.	log-err	CoV
Baselines					
CMC		6.55e-05	1	2.65e-01	4.06e-01
fRKL-u		6.30e-05	2.96e+02	1.88e-02	2.37e-02
fFKL-u		0	✗	✗	✗
fRKL-c	✓	6.32e-05	5.89e+02	1.38e-02	1.67e-02
fFKL-c	✓	6.36e-05	3.67e+03	6.41e-03	6.65e-03
AMS		6.33e-05	4.19e-01	1.04e-01	5.98e-01
Subset		6.33e-05	6.81e-01	5.61e-02	5.23e-01
Safe-ICE		6.17E-05	5.25e-02	2.93e-02	5.45e-01
Our methods					
CFI-u		0	✗	✗	✗
CFI-c	✓	6.34e-05	2.94e+03	5.17e-03	7.44e-03
CFI-c+RKL-c	✓	6.33e-05	2.90e+03	5.72e-03	7.51e-03
CFI-c+FKL-c	✓	6.33e-05	***5.56e+03***	***4.49e-03***	***5.42e-03***
CFI-c+RKL-c+F		6.32e-05	*3.82e+03*	*4.62e-03*	*6.54e-03*
CFI-c+FKL-c+F		6.34e-05	**5.81e+03**	**4.39e-03**	**5.29e-03**

Takeaway. KL fitting alone does not guarantee reliable importance sampling: proposals can appear well optimized while losing the tail overlap needed for finite variance. This motivates the covariance constraint and coverage-first initialization used below.

5 Experimental results

5.1 Experimental setup

We report results for the following methods. CFI-u denotes unconstrained coverage-first initialization and already exhibits numerical instabilities and estimator failures across most considered problems; as a consequence, further refinements built on CFI-u are omitted from the main comparison. CFI-c refers to constrained coverage-first initialization. Starting from CFI-c, we consider constrained KL refinement using either reverse or forward KL (CFI-c+RKL-c and CFI-c+FKL-c). Finally, these constrained methods can be further refined using normalizing flows (CFI-c+RKL-c+F and CFI-c+FKL-c+F).

External baselines. We additionally compare against three external rare-event baselines: Subset Simulation (SS), Adaptive Multilevel Splitting (AMS), and Safe-ICE [11, 26]. SS is implemented using `UQpy`, following the original method of Au and Beck [3]. AMS is implemented following the SMC framework of Cérou et al. [4]. For SS and AMS, we use a matched computational budget measured by the number of evaluations of g. Safe-ICE is discussed in Appendix K; because the available implementation is computationally expensive, this comparison is restricted to representative low-, moderate-, and high-dimensional benchmarks. We also consider a Hesterberg-style defensive counterpart of fRKL [14], denoted fRKL-d (d stands for *defensive*). It uses the full defensive proposal

$$p_{\boldsymbol{\theta}}^{\mathrm{def}}(\mathbf{x}) = \lambda f(\mathbf{x}) + (1-\lambda) p_{\mathrm{GMM},\boldsymbol{\theta}}(\mathbf{x}),$$

which ensures $f(\mathbf{x})/p_{\boldsymbol{\theta}}^{\mathrm{def}}(\mathbf{x}) \leq 1/\lambda$. Unlike our covariance constraint, this construction bounds the importance weights through a fixed nominal component.

Additional KL baselines. For completeness, Appendix D additionally reports results for the classical unconstrained reverse- and forward-KL baselines (fRKL-u and fFKL-u, where f stands for **f**ull/original baseline). These methods are representative of classical KL-based importance-sampling approaches and serve primarily to illustrate the instability addressed by our constrained framework: although they can occasionally perform well, they frequently fail catastrophically due to infinite or extremely large variance.

Table 2: Best methods by CoV at $R = 10^5$ simulator evaluations: baseline (CMC, fFKL-c, fRKL-c, CFI-u, AMS, Subset), theorem-covered proposed method with a final constrained GMM, and proposed method overall. Bold marks the overall lowest CoV.

g	d	I_{true}	Best baseline by CoV			Best finite-var. method			Best proposed overall		
			Method	log-err ↓	CoV ↓	Method	log-err ↓	CoV ↓	Method	log-err ↓	CoV ↓
g_1	2	4.79e-6	CMC	1.43e+00	**3.92e-03**	CFI-c+RKL-c	6.26e-03	7.24e-03	CFI-c+FKL-c+F	3.16e-03	**3.92e-03**
g_2	2	2.23e-3	CMC	4.14e-02	2.46e-03	CFI-c+FKL-c	3.39e-03	4.85e-03	CFI-c+RKL-c+F	2.87e-03	**1.86e-03**
g_3	2	4.21e-3	CMC	4.25e-02	**3.31e-03**	CFI-c+FKL-c	3.25e-03	4.35e-03	CFI-c+FKL-c+F	2.80e-03	**3.31e-03**
g_5	40	3.15e-5	fRKL-c	7.31e-02	3.19e-02	CFI-c+RKL-c	7.28e-02	4.30e-02	CFI-c+RKL-c+F	6.45e-02	**2.99e-02**
g_6	100	2.33e-4	fRKL-c	5.90e-03	6.08e-03	CFI-c+RKL-c	7.02e-03	6.12e-03	CFI-c+RKL-c+F	5.94e-03	**5.43e-03**
g_7	62	6.00e-5	fRKL-c	5.26e-02	8.88e-03	CFI-c+RKL-c	5.44e-02	6.80e-03	CFI-c+RKL-c+F	5.49e-02	**6.24e-03**
g_{10}	10	4.68e-4	fFKL-c	1.24e-02	**6.01e-03**	CFI-c+FKL-c	1.14e-02	6.80e-03	CFI-c+FKL-c+F	1.61e-02	6.22e-03

IS budget. The sample size required for importance sampling can grow exponentially with proposal–target mismatch [5], motivating our coverage-first initialization and KL refinement. We use $R = 10^5$ samples per evaluation.

Evaluation metrics. For each method, we report the rare-event probability estimate $\hat{I}_R$ together with complementary measures of accuracy, variability, and efficiency. We compare $\hat{I}_R$ against $I_{\text{true}} = \mathbb{P}\left(g(\mathbf{X}) > \gamma\right)$, denoting either an analytically available value or a numerical reference obtained from a large-scale CMC run with 5×10^7 samples. For the ResNet benchmark g_7, the probability is not analytically accessible; following [12], the high-budget CMC estimate is used only as a numerical reference for reporting log-error and is not used by FlowCE. Empirical variance, CoV, and variance reduction remain applicable without an analytic value; log-error provides an additional numerical sanity check. In the main body, we report the logarithmic error $\log\text{-err} = |\log I_{\text{true}} - \log \hat{I}_R|$ and the coefficient of variation (CoV), whereas variance and variance reduction (compared to CMC method) are reported in full tables in Appendix D. For estimators with comparable means, variance scales approximately with the squared CoV. Thus, halving the CoV corresponds to approximately fourfold lower variance, or four times fewer samples for the same relative precision. For CMC and FlowCE methods evaluated under this protocol, we use 20 independent evaluation runs. For each non-CMC FlowCE estimator evaluated under this protocol, the reported estimate $\hat{I}_R$, mean log-error, and CoV are computed using all 20 runs, including runs in which the corresponding CMC estimate is zero. The only metric-specific exclusion concerns variance reduction relative to CMC. If CMC observes no rare event in a given run, its empirical variance is zero and the corresponding variance-reduction ratio is not informative; we therefore average variance reduction only over runs in which the CMC estimate is positive. Accordingly, the "#seeds used" entry reports the number of runs used for the variance-reduction metric and does not apply to $\hat{I}_R$, log-error, or CoV. For example, $4/20$ means that CMC observed at least one rare event in four of the 20 runs, whereas all 20 corresponding FlowCE runs were retained for the other reported metrics.

5.2 Benchmark problems

We benchmark the proposed pipeline consisting of CFI initialization, followed by FKL-c or RKL-c refinement and optional flow refinement (+F), on a collection of rare–event problems of the form $\mathbb{P}\big(g(\mathbf{X}) > \gamma\big)$ with $\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\mu}, \boldsymbol{\Sigma}\right)$. The target probabilities range from 10^{-6} to 10^{-3}; full function definitions and experimental settings are given in Appendix L. We consider two covariance structures: $\boldsymbol{\Sigma} = \mathbf{I}$ for all examples besides g_{10} and $\boldsymbol{\Sigma}_{ii}^{g_{10}} = i$ and $\boldsymbol{\Sigma}_{ij}^{g_{10}} = 0.5\sqrt{ij}$ for $i \neq j$. Two mean vectors $\boldsymbol{\mu}$ are $\boldsymbol{\mu}^0 = (0, \ldots, 0)$ and $\boldsymbol{\mu}^{g_{10}} = (1-d,\, 2-d,\, \ldots,\, d-d)$, respecively.

Low–dimensional benchmarks. For g_1, g_2, and g_3 we consider the two–dimensional case ($d = 2$) with $\gamma = 0$. These benchmarks capture complementary geometric structures, including a two–mode configuration (g_1), a radial benchmark with piecewise boundary (g_2), and an anisotropic problem with a linear tilt (g_3). Function g_1 corresponds to the "Leaf" function from [12], while g_2 and g_3 are structural-reliability benchmarks studied using sequential importance sampling in [21].

Higher–dimensional benchmarks. We additionally consider higher–dimensional problems commonly used in the rare–event literature. These include the 40–dimensional Powell function g_5, a standard nonconvex benchmark considered in [12], and a linear function in $d = 100$ dimensions (g_6), which isolates high-dimensional scaling in a geometrically simple setting and was studied in [21]. For consistency with those papers, we set $\gamma = -403.6636$ for g_5 and $\gamma = 3.5$.

Function g_7 denotes the ResNet-based rare-event model introduced in [12], where rare events arise from stochastic perturbations of network parameters. We consider a ResNet-18 model subjected to a 62-dimensional Gaussian perturbation applied in a layerwise manner, with rare events defined by exceeding a prescribed loss threshold; implementation details are provided in Appendix L. To further assess scalability, we also report a higher-dimensional ResNet variant with $d = 174$ perturbation variables in Appendix E

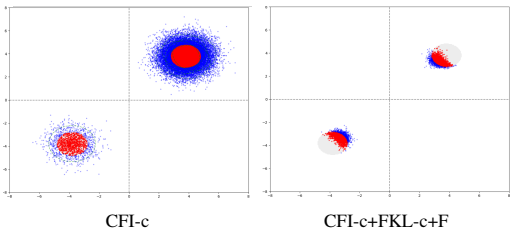

Figure 2: Snapshots for $\mathbb{P}\big(g_1(\mathbf{X}) > 0\big)$ with $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $K = 2$.

Lognormal sum benchmark. For g_{10} we consider $d = 10$ dimensional $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}^{g_{10}}, \boldsymbol{\Sigma}^{g_{10}})$ and estimate $\mathbb{P}\left(\sum_{i=1}^{10} \exp(X_i) > \gamma\right)$, corresponding to the tail probability of a sum of correlated lognormal random variables. This problem arises naturally in financial risk analysis, where $\exp(X_i)$ represents the price of the ith asset in a correlated Black–Scholes model. We consider $\gamma = 40{,}000$. In Appendix D we report also values for $\gamma = 20{,}000$ and $\gamma = 500{,}000$ following the setup from [2].

Ablations and sensitivity. Unless stated otherwise, all experiments use a shared PyTorch implementation and common hyperparameters. The reverse-KL smoothing parameter τ is selected automatically, while β-annealing allows CFI-c to reach the target level before restoring the covariance floor; implementation details are given in Appendix F. The full $K = 16$ comparisons in Tables 6–8 expose the pipeline ablations: CFI provides coverage, the covariance constraint prevents collapse, and subsequent KL or flow refinement may improve efficiency. Tables 15 and 16 show only mild sensitivity to ρ and α on g_5 ($d = 40$). We treat M separately as a simulation budget and keep it fixed across methods and benchmarks.

Visualization ($d = 2$, g_1). Fig. 2 shows representative snapshots for $\mathbb{P}\big(g_1(\mathbf{X}) > 0\big)$ with $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ using Gaussian mixtures with $K = 2$. The panels illustrate the CFI initialization and final CFI-c+FKL-c+F with normalizing flow. Additional visualizations, with alternative mixture sizes and RKL-c trajectories, are provided in Appendix N.

5.3 Summary of main results

For $K = 16$, Table 2 summarizes estimates, log-errors, and CoVs, while Table 3 reports failure counts over 20 evaluations. A run fails if its IS estimate is non-finite or zero because no positive event contribution is observed. These counts measure empirical stability; zero failures for flow-refined proposals do not establish finite variance. For fFKL-c on g_6 and g_7, all 20 estimates are positive and finite but vary substantially across seeds, especially at $K = 16$, thus theorem coverage and numerical success do not ensure small finite-budget variance. Appendix D provides full results for $K = 6$ and $K = 16$, including unconstrained baselines.

Table 3: Failure counts over 20 independent runs. This table measures reliability rather than conditional accuracy on successful runs.

Method	g_1	g_2	g_3	g_5	g_6	g_7	g_{10}
CMC	16	0	0	1	0	0	0
fRKL-c	0	0	0	0	0	0	0
fFKL-c	0	0	0	0	0	0	0
CFI-u	20	20	19	0	20	20	0
CFI-c	0	0	0	0	0	0	0
CFI-c+RKL-c	0	0	0	0	0	0	0
CFI-c+FKL-c	0	0	0	0	0	0	0
CFI-c+RKL-c+F	0	0	0	0	0	0	0
CFI-c+FKL-c+F	0	0	0	0	0	0	0

Unconstrained CFI-u frequently collapses. Likewise, fFKL-u and fRKL-u can perform well but often fail, particularly in high-dimensional or anisotropic settings. In contrast, CFI-c provides a

stable starting point for constrained KL refinement. FKL-c often achieves greater variance reduction, notably on g_1 and g_3, while RKL-c remains competitive.

Flow refinement (+F) often further reduces variance, especially for multimodal and high-dimensional benchmarks g_5 and g_7. These results support constrained CFI initialization followed by KL refinement, with optional flows when empirically beneficial.

Computational cost. Representative timings from our reported setup illustrate the computational cost trade-off. In these measurements, CFI is comparatively inexpensive, whereas full adaptive KL fitting costs substantially more because it repeats proposal optimization across intermediate levels. On g_7, the optional flow stage adds approximately 38–42 seconds and yields a moderate CoV improvement. The implementations were not aggressively optimized, and wall-clock timings are setup- and hardware-dependent; we therefore use them to illustrate trade-offs rather than establish a universal runtime ranking. Appendix D.1 reports detailed training times for g_1, g_6, and g_7.

Defensive importance sampling. We additionally evaluate fRKL-d on g_1 using a $K = 16$ trainable GMM and the full defensive proposal throughout adaptation. Following the guidance of Hesterberg [14, Sec. 6.2] that defensive-mixture results are not unduly sensitive when λ is bounded away from zero and that values between 0.1 and 0.5 are typically used, Table 4 reports the representative choice $\lambda = 0.1$.

Table 4: Comparison with Hesterberg-style defensive importance sampling on g_1 using $K = 16$, $R = 10^5$, and 20 independent evaluation runs. The non-defensive rows are reproduced from Table 6.

Method	Estimate	Var. reduction	log-error	CoV
fRKL-d, $\lambda = 0.1$	4.773e-06	1.73e+03	5.06e-02	7.53e-02
fRKL-c	4.80e-06	2.95e+04	1.59e-02	1.82e-02
CFI-c+RKL-c	**4.79e-06**	**1.87e+05**	**6.26e-03**	**7.24e-03**
CFI-c+RKL-c+F	**4.79e-06**	**7.77e+05**	**3.55e-03**	**3.95e-03**

We also evaluated $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$; the full sensitivity results are reported in Appendix D. All 20 fRKL-d evaluation estimates were nonzero. Relative to fRKL-d, fRKL-c reduces CoV by approximately 76%, increases variance reduction by a factor of 17, and reduces log-error by a factor of 3.2; CFI and optional flow refinement improve these metrics further.

Limitations

The finite-variance guarantee assumes a Gaussian nominal distribution and applies only when the final proposal remains a constrained GMM; it does not extend automatically to the optional flow refinement. For non-Gaussian nominal distributions, analogous guarantees would require a separate tail analysis. CFI and a finite GMM may miss modes or approximate narrow and irregular rare-event regions poorly, while the covariance floor can limit concentration around the rare-event set. Full-covariance mixtures require $O(Kd^2)$ parameters and enough elite samples to estimate them; the covariance floor does not remove this sample complexity, and low-rank or structured covariances are a natural extension. CFI reduces cold-start failures but may require many intermediate levels when the rare-event region is difficult to reach. The method is therefore most useful when cold start and covariance collapse dominate, and less attractive when crude Monte Carlo is adequate or the available budget cannot support full-covariance estimation.

6 Conclusions

We proposed a rare-event importance-sampling framework combining coverage-first initialization, constrained KL refinement, and optional normalizing flows. Theorem 1 guarantees finite variance for final constrained GMM proposals satisfying its strict covariance condition, but not necessarily small finite-budget variance. Across the tested benchmarks, the framework achieves favorable empirical accuracy and stability compared with KL baselines, Subset Simulation, AMS, and selected Safe-ICE comparisons, while unconstrained KL proposals frequently collapse.

Acknowledgments and Disclosure of Funding

We gratefully acknowledge Polish high-performance computing infrastructure PLGrid (HPC Center: WCSS) for providing computer facilities and support within computational grant no. PLG/2026/019627. Rafał Topolnicki acknowledges the support of the project "Center for Trustworthy Artificial Intelligence for Life Sciences" implemented under the International Research Agendas programme of the Foundation for Polish Science co-financed by the European Union under the European Funds for Smart Economy 2021-2027 (FENG). The work of Tomasz Trzciński was supported by the National Science Centre, Poland (NCN), under grant no. 2022/45/B/ST6/02817.

References

[1] A. Agrawal, D. R. Sheldon, and J. Domke. Advances in Black-Box VI: Normalizing Flows, Importance Weighting, and Optimization. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, pages 17358–17369, 2020.

[2] S. Asmussen, J. Blanchet, S. Juneja, and L. Rojas-Nandayapa. Efficient simulation of tail probabilities of sums of correlated lognormals. *Annals of Operations Research*, 189:5–23, 2011.

[3] S.-K. Au and J. L. Beck. Estimation of small failure probabilities in high dimensions by subset simulation. *Probabilistic Engineering Mechanics*, 16(4):263–277, 2001.

[4] F. Cérou, P. Del Moral, T. Furon, and A. Guyader. Sequential Monte Carlo for rare event estimation. *Statistics and Computing*, 22(3):795–808, 2012.

[5] S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

[6] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein. A Tutorial on the Cross-Entropy method. *Annals of Operations Research*, 134(1):19–67, 2005.

[7] B. Delyon and F. Portier. Safe adaptive importance sampling: A mixture approach. *The Annals of Statistics*, 49(2):885–917, 2021.

[8] L. Dinh, D. Krueger, and Y. Bengio. NICE: Non-linear Independent Components Estimation. In *Proceedings of the 3rd International Conference on Learning Representations (ICLR), Workshop Track*, 2015.

[9] L. Dinh, J. Sohl-Dickstein, and S. Bengio. Density estimation using Real NVP. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, 2017.

[10] R. Douc, A. Guillin, J.-M. Marin, and C. P. Robert. Convergence of adaptive mixtures of importance sampling schemes. *The Annals of Statistics*, 35(1):420–448, 2007.

[11] Z. Gao and G. E. Karniadakis. Safe cross-entropy-based importance sampling for rare event simulations. arXiv:2509.07160, 2025.

[12] Z. Gao, D. Zhang, L. Daniel, and D. S. Boning. Rare Event Probability Learning by Normalizing Flows. arXiv:2310.19167, 2023.

[13] L. Gibson, M. Hoerger, and D. P. Kroese. A Flow-Based Generative Model for Rare-Event Simulation. arXiv:2305.07863, 2023.

[14] T. Hesterberg. Weighted Average Importance Sampling and Defensive Mixture Distributions. *Technometrics*, 37(2):185–194, 1995.

[15] T. Homem-de Mello and R. Y. Rubinstein. Estimation of rare event probabilities using cross-entropy. In *Proceedings of the 2002 Winter Simulation Conference (WSC)*, pages 310–319. IEEE, 2002.

[16] G. Jerfel, S. Wang, C. Wong-Fannjiang, K. A. Heller, Y. Ma, and M. I. Jordan. Variational Refinement for Importance Sampling Using the Forward Kullback-Leibler Divergence. In *Proceedings of the 37th Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1819–1829. PMLR, 2021.

[17] X. Li, C. Li, J. Chi, and J. Ouyang. Variance Reduction in Black-box Variational Inference by Adaptive Importance Sampling. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2404–2410, 2018.

[18] P. Lorek and T. Rolski. *Lectures on Monte Carlo Theory*. Springer Nature, 2025.

[19] A. B. Owen. Importance sampling. https://artowen.su.domains/mc/Ch-var-is.pdf, 2013. Lecture notes, Stanford University.

[20] A. B. Owen and Y. Zhou. Safe and Effective Importance Sampling. *Journal of the American Statistical Association*, 95(449):135–143, 2000.

[21] I. Papaioannou, C. Papadimitriou, and D. Straub. Sequential importance sampling for structural reliability analysis. *Structural Safety*, 62:66–75, 2016.

[22] G. Papamakarios, T. Pavlakou, and I. Murray. Masked autoregressive flow for density estimation. In *Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS)*, pages 2338–2347, 2017.

[23] R. Ranganath, S. Gerrish, and D. Blei. Black Box Variational Inference. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 814–822. PMLR, 2014.

[24] B. Recht, R. Roelofs, L. Schmidt, and V. Shankar. Do ImageNet classifiers generalize to ImageNet? In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97 of *Proceedings of Machine Learning Research*, pages 5389–5400. PMLR, 2019.

[25] D. Rezende and S. Mohamed. Variational Inference with Normalizing Flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pages 1530–1538. PMLR, 2015.

[26] D. Ribeiro. Safe-ICE. https://github.com/DiogoRibeiro7/adaptive-importance-sampling/, 2025. GitHub repository.

[27] R. Y. Rubinstein. The Cross-Entropy Method for Combinatorial and Continuous Optimization. *Methodology and Computing in Applied Probability*, 1(2):127–190, 1999.

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

[29] F. J. R. Ruiz, M. K. Titsias, and D. M. Blei. Overdispersed Black-box Variational Inference. In *Proceedings of the 32nd Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 647–656. AUAI Press, 2016.

[30] X. Su and Y. Chen. Variational approximation for importance sampling. *Computational Statistics*, 36(3):1901–1930, 2021.

[31] S. Tong and G. Stadler. Large Deviation Theory-Based Adaptive Importance Sampling for Rare Events in High Dimensions. *SIAM/ASA Journal on Uncertainty Quantification*, 11(3):788–813, 2023.

[32] TorchVision maintainers and contributors. `resnet18` — TorchVision main documentation. https://docs.pytorch.org/vision/main/models/generated/torchvision.models.resnet18. Official documentation for `ResNet18_Weights.IMAGENET1K_V1`; accessed 2026-10-03.

[33] F. Uribe, I. Papaioannou, Y. M. Marzouk, and D. Straub. Cross-Entropy-Based Importance Sampling with Failure-Informed Dimension Reduction for Rare Event Simulation. *SIAM/ASA Journal on Uncertainty Quantification*, 9(2):818–847, 2021.

[34] Z. Wang and J. Song. Cross-entropy-based adaptive importance sampling using von Mises-Fisher mixture for high dimensional reliability analysis. *Structural Safety*, 59:42–52, 2016.

[35] R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction summarize the coverage-first initialization, constrained KL refinement, and optional flow refinement. They accurately describe the theoretical contribution as extending a sufficient finite-variance condition from single-Gaussian to Gaussian mixture proposals and making it enforceable during proposal learning. They also state that the guarantee assumes a Gaussian nominal distribution and applies only when the final proposal is a constrained GMM, and accurately describe the scope of the experimental evaluation.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The dedicated Limitations section discusses the Gaussian nominal-distribution assumption, the scope of the finite-variance guarantee, the exclusion of flow-refined proposals from that guarantee, limitations of finite GMMs, and computational and statistical scaling in high dimensions.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: Proposition 1 and Theorem 1 state their assumptions formally; complete proofs are provided in Appendix B.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper describes the proposed algorithms, benchmark definitions, model architectures, hyperparameters, sampling budgets, and evaluation protocol in Section 5.1 and the appendices.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The repository linked in the abstract provides code, experiment configurations, saved GMM proposals, and instructions for reproducing the main results on the analytic benchmarks. The g_7 benchmark additionally requires external prepared data and pretrained ResNet-18 weights that are not included in the repository.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Detailed descriptions of functions g, dimensions, covariance matrices, and experimental conditions are provided. Hyperparameters such as ρ and α are detailed. We also specify sample sizes, evaluation protocol, and model structure, enabling full understanding of the setting.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For CMC and the proposed estimators evaluated under this protocol, we use 20 independent evaluation runs and report averaged estimates, log-errors, coefficients of variation, and empirical failure counts. Section 5.1 specifies the run accounting and the metric-specific exclusion used only for the CMC-dependent variance-reduction ratio.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Experiments were run on a machine with an AMD Ryzen 5 5500 CPU, 16 GB RAM, and an NVIDIA GeForce RTX 4060 Ti with 16 GB memory. Appendix D.1 reports representative running times and describes their hardware- and implementation-dependent nature.

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: The work studies rare-event estimation using synthetic benchmarks and, for g_7, a pretrained ResNet-18 with ImageNetV2-derived data. It involved no human-subject research or collection of sensitive personal data.

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: This is foundational work on Monte Carlo estimation. The rare-event variables are Gaussian perturbations, and g_7 additionally uses a pretrained ResNet-18 with a fixed ImageNetV2-derived tensor as an external benchmark input. The work does not propose a specific deployment scenario with identifiable direct societal impacts.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This work does not release data or models posing a high risk of misuse.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We credit ImageNetV2 and the pretrained torchvision ResNet-18 weights used for g_7. These external assets remain subject to their upstream terms and are not redistributed with our code.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

Justification: The repository documents the proposed methods, experiment configurations, and 90 saved final GMM proposals, including instructions for evaluating the stored proposals without retraining.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The work did not involve crowdsourcing or human-subject research, and no study participants were recruited. The g_7 benchmark uses pre-existing ImageNetV2-derived inputs rather than data collected from participants in this study.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The work did not involve human-subject research or recruitment of study participants.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not a component of the proposed method or experimental evaluation.

Robust Importance Sampling for Rare Events via Constrained Gaussian Mixtures: Appendix

A Additional background and baseline methods

This appendix summarizes classical Monte Carlo and KL-based baselines for rare-event probability estimation. These methods serve as reference points for the constrained and coverage-first approaches introduced in the main body.

Crude Monte Carlo (CMC). The target probability can be written as

$$I = \mathbb{E}_{\mathbf{X}\sim f}\big[\mathbb{1}(g(\mathbf{X}) > \gamma)\big],$$

and estimated by the empirical fraction of R samples falling in the rare set. The variance of the estimator equals $I(1-I)/R$, which becomes prohibitive in the rare-event regime unless $R = \mathcal{O}(1/I)$.

Importance Sampling and KL fitting. Importance sampling introduces a proposal distribution h and rewrites

$$I = \mathbb{E}_{\mathbf{X}\sim h}\left[\mathbb{1}(g(\mathbf{X}) > \gamma)\,\frac{f(\mathbf{X})}{h(\mathbf{X})}\right].$$

The zero-variance proposal is

$$h^{\mathrm{opt}}(\mathbf{x}) = \frac{\mathbb{1}(g(\mathbf{x}) > \gamma)\, f(\mathbf{x})}{I},$$

which is not directly usable since it depends on the unknown I (except some special cases where MCMC may be employed – but breaks i.i.d. draws and standard unbiased IS). We therefore restrict to a parametric family $h \equiv p_{\boldsymbol{\theta}}$ and fit $p_{\boldsymbol{\theta}}$ to approximate h^{opt} by minimizing a KL divergence. Due to the asymmetry of KL, this leads to two classical baselines depending on its orientation.

Reverse KL (RKL – variational inference). Minimizing $\mathrm{KL}(p_{\boldsymbol{\theta}} \| h^{\mathrm{opt}})$ is ill-posed because h^{opt} contains an indicator. A standard remedy replaces the indicator with a smooth gate

$$s_\tau(g(\mathbf{x}), \gamma) = \sigma\left(\frac{g(\mathbf{x}) - \gamma}{\tau}\right), \qquad \sigma(z) = \frac{1}{1+e^{-z}},$$

leading to the surrogate target

$$h_s^{\mathrm{opt}}(\mathbf{x}) = \frac{s_\tau(g(\mathbf{x}), \gamma)\, f(\mathbf{x})}{I_s}, \qquad I_s = \int s_\tau(g(\mathbf{x}), \gamma) f(\mathbf{x})\, d\mathbf{x}.$$

The resulting reverse-KL objective is

$$\mathbb{E}_{\mathbf{X}\sim p_{\boldsymbol{\theta}}}\Big(\log p_{\boldsymbol{\theta}}(\mathbf{X}) - \log s_\tau(g(\mathbf{X}), \gamma) - \log f(\mathbf{X})\Big). \tag{8}$$

Using samples $\mathbf{X}_i \sim p_{\boldsymbol{\theta}}$ yields the empirical surrogate

$$\hat{L}(\boldsymbol{\theta}) = \frac{1}{M}\sum_{i=1}^{M}\Big(\log p_{\boldsymbol{\theta}}(\mathbf{X}_i) - \log f(\mathbf{X}_i) - \log s_\tau(g(\mathbf{X}_i), \gamma)\Big).$$

Gradients are estimated via the log-derivative (REINFORCE) estimator [35]. Writing $\eta_{\boldsymbol{\theta}}(\mathbf{x}) = \log p_{\boldsymbol{\theta}}(\mathbf{x}) - \log f(\mathbf{x}) - \log s_\tau(g(\mathbf{x}), \gamma)$,

$$\nabla_{\boldsymbol{\theta}} L(\boldsymbol{\theta}) = \mathbb{E}_{p_{\boldsymbol{\theta}}}\big[(1 + \eta_{\boldsymbol{\theta}}(\mathbf{X}) - b)\, \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{X})\big],$$

where b is a control variate used to reduce variance. This yields the practical update summarized in Alg. 2.

Algorithm 2 RKL – Reverse-KL (Variational Inference)

Require: Family $\{p_{\boldsymbol{\theta}}\}$, init $\boldsymbol{\theta}_0$, base f, function g, level γ, batch M, learning rates $\{\kappa_t\}$, schedule $\{\tau_t\}$ (or fixed τ), smoothing α, baseline momentum β

Initialize $\boldsymbol{\theta}'_0 = \boldsymbol{\theta}_0$, $b_0 = 0$, $t = 0$

repeat

  $t \leftarrow t+1$; $\tau \leftarrow \tau_t$; sample $\mathbf{X}_{1:M} \sim p_{\boldsymbol{\theta}'_{t-1}}$

  $z_i \leftarrow (g(\mathbf{X}_i) - \gamma)/\tau$, $\log s_i \leftarrow \log \sigma(z_i)$

  $\eta_i \leftarrow \log p_{\boldsymbol{\theta}'_{t-1}}(\mathbf{X}_i) - \log f(\mathbf{X}_i) - \log s_i$

  Compute

$$\widehat{\nabla}_{\boldsymbol{\theta}} L_t = \frac{1}{M} \sum_{i=1}^{M} (1 + \eta_i - b_{t-1}) \, \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{X}_i) \Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}'_{t-1}}$$

  Update: $\boldsymbol{\theta}'_t \leftarrow \boldsymbol{\theta}'_{t-1} - \kappa_t \, \widehat{\nabla}_{\boldsymbol{\theta}} L_t$; $b_t \leftarrow (1-\beta) b_{t-1} + \beta \, \frac{1}{M} \sum_i \eta_i$

  Smoothing: $\boldsymbol{\theta}_t \leftarrow \alpha \, \boldsymbol{\theta}'_t + (1-\alpha) \, \boldsymbol{\theta}'_{t-1}$

until convergence (or $t = T_{\max}$)

Return $p_{\boldsymbol{\theta}_t}$

Algorithm 3 FKL-u: forward KL, unconstrained (baseline)

Require: Family (GMM), batch M, epochs T, smoothing α

Sample $\mathbf{X}_{1:M} \sim f$

Maximize

$$\hat{J}(\boldsymbol{\theta}) = \frac{1}{M} \sum_{i=1}^{M} \mathbb{1}(g(\mathbf{X}_i) \geq \gamma) \log p_{\boldsymbol{\theta}}(\mathbf{X}_i)$$

by gradient steps for T epochs (with smoothing)

Return $p_{\boldsymbol{\theta}'_T}$

Algorithm 4 fFKL: full adaptive CE (baseline)

Require: Family $\{p_{\boldsymbol{\theta}}\}$ with $\boldsymbol{\theta}_0$, nominal density f, function g, target level γ, quantile ρ, batch M, smoothing $\alpha \in [0, 1)$

$t = 0$

repeat

  $t \leftarrow t+1$; sample $\mathbf{X}_{1:M} \sim p_{\boldsymbol{\theta}_{t-1}}$

  Set $\eta_i = g(\mathbf{X}_i)$, sort: $\eta_{(1)} \leq \cdots \leq \eta_{(M)}$

  $(1-\rho)$ quantile: $\gamma_t \leftarrow \eta_{(\lceil (1-\rho)M \rceil)}$

  Compute likelihood-ratio weights $w_i \leftarrow f(\mathbf{X}_i)/p_{\boldsymbol{\theta}_{t-1}}(\mathbf{X}_i)$

  Weighted CE update:

$$\boldsymbol{\theta}'_t = \arg\max_{\boldsymbol{\theta}} \frac{1}{M} \sum_{i=1}^{M} w_i \, \mathbb{1}\big(g(\mathbf{X}_i) \geq \gamma_t\big) \log p_{\boldsymbol{\theta}}(\mathbf{X}_i)$$

  $\boldsymbol{\theta}_t \leftarrow \alpha \boldsymbol{\theta}'_t + (1-\alpha) \boldsymbol{\theta}_{t-1}$

until $\gamma_t \geq \gamma$

Return $p_{\boldsymbol{\theta}_t}$

Forward KL (FKL – cross-entropy). Minimizing $\mathrm{KL}(h^{\mathrm{opt}} \| p_{\boldsymbol{\theta}})$ yields

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{X} \sim f}\big[\mathbb{1}(g(\mathbf{X}) > \gamma) \, \log p_{\boldsymbol{\theta}}(\mathbf{X})\big]. \tag{9}$$

Using samples $\mathbf{X}_j \sim f$ leads to the surrogate

$$\boldsymbol{\theta}^* \approx \arg\max_{\boldsymbol{\theta}} \frac{1}{M} \sum_{j=1}^{M} \mathbb{1}(g(\mathbf{X}_j) > \gamma) \, \log p_{\boldsymbol{\theta}}(\mathbf{X}_j), \tag{10}$$

corresponding to the standard non-adaptive cross-entropy method. This baseline suffers from severe sample starvation when I is small.

Adaptive forward KL (fFKL – full cross-entropy). The adaptive CE method iteratively increases the threshold γ_t and updates $p_{\boldsymbol{\theta}}$ using importance-weighted samples, ensuring that a fixed fraction of mass lies above the current level. While this approach alleviates cold-start issues, it does not guarantee finite variance.

B Proofs

B.1 Proof of Proposition 1

Proof. Let $A_\gamma = \{g(\mathbf{X}) > \gamma\}$ and $I = \mathbb{P}(A_\gamma)$, where $\mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, denote its density by f. Let the proposal be $p_{\boldsymbol{\theta}} = \mathcal{N}(\boldsymbol{\mu}', \boldsymbol{\Sigma}')$. Then

$$\mathbb{E}_f\left[\mathbf{1}_{A_\gamma} \log p_{\boldsymbol{\theta}}(\mathbf{X})\right] = I\,\mathbb{E}\left[\log p_{\boldsymbol{\theta}}(\mathbf{X}) \mid A_\gamma\right].$$

Since I does not depend on $\boldsymbol{\theta}$, maximizing the left-hand side is equivalent to maximum-likelihood fitting a Gaussian distribution to the conditional law $\mathcal{L}(\mathbf{X} \mid A_\gamma)$. The Gaussian maximum-likelihood solution matches the first two moments of this conditional law. Hence

$$\boldsymbol{\mu}'_{\text{opt}} = \mathbb{E}\left[\mathbf{X} \mid A_\gamma\right], \qquad \boldsymbol{\Sigma}'_{\text{opt}} = \text{Cov}(\mathbf{X} \mid A_\gamma).$$

Consequently, if

$$\lambda_{\min}\left(\text{Cov}(\mathbf{X} \mid A_\gamma)\right) \to 0 \quad \text{as } \gamma \to \infty,$$

then

$$\lambda_{\min}(\boldsymbol{\Sigma}'_{\text{opt}}) \to 0.$$

Thus, unconstrained KL fitting inherits the concentration of the rare-event conditional distribution, which explains the covariance-collapse mechanism. □

As mentioned earlier, such concentration is typical in rare-event regimes: as γ increases, the event $\{g(\mathbf{X}) > \gamma\}$ is dominated by a small subset of the space (e.g., near a boundary or a dominating point), leading to effectively lower-dimensional support of the conditional distribution.

B.2 Proof of Theorem 1

Proof. Let $f(\mathbf{x}) = \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denote the target density, and let $p(\mathbf{x}) = \sum_k \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ be the proposal GMM.

Define the importance weight:

$$w(\mathbf{x}) = \mathbb{1}(g(\mathbf{x}) > \gamma) \cdot \frac{f(\mathbf{x})}{p(\mathbf{x})}.$$

Since $\mathbb{E}_p[w(\mathbf{X})] = I < \infty$, finiteness of the variance is equivalent to finiteness of the second moment:

$$\text{Var}_p[w(\mathbf{X})] < \infty \quad \Longleftrightarrow \quad \mathbb{E}_p[w(\mathbf{X})^2] < \infty.$$

Moreover,

$$\mathbb{E}_p[w(\mathbf{X})^2] = \int \left(\frac{f(\mathbf{x})}{p(\mathbf{x})}\right)^2 \mathbb{1}(g(\mathbf{x}) > \gamma)\, p(\mathbf{x})\, d\mathbf{x} = \int_{\{g(\mathbf{x})>\gamma\}} \frac{f(\mathbf{x})^2}{p(\mathbf{x})}\, d\mathbf{x}.$$

Choose any mixture component k with $\pi_k > 0$. Then

$$p(\mathbf{x}) = \sum_j \pi_j \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j) \geq \pi_k \mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Therefore,

$$\int_{\{g(\mathbf{x})>\gamma\}} \frac{f(\mathbf{x})^2}{p(\mathbf{x})}\, d\mathbf{x} \leq \frac{1}{\pi_k} \int_{\mathbb{R}^d} \frac{f(\mathbf{x})^2}{\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}\, d\mathbf{x}.$$

Up to a positive multiplicative constant independent of $\mathbf{x}$,

$$\frac{f(\mathbf{x})^2}{\mathcal{N}(\mathbf{x} \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \propto \exp\!\left(-(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k)\right).$$

If $\boldsymbol{\Sigma}_k \succ \frac{1}{2}\boldsymbol{\Sigma}$, then $\boldsymbol{\Sigma}_k^{-1} \prec 2\boldsymbol{\Sigma}^{-1}$, and hence

$$\boldsymbol{\Sigma}^{-1} - \frac{1}{2}\boldsymbol{\Sigma}_k^{-1} \succ 0.$$

Consequently,

$$(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}) - \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_k)^\top \boldsymbol{\Sigma}_k^{-1}(\mathbf{x}-\boldsymbol{\mu}_k) \longrightarrow +\infty$$

as $\|\mathbf{x}\| \to \infty$. Thus the ratio above has Gaussian-decaying tails and is integrable over $\mathbb{R}^d$.

Thus $\mathbb{E}_p[w(\mathbf{X})^2] < \infty$, and consequently $\mathrm{Var}_p[w(\mathbf{X})] < \infty$. Since $w(\mathbf{X}_1), \ldots, w(\mathbf{X}_R)$ are i.i.d.,

$$\mathrm{Var}_p(\hat{I}_R^{\mathrm{IS}}) = \frac{1}{R}\mathrm{Var}_p[w(\mathbf{X})] < \infty.$$

□

C Motivating example: full derivation

We consider the rare-event probability

$$I = \mathbb{P}(X > 4), \qquad X \sim \mathcal{N}(0,1).$$

For crude Monte Carlo (CMC), the estimator variance equals

$$\mathrm{Var}(\hat{I}^{\mathrm{CMC}}) = \frac{I(1-I)}{R},$$

which is prohibitive in rare-event regimes unless R is of order $1/I$. We compare the classical forward-KL approach (FKL, equivalently cross-entropy) and the reverse-KL approach (RKL, variational inference with a smoothed gate s_τ) for two Gaussian proposal families:

$$p_{\boldsymbol{\theta}} = \mathcal{N}(\mu, 1) \qquad \text{and} \qquad p_{\boldsymbol{\theta}} = \mathcal{N}(\mu, \sigma^2).$$

Fixed variance: $p_{\boldsymbol{\theta}} = \mathcal{N}(\mu, 1)$.

- **FKL (CE).** From the forward-KL objective (9), the optimal mean satisfies

$$\mu^*_{\mathrm{FKL}} = \operatorname*{argmin}_{\mu} \int_4^\infty (x-\mu)^2 e^{-x^2/2}\, dx = 4.2256. \tag{11}$$

  The resulting importance sampling variance equals 4.5162×10^{-9}.

- **RKL (smoothed).** For the reverse-KL objective with smoothed gate

$$s_\tau(x, 4) = \sigma((x-4)/\tau),$$

  the optimization reduces to minimizing

$$r(\mu;\tau) = \tfrac{1}{2}\mu^2 - \mathbb{E}\big[\log \sigma((X-4)/\tau)\big].$$

  The stationarity condition yields

$$\mu^*_{\mathrm{RKL}} = \tau^{-1}\mathbb{E}\big[1 - \sigma((X-4)/\tau)\big].$$

  For $\tau = 0.1$, this gives $\mu^*_{\mathrm{RKL}} = 4.2036$, with IS variance 4.5045×10^{-9}.

Thus, with fixed variance, both FKL and RKL lead to well-behaved proposals and comparable variance reduction over crude Monte Carlo.

Free mean and variance: $p_{\boldsymbol{\theta}} = \mathcal{N}(\mu, \sigma^2)$. The behavior changes qualitatively once the proposal variance is also optimized.

- **FKL (CE).** Optimizing (9) over (μ, σ^2) yields

$$(\mu^*_{\mathrm{FKL}}, \sigma^{2*}_{\mathrm{FKL}}) = (4.2256,\ 0.046671).$$

  The second moment of the IS estimator is finite only if

$$\mathbb{E}_p\left[\mathbf{1}(X>4)\frac{f(X)^2}{p(X)^2}\right] = \int_4^\infty \frac{f(x)^2}{p(x)}\, dx < \infty.$$

For a Gaussian proposal $p = \mathcal{N}(\mu, \sigma^2)$, a direct calculation gives

$$\int_4^\infty \frac{f(x)^2}{p(x)}\,dx = \frac{\sigma}{\sqrt{2\pi}} \exp\Big(\frac{\mu^2}{2\sigma^2}\Big) \int_4^\infty \exp\Big(\Big(\tfrac{1}{2\sigma^2} - 1\Big)x^2 - \mu x\Big)\,dx.$$

This integral diverges whenever $\sigma^2 < \frac{1}{2}$ (and also for $\sigma^2 = \frac{1}{2}$ if $\mu \le 0$). Thus, the threshold $1/2$ is sharp in this one-dimensional example, up to the boundary case $\sigma^2 = 1/2$, whose behavior depends on the proposal mean and the geometry of the event. For multivariate GMM proposals, however, the componentwise condition in Theorem 1 is deliberately conservative rather than necessary. Since $\sigma^{2*}_{\mathrm{FKL}} = 0.046671 < \frac{1}{2}$, the resulting IS estimator has infinite variance, despite optimizing the forward-KL objective.

- **RKL (smoothed).** For the reverse-KL objective, we minimize

$$r(\mu, \sigma^2; \tau) = \underbrace{\tfrac{1}{2}\Big(\mu^2 + \sigma^2 - \log\sigma^2 - 1\Big)}_{\mathrm{KL}(\mathcal{N}(\mu,\sigma^2)\,\|\,\mathcal{N}(0,1))} - \mathbb{E}_{X\sim\mathcal{N}(\mu,\sigma^2)}\big[\log\sigma((X-4)/\tau)\big].$$

The first-order conditions are

$$\partial_\mu r = \mu - \tau^{-1}\mathbb{E}\Big[1 - \sigma\Big(\tfrac{\mu+\sigma\varepsilon-4}{\tau}\Big)\Big] = 0,$$

$$\partial_{\sigma^2} r = \tfrac{1}{2}\Big(1 - \tfrac{1}{\sigma^2}\Big) - \frac{1}{2\tau\sqrt{\sigma^2}}\mathbb{E}\Big[\varepsilon\Big(1 - \sigma\Big(\tfrac{\mu+\sigma\varepsilon-4}{\tau}\Big)\Big)\Big] = 0,$$

with $\varepsilon \sim \mathcal{N}(0,1)$. For small τ, solutions with $\sigma^{2*}_{\mathrm{RKL}} < \frac{1}{2}$ may occur; for example, with $\tau = 0.1$ one obtains

$$(\mu^*_{\mathrm{RKL}}, \sigma^{2*}_{\mathrm{RKL}}) = (3.2782,\ 0.4306),$$

again implying infinite IS variance.

Numerical summary. Table 5 summarizes the optimal proposals and the corresponding IS variances for the four cases discussed above.

Table 5: Motivating example: optimal proposals obtained by FKL and RKL and the resulting IS variance for $\mathbb{P}(X > 4)$.

Model		Optimum	IS variance
$\mathcal{N}(\mu, 1)$	FKL	$\mu^* = 4.2256$	$4.5162 \cdot 10^{-9}$
$\mathcal{N}(\mu, 1)$	RKL ($\tau = 0.1$)	$\mu^* = 4.2036$	$4.5045 \cdot 10^{-9}$
$\mathcal{N}(\mu, \sigma^2)$	FKL	$(4.2256,\ 0.0467)$	∞
$\mathcal{N}(\mu, \sigma^2)$	RKL ($\tau = 0.1$)	$(3.2782,\ 0.4306)$	∞

Interpretation. The example shows that KL fitting can concentrate the proposal too aggressively around the rare-event conditional distribution. Such concentration may improve the KL objective, but it can destroy the tail overlap required for stable importance sampling. In particular, when variance parameters are unconstrained, both FKL and RKL can drive σ^2 below the critical threshold $\frac{1}{2}$, leading to infinite estimator variance.

This mechanism is not specific to the one-dimensional Gaussian example. In higher dimensions and for Gaussian mixture proposals, individual components may collapse onto rare-event modes, producing covariance matrices that are too small in some directions. This is the covariance-collapse mechanism formalized in Proposition 1. The constrained proposal family rules out this particular infinite-variance mechanism by enforcing the covariance floor from Theorem 1, namely

$$\boldsymbol{\Sigma}_k \succ \tfrac{1}{2}\,\boldsymbol{\Sigma},$$

where $\boldsymbol{\Sigma}$ is the covariance matrix of the original sampling distribution.

Empirical GMM illustration. We also test the related two-sided rare-event problem

$$I = \mathbb{P}(|X| > 4), \qquad X \sim \mathcal{N}(0, 1),$$

using a Gaussian mixture proposal with $K = 20$ components. This setting is shown in Fig. 1, and the corresponding quantitative results are reported in Table 1. Consistent with the derivation above, the unconstrained CFI-u proposal exhibits severe variance pathologies, whereas the constrained CFI-c proposal remains stable and accurate.

Summary. This example illustrates two key points. First, minimizing reverse or forward KL alone does not guarantee a finite-variance IS estimator. Second, the issue is not specific to the one-dimensional Gaussian case: the same covariance-collapse mechanism appears in Gaussian mixtures and higher-dimensional settings. This motivates the covariance constraints imposed in the main body.

D Complete numerical results

This section reports the complete numerical results for all benchmark problems and model variants. The main paper reports compact summaries to preserve the flow of the experimental section; the tables below provide the full per-benchmark numerical results. Compared to the main body, the tables below additionally include the unconstrained baseline methods fRKL-u and fFKL-u, which are omitted from the primary comparison due to their lack of robustness.

In all tables below, FG marks methods whose final proposal is a constrained GMM covered by the finite-variance guarantee of Theorem 1.

Reference values. For selected benchmarks previously studied in the literature, we additionally report reference values taken from the corresponding publications (denoted as "Ref. results$^{(*)}$" or "CEVE$^{(*)}$" in the tables, with sources specified in the captions). These values are used solely as external calibration points and are not meant to enable direct efficiency comparisons, as experimental protocols (e.g., IS sample sizes, number of independent runs, or variance reporting conventions) often differ across studies and are not always fully specified. In contrast, CMC and FlowCE results evaluated under our protocol use a fixed IS budget and 20 independent evaluation runs, allowing for a consistent assessment of estimator stability and variability. As specified in Section 5.1, only the CMC-dependent variance-reduction ratio excludes runs with zero CMC estimates; the corresponding FlowCE estimates, mean log-errors, and CoVs use all 20 runs.

Results for $K = 6$. For $K = 6$, we report full results for all benchmarks:

- g_1, g_2, g_3: Table 9,
- g_5, g_6, g_7: Table 10,
- g_{10}: Table 11.

These tables provide a comprehensive view of estimator performance, including methods that occasionally fail numerically or produce degenerate finite-sample estimates. Such observed failures are distinct from a theoretically established infinite-variance estimator.

Results for $K = 16$. For $K = 16$, the appendix tables provide the detailed results underlying the compact summaries in the main body and additionally include the unconstrained baselines:

- g_1, g_2, g_3: Table 6,
- g_5, g_6, g_7: Table 7,
- g_{10}: Table 8.

Finite-budget behavior of fFKL-c on g_6 and g_7. For each fFKL-c setting with $K \in \{6, 16\}$, the learned final proposal uses training seed 100; the reported metrics average 20 independent ordinary, unnormalized IS evaluations with seeds $100, 200, \ldots, 2000$ and $R = 10^5$ each, without pooling samples across runs. All 20 evaluations completed with positive event counts, but this does not imply small finite-sample variance. In particular, the $K = 16$ estimates range from 1.9560×10^{-9} to 7.7778×10^{-4} for g_6 and from 3.7277×10^{-6} to 1.0757×10^{-3} for g_7.

Ranking and highlighting. In contrast to the compact main-body comparison, which reports selected baselines and CFI-based methods, the appendix tables rank *all* methods with reportable, nondegenerate numerical estimates, including fRKL-u and fFKL-u. As a result, the highlighted entries (best, second-best, and third-best rows) may differ from those shown in the main body. This presentation is intended to provide a complete quantitative record, while the main paper focuses on the robustness–efficiency trade-off.

Table 6: Estimating $\mathbb{P}\left(g_i(\mathbf{X}) > \gamma\right), i = 1, 2, 3, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}^{(A)}, \boldsymbol{\Sigma}^{(A)})$, $K = 16$, $R = 10^5$, $d = 2$.
Ref. results$^{(*)}$: results from [12] for g_1 and [21] for g_2, g_3.

$I_{\text{true}} =$		g_1 [12] 4.7934e-06					g_2 [21] 2.22668e-03					g_3 [21] 4.2073e-03				
	FG	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$
CMC		4.50e-06	1	1.43e+00	*3.92e-03*	1.43e+00	2.25e-03	1	4.14e-02	*2.46e-03*	4.14e-02	4.15e-03	1	4.25e-02	***3.31e-03***	4.25e-02
Ablation																
CFI-u		0	✗	✗	✗	✗	0	✗	✗	✗	✗	3.90e-06	**4.57e+04**	7.02e+00	1.15e-01	1.53e-04
fRKL-c	✓	4.80e-06	2.95e+04	1.59e-02	1.82e-02	5.38e-07	2.22e-03	6.12e+01	5.18e-03	8.61e-03	8.46e-05	4.21e-03	4.89e+01	5.98e-03	6.91e-03	1.22e-04
fFKL-c	✓	4.80e-06	1.41e+05	6.65e-03	8.31e-03	4.71e-08	2.22e-03	1.89e+02	5.52e-03	4.91e-03	2.93e-05	4.21e-03	1.20e+02	4.90e-03	4.41e-03	4.09e-05
Baseline																
AMS		5.47e-06	6.39e-01	1.31e-01	1.34e+00	2.06e-01	2.24e-03	5.47e-01	6.75e-03	7.41e-02	1.23e-02	4.23e-03	3.73e-01	4.29e-03	6.77e-02	1.15e-02
Subset		3.41e-06	1.48e+00	3.41e-01	1.45e+00	2.31e-01	2.06e-03	7.81e-01	7.79e-02	8.42e-02	9.98e-02	3.80e-03	7.93e-01	1.02e-01	5.41e-02	1.28e-01
Our: CFI-based																
CFI-c	✓	4.82e-06	9.85e+04	7.58e-03	9.87e-03	7.69e-08	2.23e-03	5.57e+01	6.93e-03	9.00e-03	1.24e-04	4.21e-03	5.76e+01	*3.77e-03*	6.37e-03	6.55e-05
CFI-c+RKL-c	✓	4.79e-06	*1.87e+05*	6.26e-03	7.24e-03	3.35e-08	2.22e-03	1.44e+02	5.77e-03	5.62e-03	4.00e-05	**4.20e-03**	1.05e+02	**1.87e-03**	4.71e-03	***1.78e-05***
CFI-c+FKL-c	✓	*4.80e-06*	1.86e+05	*6.23e-03*	7.24e-03	*3.34e-08*	*2.22e-03*	*1.93e+02*	*3.39e-03*	*4.85e-03*	*1.76e-05*	*4.21e-03*	*1.23e+02*	*3.25e-03*	4.35e-03	*2.63e-05*
CFI-c+RKL-c+F		***4.79e-06***	**7.77e+05**	***3.55e-03***	***3.95e-03***	***4.57e-09***	***2.22e-03***	**1.29e+03**	***2.87e-03***	**1.86e-03**	**2.22e-06**	4.20e-03	7.12e+01	3.51e-03	*4.15e-03*	4.93e-05
CFI-c+FKL-c+F		**4.79e-06**	***7.51e+05***	**3.16e-03**	**3.92e-03**	**4.21e-09**	**2.22e-03**	***6.54e+02***	**2.22e-03**	***2.46e-03***	***3.39e-06***	***4.20e-03***	***2.07e+02***	***2.80e-03***	**3.31e-03**	**1.35e-05**
#seeds used				4/20					20/20					20/20		
Ref. results$^{(*)}$		–	–	1.10e-01	–	–	2.23e-03	–	–	1.10e-01	–	4.20e-03	–	–	6.00e-02	–

Table 7: Estimating $\mathbb{P}\left(g_i(\mathbf{X}) > \gamma\right), i = 5, 6, 7$, with $K = 16$ and $R = 10^5$.
Ref. results$^{(*)}$: results from [12] for g_5, g_7 and [21] for g_6.

$I_{\text{true}} =$		g_5 [12] 3.1500e-05					g_6 [21] 2.3262e-04					g_7 [12] 6.00e-05				
	FG	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$
CMC		3.30e-05	1	4.59e-01	1.17e-01	4.59e-01	2.27e-04	1	1.67e-01	1.80e-01	1.67e-01	5.00e-05	1	3.35e-01	1.65e-02	3.35e-01
Ablation																
fRKL-c	✓	*3.39e-05*	***2.97e+02***	*7.31e-02*	***3.19e-02***	***2.46e-04***	**2.32e-04**	*1.14e+03*	**5.90e-03**	***6.08e-03***	*5.20e-06*	5.69e-05	*1.96e+03*	5.26e-02	*8.88e-03*	*2.69e-05*
fFKL-c	✓	3.39e-05	*2.82e+02*	7.32e-02	*3.27e-02*	*2.60e-04*	3.90e-05	**3.82e+08**	8.22e+00	6.36e-01	**2.15e-08**	1.19e-04	4.66e+01	1.60e+00	5.96e-01	3.42e-02
CFI-u		5.24e-07	1.96e+02	5.33e+00	4.54e-01	2.72e-02	0	✗	✗	✗	✗	0	✗	✗	✗	✗
Baseline																
AMS		2.75e-05	9.26e-01	1.34e-01	9.51e-01	1.45e-01	2.30e-04	6.96e-02	9.20e-03	6.61e-01	1.32e-01	6.48e-05	1.50e-01	7.73e-02	8.15e-01	5.14e-01
Subset		3.71e-05	7.07e-01	1.63e-01	5.93e-01	2.31e-01	2.14e-04	1.78e-01	8.41e-02	3.92e-01	4.73e-01	**6.22e-05**	5.44e-01	**3.57e-02**	4.93e-01	6.56e-02
Our: CFI-based																
CFI-c	✓	3.27e-05	4.24e+01	8.07e-02	8.89e-02	1.90e-03	2.34e-04	8.33e+00	5.09e-02	5.81e-02	6.11e-03	***5.72e-05***	2.28e+02	***5.00e-02***	2.42e-02	2.19e-04
CFI-c+RKL-c	✓	***3.39e-05***	1.49e+02	***7.28e-02***	4.30e-02	4.88e-04	*2.32e-04*	1.12e+03	*7.02e-03*	*6.12e-03*	6.26e-06	5.68e-05	***3.35e+03***	5.44e-02	***6.80e-03***	***1.62e-05***
CFI-c+FKL-c	✓	3.40e-05	7.60e+00	1.48e-01	1.80e-01	1.94e-02	2.34e-04	4.14e-01	1.77e-01	1.88e-01	4.29e-01	*5.70e-05*	3.26e+02	*5.09e-02*	2.14e-02	1.56e-04
CFI-c+RKL-c+F		**3.36e-05**	**3.15e+02**	**6.45e-02**	**2.99e-02**	**2.05e-04**	***2.33e-04***	***1.42e+03***	***5.94e-03***	**5.43e-03**	***4.19e-06***	5.68e-05	**3.97e+03**	5.49e-02	**6.24e-03**	**1.38e-05**
CFI-c+FKL-c+F		3.22e-05	1.32e+01	9.17e-02	1.17e-01	6.96e-03	2.28e-04	5.48e-01	1.79e-01	1.80e-01	3.26e-01	5.65e-05	5.72e+02	6.11e-02	1.65e-02	1.07e-04
#seeds used				19/20					20/20					20/20		
Ref. results$^{(*)}$		–	–	3.80e-01	–	–	2.33e-04	–	–	2.80e-01	–	–	–	6.20e-01	–	–

Table 8: Estimating $\mathbb{P}\left(g_{10}(\mathbf{X}) > \gamma\right)$ for different thresholds, with $K = 16$ and $R = 10^5$.
CEVE$^{(*)}$: results from [2].

$I_{\text{true}} =$		$\gamma = 20{,}000$ 1.03526e-03					$\gamma = 40{,}000$ 4.683e-04					$\gamma = 500{,}000$ 1.77400e-05				
	FG	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$
CMC		1.03e-03	1	7.22e-02	*6.22e-03*	7.22e-02	4.69e-04	1	1.12e-01	*6.22e-03*	1.12e-01	2.30e-05	1	4.28e-01	***1.68e-02***	4.28e-01
Ablation																
CFI-u		0	✗	✗	✗	✗	0	✗	✗	✗	✗	0	✗	✗	✗	✗
fRKL-c	✓	1.03e-03	2.76e+01	1.43e-02	1.88e-02	5.18e-04	4.61e-04	4.89e+01	2.27e-02	2.12e-02	4.64e-04	1.78e-05	***1.25e+03***	2.94e-02	2.53e-02	*2.34e-05*
fFKL-c	✓	**1.02e-03**	**3.02e+02**	**1.09e-02**	**5.70e-03**	**3.62e-05**	***4.63e-04***	**6.07e+02**	***1.24e-02***	**6.01e-03**	**2.04e-05**	**1.79e-05**	**1.30e+04**	**1.19e-02**	**7.81e-03**	**9.15e-07**
Baseline																
AMS		*1.02e-03*	1.72e-01	*1.17e-02*	2.02e-01	6.82e-02	**4.71e-04**	2.49e-01	**5.48e-03**	3.31e-01	2.20e-02	1.70e-05	8.09e-01	4.15e-02	1.00e+00	5.12e-02
Subset		9.72e-04	9.51e-01	6.29e-02	1.08e-01	6.62e-02	4.58e-04	1.78e+00	2.33e-02	1.21e-01	1.31e-02	1.48e-05	2.96e+00	1.79e-01	5.47e-01	6.06e-02
Our: CFI-based																
CFI-c	✓	1.02e-03	5.28e+01	1.26e-02	1.33e-02	2.39e-04	4.65e-04	8.76e+01	1.35e-02	1.37e-02	1.54e-04	1.79e-05	2.16e+02	3.04e-02	3.25e-02	1.41e-04
CFI-c+RKL-c	✓	1.02e-03	7.63e+00	2.87e-02	3.12e-02	3.77e-03	4.66e-04	5.33e+01	1.39e-02	1.98e-02	2.61e-04	1.86e-05	6.26e+01	6.11e-02	6.36e-02	9.76e-04
CFI-c+FKL-c	✓	1.02e-03	*1.66e+02*	1.24e-02	7.46e-03	7.47e-05	*4.63e-04*	***4.68e+02***	1.14e-02	6.80e-03	***2.44e-05***	***1.78e-05***	1.19e+03	***1.31e-02***	*1.80e-02*	***1.10e-05***
CFI-c+RKL-c+F		1.01e-03	1.79e+01	2.52e-02	2.26e-02	1.41e-03	4.59e-04	1.39e+02	2.25e-02	*1.14e-02*	1.62e-04	1.76e-05	1.15e+02	5.15e-02	3.22e-02	4.48e-04
CFI-c+FKL-c+F		***1.02e-03***	***1.90e+02***	***1.16e-02***	***6.22e-03***	***6.10e-05***	4.61e-04	*4.02e+02*	1.61e-02	***6.22e-03***	*3.99e-05*	*1.78e-05*	*3.69e+03*	*1.46e-02*	**1.68e-02**	**3.96e-06**
#seeds used				20/20					20/20					18/20		
CEVE$^{(*)}$		1.02e-03	–	–	1.28e-01	–	4.63e-04	–	–	8.60e-02	–	1.79e-05	–	–	4.46e-02	–

Table 9: Estimating $\mathbb{P}\left(g_i(\mathbf{X}) > \gamma\right), i = 1, 2, 3, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}^{(A)}, \boldsymbol{\Sigma}^{(A)})$, $K = 6$, $R = 10^5$, $d = 2$.
Ref. results$^{(*)}$: results from [12] for g_1 and [21] for g_2, g_3.

$I_{\text{true}} =$		g_1 [12] 4.7934e-06					g_2 [21] 2.22668e-03					g_3 [21] 4.2073e-03				
	FG	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$
CMC		4.50e-06	1	1.43e+00	3.92e-03	1.43e+00	2.25e-03	1	4.14e-02	2.46e-03	4.14e-02	4.15e-03	1	4.25e-02	3.31e-03	4.25e-02
Baselines (no CFI)																
fRKL-u		4.74e-06	1.96e+03	5.56e-02	6.95e-02	2.83e-05	2.22e-03	6.43e+01	7.68e-03	8.40e-03	1.19e-04	4.21e-03	4.30e+01	5.73e-03	7.02e-03	1.33e-04
fFKL-u		*4.80e-06*	*3.15e+05*	*5.65e-03*	*5.56e-03*	*1.79e-08*	2.21e-03	2.69e+02	8.05e-03	4.10e-03	3.00e-05	**4.21e-03**	**5.07e+02**	**1.38e-03**	**2.14e-03**	**2.72e-06**
Ablation																
CFI-u		0	✗	✗	✗	✗	0	✗	✗	✗	✗	3.90e-06	4.57e+04	7.02e+00	1.15e-01	1.53e-04
fRKL-c	✓	4.80e-06	2.95e+04	1.59e-02	1.82e-02	5.38e-07	2.22e-03	6.12e+01	5.18e-03	8.61e-03	8.46e-05	4.21e-03	4.89e+01	5.98e-03	6.91e-03	1.22e-04
fFKL-c	✓	4.80e-06	1.41e+05	6.65e-03	8.31e-03	4.71e-08	2.22e-03	1.89e+02	5.52e-03	4.91e-03	2.93e-05	4.21e-03	1.20e+02	4.90e-03	4.41e-03	4.09e-05
Our: CFI-based																
CFI-c	✓	4.82e-06	9.85e+04	7.58e-03	9.87e-03	7.69e-08	2.23e-03	5.57e+01	6.93e-03	9.00e-03	1.24e-04	4.21e-03	5.76e+01	3.77e-03	6.37e-03	6.55e-05
CFI-c+RKL-c	✓	4.79e-06	1.87e+05	6.26e-03	7.24e-03	3.35e-08	2.22e-03	1.44e+02	5.77e-03	5.62e-03	4.00e-05	*4.20e-03*	*1.05e+02*	*1.87e-03*	*4.71e-03*	*1.78e-05*
CFI-c+FKL-c	✓	4.80e-06	1.86e+05	6.23e-03	7.24e-03	3.34e-08	*2.22e-03*	*1.93e+02*	*3.39e-03*	*4.85e-03*	*1.76e-05*	4.21e-03	1.23e+02	3.25e-03	4.35e-03	2.63e-05
CFI-c+RKL-c+F		***4.79e-06***	***7.77e+05***	***3.55e-03***	***3.95e-03***	***4.57e-09***	**2.22e-03**	**1.29e+03**	**2.87e-03**	**1.86e-03**	**2.22e-06**	4.20e-03	7.12e+01	3.51e-03	4.15e-03	4.93e-05
CFI-c+FKL-c+F		**4.79e-06**	**7.51e+05**	**3.16e-03**	**3.92e-03**	**4.21e-09**	***2.22e-03***	***6.54e+02***	***2.22e-03***	***2.46e-03***	***3.39e-06***	***4.20e-03***	***2.07e+02***	***2.80e-03***	***3.31e-03***	***1.35e-05***
#seeds used				4/20					20/20					20/20		
Ref. results$^{(*)}$		–	–	1.10e-01	–	–	2.23e-03	–	–	1.10e-01	–	4.20e-03	–	–	6.00e-02	–

Table 10: Estimating $\mathbb{P}\left(g_i(\mathbf{X}) > \gamma\right), i = 5, 6, 7, \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}^{(A)}, \boldsymbol{\Sigma}^{(A)})$, $K = 6$, $R = 10^5$.
Ref. results$^{(*)}$: results from [12] for g_5, g_7 and [21] for g_6.

dimension $d =$ $I_{\text{true}} =$		g_5 [12] 40 3.1500e-05					g_6 [21] 100 2.3262e-04					g_7 (ResNet) [12] 62 6.00e-05				
	FG	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$
CMC		3.30e-05	1	4.59e-01	1.17e-01	4.59e-01	2.27e-04	1	1.67e-01	1.80e-01	1.67e-01	5.00e-05	1	3.35e-01	1.65e-02	3.35e-01
Baselines (no CFI)																
fRKL-u		3.63e-05	1.25e+00	2.31e-01	2.41e-01	1.84e-01	**2.33e-04**	***5.87e+03***	**2.12e-03**	**2.67e-03**	**3.61e-07**	*5.68e-05*	*2.06e+03*	*5.53e-02*	*8.67e-03*	*2.68e-05*
fFKL-u		**3.39e-05**	**4.03e+03**	**7.37e-02**	**8.64e-03**	**1.83e-05**	0	✗	✗	✗	✗	0	✗	✗	✗	✗
Ablation																
CFI-u		5.24e-07	1.96e+02	5.33e+00	4.54e-01	2.72e-02	0	✗	✗	✗	✗	0	✗	✗	✗	✗
fRKL-c	✓	*3.39e-05*	*2.97e+02*	*7.31e-02*	*3.19e-02*	*2.46e-04*	*2.32e-04*	1.14e+03	*5.90e-03*	*6.08e-03*	*5.20e-06*	5.69e-05	1.96e+03	5.26e-02	8.88e-03	2.69e-05
fFKL-c	✓	3.39e-05	2.82e+02	7.32e-02	3.27e-02	2.60e-04	7.50e-06	**1.16e+04**	4.17e+00	5.76e-01	3.59e-04	4.81e-05	1.70e+01	5.33e-01	2.54e-01	3.13e-02
Our: CFI-based																
CFI-c	✓	3.27e-05	4.24e+01	8.07e-02	8.89e-02	1.90e-03	2.34e-04	8.33e+00	5.09e-02	5.81e-02	6.11e-03	5.72e-05	2.28e+02	5.00e-02	2.42e-02	2.19e-04
CFI-c+RKL-c	✓	3.39e-05	1.49e+02	7.28e-02	4.30e-02	4.88e-04	2.32e-04	1.12e+03	7.02e-03	6.12e-03	6.26e-06	***5.68e-05***	***3.35e+03***	***5.44e-02***	***6.80e-03***	***1.62e-05***
CFI-c+FKL-c	✓	3.40e-05	7.60e+00	1.48e-01	1.80e-01	1.94e-02	2.34e-04	4.14e-01	1.77e-01	1.88e-01	4.29e-01	5.70e-05	3.26e+02	5.09e-02	2.14e-02	1.56e-04
CFI-c+RKL-c+F		***3.36e-05***	***3.15e+02***	***6.45e-02***	***2.99e-02***	***2.05e-04***	***2.33e-04***	*1.42e+03*	***5.94e-03***	***5.43e-03***	***4.19e-06***	**5.68e-05**	**3.97e+03**	**5.49e-02**	**6.24e-03**	**1.38e-05**
CFI-c+FKL-c+F		3.22e-05	1.32e+01	9.17e-02	1.17e-01	6.96e-03	2.28e-04	5.48e-01	1.79e-01	1.80e-01	3.26e-01	5.65e-05	5.72e+02	6.11e-02	1.65e-02	1.07e-04
#seeds used				19/20					20/20					20/20		
Ref. results$^{(*)}$		–	–	3.80e-01	–	–	2.33e-04	–	–	2.80e-01	–	–	–	6.20e-01	–	–

Table 11: Estimating $\mathbb{P}\left(g_{10}(\mathbf{X}) > \gamma\right), \mathbf{X} \sim \mathcal{N}(\boldsymbol{\mu}^{(B)}, \boldsymbol{\Sigma}^{(B)})$, $K = 6$, $R = 10^5$.
CEVE$^{(*)}$: results from [2].

$\gamma =$ $I_{\text{true}} =$		20,000 1.03526e-03					40,000 4.683e-04					500,000 1.77400e-05				
	FG	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$	$\hat{I}_R$	Var red.	log-err	CoV	$\frac{\text{log-err}}{\text{Var red.}}$
CMC		1.03e-03	1	7.22e-02	6.97e-03	7.22e-02	4.69e-04	1	1.12e-01	7.98e-03	1.12e-01	2.30e-05	1	4.28e-01	1.36e-02	4.28e-01
Ablation																
CFI-u		4.37e-05	1.60e+01	3.26e+00	3.65e-01	2.04e-01	3.40e-06	1.01e+03	5.10e+00	3.30e-01	5.05e-03	0	✗	✗	✗	✗
Our: CFI-based																
fRKL-c	✓	1.02e-03	2.48e+01	1.66e-02	2.00e-02	6.70e-04	4.59e-04	5.17e+01	2.46e-02	2.07e-02	4.76e-04	1.78e-05	7.70e+02	2.97e-02	3.24e-02	3.86e-05
fFKL-c	✓	**1.02e-03**	**2.29e+02**	**1.04e-02**	**6.55e-03**	**4.53e-05**	***4.62e-04***	***4.20e+02***	***1.36e-02***	***7.23e-03***	***3.24e-05***	**1.79e-05**	**7.71e+03**	**9.87e-03**	**1.01e-02**	**1.28e-06**
CFI-c	✓	1.02e-03	3.51e+01	2.06e-02	1.59e-02	5.85e-04	4.62e-04	8.88e+01	1.69e-02	1.55e-02	1.90e-04	***1.80e-05***	***3.09e+03***	***1.62e-02***	***1.56e-02***	***5.24e-06***
CFI-c+RKL-c	✓	9.98e-04	1.09e+00	7.10e-02	4.50e-02	6.53e-02	4.67e-04	9.51e+00	3.31e-02	3.88e-02	3.48e-03	1.80e-05	9.39e+02	2.65e-02	2.75e-02	2.83e-05
CFI-c+FKL-c	✓	***1.02e-03***	***1.52e+02***	***1.31e-02***	***8.03e-03***	***8.64e-05***	**4.63e-04**	**4.08e+02**	**1.12e-02**	**7.30e-03**	**2.75e-05**	1.81e-05	6.96e+02	2.03e-02	1.83e-02	2.92e-05
CFI-c+RKL-c+F		9.90e-04	2.70e+01	4.49e-02	1.90e-02	1.66e-03	4.56e-04	3.84e+00	6.28e-02	3.81e-02	1.63e-02	*1.80e-05*	*2.81e+03*	*1.47e-02*	*1.25e-02*	*5.24e-06*
CFI-c+FKL-c+F		*1.02e-03*	*1.61e+02*	*1.40e-02*	*6.97e-03*	*8.74e-05*	*4.63e-04*	*2.47e+02*	*1.22e-02*	*7.98e-03*	*4.92e-05*	1.79e-05	2.35e+03	1.37e-02	1.36e-02	5.83e-06
#seeds used				20/20					20/20					18/20		
CEVE$^{(*)}$		1.02e-03	–	–	1.28e-01	–	4.63e-04	–	–	8.60e-02	–	1.79e-05	–	–	4.46e-02	–

Defensive importance sampling and sensitivity to λ. For the g_1 benchmark, we additionally evaluate the Hesterberg-style defensive counterpart of fRKL, denoted fRKL-d. We use a $K = 16$ trainable GMM and the full defensive proposal

$$p_{\boldsymbol{\theta}}^{\mathrm{def}}(\mathbf{x}) = \lambda f(\mathbf{x}) + (1-\lambda) p_{\mathrm{GMM},\boldsymbol{\theta}}(\mathbf{x})$$

throughout adaptation and evaluation. The fixed nominal component implies $f(\mathbf{x})/p_{\boldsymbol{\theta}}^{\mathrm{def}}(\mathbf{x}) \leq 1/\lambda$. Following Hesterberg [14, Sec. 6.2], we evaluate $\lambda \in \{0.1, 0.2, 0.3, 0.4, 0.5\}$. For each value, one trained proposal is held fixed and evaluated using 20 independent seeds; all reported defensive-IS estimates are nonzero.

Table 12 shows that the defensive baseline is not highly sensitive to λ over this range, and the conclusions of the main comparison remain unchanged. The fixed component bounds the importance weights, but in this experiment it does not prevent concentration of the adaptive GMM component; this is an empirical observation for g_1, not a general limitation of defensive importance sampling.

Table 12: Sensitivity of the Hesterberg-style defensive baseline on g_1 with $K = 16$, $R = 10^5$, and 20 independent evaluation runs.

Method	Estimate	Var. reduction	log-error	CoV
fRKL-d, $\lambda = 0.1$	4.773e-06	1.73e+03	5.06e-02	7.53e-02
fRKL-d, $\lambda = 0.2$	4.810e-06	1.71e+03	6.52e-02	7.55e-02
fRKL-d, $\lambda = 0.3$	4.866e-06	1.07e+03	8.54e-02	9.33e-02
fRKL-d, $\lambda = 0.4$	4.828e-06	1.03e+03	6.24e-02	9.49e-02
fRKL-d, $\lambda = 0.5$	4.955e-06	8.29e+02	8.66e-02	9.76e-02

D.1 Indicative running times

We report indicative training times for g_1 ($d = 2$), g_6 ($d = 100$), and g_7 ($d = 62$, ResNet). The measurements use our shared PyTorch implementation, but wall-clock time depends on implementation details such as batching, parallelization, and hardware. We did not optimize the code for wall-clock time and used largely shared, conservative epoch counts across benchmarks.

Table 13: Indicative training times in seconds. Panel (a) reports total training time; panel (b) reports separately measured operations on g_7.

(a) Total training time

Method	g_1 ($d = 2$)	g_6 ($d = 100$)
fFKL-c	~ 3000	~ 1570
fRKL-c	~ 55	~ 340
CFI-c	~ 5	~ 8
CFI-c+FKL-c	~ 60	~ 234

(b) Measured operations on g_7 ($d = 62$, ResNet)

Operation	Time
Full fRKL-c training	~ 130
CFI-c	~ 76
FKL-c refinement after CFI-c	~ 13
RKL-c refinement after CFI-c	~ 52
Flow after CFI-c+FKL-c	~ 38
Flow after CFI-c+RKL-c	~ 42

CFI is inexpensive because each intermediate level requires only an unweighted GMM fit to the current elite sample. Full adaptive KL/CE baselines instead repeat proposal optimization, including many gradient-based epochs, at every level. For example, on g_6, CFI-c+FKL-c takes approximately 234 seconds, compared with 1570 seconds for fFKL-c, while CFI-c alone takes approximately 8 seconds.

On g_7, flow refinement after CFI-c+RKL-c adds approximately 42 seconds and reduces CoV from 6.80×10^{-3} to 6.24×10^{-3}. After CFI-c+FKL-c, it adds approximately 38 seconds and reduces

CoV from 2.14×10^{-2} to 1.65×10^{-2}. These gains are useful but moderate relative to the extra training time, which motivates treating flow refinement as optional. The final KL and flow stages could likely use fewer epochs than the conservative shared settings used here.

E Scalability: higher-dimensional ResNet experiment

To further assess scalability, we consider a higher-dimensional variant of the ResNet-based rare-event benchmark g_7. The mechanism discussed in the main text becomes especially relevant in higher dimensions: in rare-event regimes, the conditional distribution $f(\mathbf{x} \mid g(\mathbf{x}) > \gamma)$ may concentrate near lower-dimensional regions or along a small number of dominant directions. This can lead to increasingly ill-conditioned covariance structures, making unconstrained KL-based optimization more prone to covariance collapse and degenerate proposals.

We therefore repeat the ResNet experiment with a finer grouping of network weights, increasing the perturbation dimension from $d = 62$ to $d = 174$. The setup is otherwise analogous to the g_7 benchmark described in Appendix L. Results are averaged over 20 independent seeds and use a GMM proposal with $K = 16$ components.

Table 14: Scalability experiment on the high-dimensional ResNet benchmark ($d = 174$, $K = 16$). Results are averaged over 20 independent seeds.

Method	log-err	CoV	log-err/var_red
CMC	3.89e-01	1.45e-01	3.89e-01
CFI-c	1.42e-01	1.34e-01	3.48e-02
CFI-c+RKL-c	1.57e-02	6.79e-03	5.37e-06
CFI-c+FKL-c	1.76e-01	1.45e-01	9.89e-02

All CFI-based methods outperform the CMC baseline on this higher-dimensional problem. In particular, CFI-c+RKL-c achieves the strongest performance, reducing the composite metric log-err/var_red by approximately four orders of magnitude relative to CMC. These results are consistent with the lower-dimensional ResNet experiment and indicate that the proposed constrained, coverage-first pipeline remains effective as the perturbation dimension increases.

F Hyperparameters and training details

All experiments are run using a largely shared set of hyperparameters. Thanks to the automatic selection of the smoothing parameter τ and β-annealing for the covariance constraint, no manual tuning is required across benchmarks, and the same configuration is used throughout unless stated otherwise.

Common hyperparameters. For all GMM-based stages (CFI, constrained FKL, constrained RKL), we use:

- initial sample size for model initialization: 10,000,
- surrogate batch size: $M = 25{,}000$,
- number of gradient steps: 500,
- learning rate: 5×10^{-3},
- smoothing factor: $\alpha = 0.75$,
- maximum number of outer iterations: 100,
- elite fraction (when applicable): $\rho = 0.2$.

Sensitivity to the smoothing parameter τ. The reverse-KL objectives are sensitive to the selection of smoothing parameter τ. This sensitivity is illustrated in Fig. 3 for the g_1 benchmark with $K = 6$ and $\mathbf{\Sigma} = \mathbf{\Sigma}^{(A)}$, where variance reduction relative to CMC is reported as a function of τ.

As seen in the figure, good performance is attained only within a relatively narrow range of τ, while overly small or large values lead to a noticeable degradation. Our adaptive procedure stabilizes

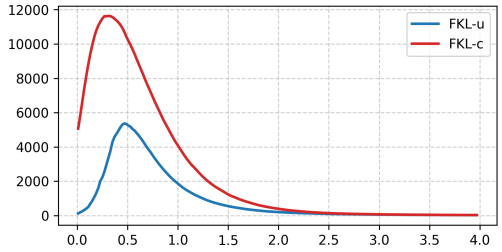

Figure 3: Variance reduction (y-axis) relative to CMC, as a function of the smoothing parameter τ (x-axis) for RKL-c (red) and RKL-u (blue).

early training and automatically selects values of τ within the effective range without any problem-specific tuning. In all experiments reported in this paper, a single choice of $c = 0.6$ and bounds $[\tau_{\min}, \tau_{\max}] = [0.02, 0.8]$ was used for all benchmarks.

In addition, β-annealing gradually increases β to avoid premature variance collapse and recovers the covariance floor required for the guarantee in the final constrained-GMM proposal.

Flow-specific hyperparameters. For the RealNVP refinement stage, we use the same training protocol across all benchmarks, with the following settings:

- initial sample size: 25,000,
- surrogate batch size: $M = 25{,}000$,
- number of gradient steps: 1,000,
- learning rate: 10^{-5}.

The only differences across benchmarks concern the architecture of the coupling networks. For low-dimensional problems (g_1, g_2, g_3, with $d = 2$), we use 48 coupling layers with hidden dimension 1024. For all higher-dimensional benchmarks, we use 32 coupling layers with hidden dimension 32. All other flow hyperparameters are kept fixed.

Parameter smoothing. After each refitting step in CFI, we smooth the newly fitted parameters with their previous values. Let primes denote the newly fitted, pre-smoothed quantities and let $\alpha \in (0, 1]$ be the weight assigned to the new fit.

Means and mixture weights are updated as

$$\boldsymbol{\mu}_k^{(t)} = \alpha \boldsymbol{\mu}_k' + (1-\alpha)\boldsymbol{\mu}_k^{(t-1)}, \qquad \pi_k^{(t)} = \alpha \pi_k' + (1-\alpha)\pi_k^{(t-1)}.$$

The mixture logits are then set to $\ell_k^{(t)} = \log(\pi_k^{(t)} + \varepsilon)$, with a small $\varepsilon > 0$ for numerical stability.

For unconstrained GMMs, we smooth the covariance matrices and recompute their Cholesky factors:

$$\boldsymbol{\Sigma}_k^{(t)} = \alpha \boldsymbol{\Sigma}_k' + (1-\alpha)\boldsymbol{\Sigma}_k^{(t-1)}, \qquad \mathbf{L}_k^{(t)} = \mathrm{Cholesky}(\boldsymbol{\Sigma}_k^{(t)}).$$

For constrained GMMs, we smooth only the free covariance component. Writing $\mathbf{S}_k' = \mathbf{L}_k' \mathbf{L}_k'^{\top}$ and $\mathbf{S}_k^{(t-1)} = \mathbf{L}_k^{(t-1)} (\mathbf{L}_k^{(t-1)})^{\top}$, we set

$$\mathbf{S}_k^{(t)} = \alpha \mathbf{S}_k' + (1-\alpha)\mathbf{S}_k^{(t-1)}, \qquad \mathbf{L}_k^{(t)} = \mathrm{Cholesky}(\mathbf{S}_k^{(t)}),$$

and form

$$\boldsymbol{\Sigma}_k^{(t)} = \mathbf{L}_k^{(t)} (\mathbf{L}_k^{(t)})^{\top} + \beta \boldsymbol{\Sigma}.$$

This preserves the covariance floor at the current value of β.

G Hyperparameter sensitivity analysis

This section studies the sensitivity of the proposed method to the main CFI hyperparameters. The parameter ρ controls the fraction of samples used in the level update and therefore affects the

exploration–exploitation trade-off: smaller values use a more selective elite set, while larger values produce more conservative level updates. The smoothing parameter α regularizes the parameter updates: smaller values lead to slower, more stable updates, while larger values make the proposal adapt more aggressively.

We evaluate sensitivity on the g_5 benchmark ($d = 40$) using CFI-c+RKL-c. We report the coefficient of variation (CoV) and the composite metric log-err/var_red used in the main experiments.

Sensitivity to ρ. Table 15 varies ρ while keeping $\alpha = 0.75$ fixed.

Table 15: Sensitivity to ρ for CFI-c+RKL-c on g_5 ($d = 40$), with fixed $\alpha = 0.75$.

ρ	CoV	log-err/var_red
0.05	$4.28 \cdot 10^{-2}$	$6.40 \cdot 10^{-4}$
0.10	$4.02 \cdot 10^{-2}$	$4.23 \cdot 10^{-4}$
0.25	$4.30 \cdot 10^{-2}$	$4.88 \cdot 10^{-4}$
0.30	$3.76 \cdot 10^{-2}$	$3.02 \cdot 10^{-4}$
0.50	$3.81 \cdot 10^{-2}$	$4.10 \cdot 10^{-4}$

The performance varies only mildly across a broad range of values. We use $\rho = 0.25$ as a representative default value: it is well within the stable region and avoids relying on the best value from this particular ablation.

Sensitivity to α. Table 16 varies α while keeping $\rho = 0.25$ fixed.

Table 16: Sensitivity to α for CFI-c+RKL-c on g_5 ($d = 40$), with fixed $\rho = 0.25$.

α	CoV	log-err/var_red
0.25	$4.77 \cdot 10^{-2}$	$4.44 \cdot 10^{-4}$
0.50	$4.62 \cdot 10^{-2}$	$4.90 \cdot 10^{-4}$
0.75	$4.30 \cdot 10^{-2}$	$4.88 \cdot 10^{-4}$
0.90	$3.86 \cdot 10^{-2}$	$3.95 \cdot 10^{-4}$

Again, the results are stable across moderate values of α. The value $\alpha = 0.75$ used in our experiments lies in this stable region and provides a balanced trade-off between update stability and adaptation speed. This choice is also consistent with standard cross-entropy practice, where smoothing is commonly used to stabilize parameter updates, with typical values around $\alpha = 0.7$ and recommended ranges such as $0.4 \leq \alpha \leq 0.9$ reported in [27, 28].

Role of the sample size M. The surrogate sample size M plays a different role from ρ and α. It is primarily a simulation budget rather than a stabilization hyperparameter: increasing M reduces Monte Carlo variability but increases computational cost, without changing the structure of the algorithm. For this reason, and to ensure fair comparison across methods, we keep M fixed throughout the experiments.

Summary. Overall, the ablation indicates that the method is not highly sensitive to the precise choice of ρ or α. This supports the use of a shared hyperparameter configuration across benchmarks.

H Alternative divergence measures

While alternative discrepancies such as Wasserstein distance and kernel Stein discrepancy (KSD) have been proposed for comparing distributions, their use in rare-event importance sampling is less direct. In our setting, the optimal target distribution is

$$h^{\mathrm{opt}}(\mathbf{x}) \propto \mathbf{1}(g(\mathbf{x}) > \gamma) f(\mathbf{x}),$$

which is not directly sampleable. KL-based objectives naturally lead to tractable estimators such as

$$\mathbb{E}\left[\mathbf{1}(g(\mathbf{X}) > \gamma) \log p_{\boldsymbol{\theta}}(\mathbf{X})\right],$$

which can be approximated using samples from f or from an importance proposal.

In contrast, Wasserstein-based approaches require solving an optimal transport problem between $p_{\boldsymbol{\theta}}$ and an empirical approximation of h^{opt}, leading to nested optimization procedures such as Sinkhorn iterations. These can be computationally demanding and statistically unstable in rare-event regimes, where the number of effective samples from the target region is small. Similarly, KSD-based methods require access to the score function $\nabla \log h^{\mathrm{opt}}(\mathbf{x})$, which is not directly available because h^{opt} contains a hard indicator. Smoothing the indicator is possible, but introduces an additional temperature parameter and corresponding sensitivity.

For these reasons, KL-based formulations remain the most natural choice in our setting: they connect directly to cross-entropy and variational objectives, admit simple Monte Carlo estimators, and are compatible with the constrained proposal families used in this work.

I Alternative rare-event sampling methods

A number of alternative approaches have been proposed for rare-event estimation. Large-deviation-based methods exploit problem-specific structure to construct asymptotically optimal proposals, but typically rely on strong assumptions and are not easily applicable in general black-box settings.

Sequential Monte Carlo (SMC) and related splitting methods construct a sequence of intermediate distributions or level sets to gradually reach the rare-event region. These methods can be effective and are natural baselines for rare-event simulation, but they often require careful design of intermediate levels, resampling, and rejuvenation steps. Neural SMC variants further increase proposal flexibility, at the cost of additional modeling and optimization complexity.

Adjoint-based sampling methods provide another route to rare-event estimation by using gradient or sensitivity information to guide sampling toward important regions. Such methods can be powerful when accurate derivatives and suitable problem structure are available, but they are less directly applicable in general black-box settings and may require substantially more algorithmic machinery.

In contrast, our approach focuses on a lightweight importance-sampling framework with explicit variance control and minimal structural assumptions. By combining coverage-first initialization with constrained KL refinement, we obtain a robust and computationally efficient pipeline applicable across a wide range of rare-event problems.

J CFI vs gradient ascent: detailed comparison

A natural question is whether the coverage stage of CFI can be replaced by gradient-based optimization of $g(\mathbf{x})$, for example via multiple gradient ascent restarts. We show that this approach suffers from two qualitatively different failure modes:

(i) **Local failure in multimodal functions**, where some modes are systematically missed;

(ii) **Global failure for unbounded ("runaway") functions**, where trajectories diverge to infinity.

(i) Multimodal functions. We consider a two-dimensional function with three modes located at $\boldsymbol{\mu}_1 = (4, 0)$, $\boldsymbol{\mu}_2 = (0, 4)$, and $\boldsymbol{\mu}_3 = (3.3, 3.3)$. The function is defined as

$$g(\mathbf{x}) = \max\{h_1(\mathbf{x}), h_2(\mathbf{x}), h_3(\mathbf{x})\},$$

where

$$h_1(\mathbf{x}) = \exp\left(-\frac{(x_1-4)^2 + x_2^2}{8}\right), \quad h_2(\mathbf{x}) = \exp\left(-\frac{x_1^2 + (x_2-4)^2}{8}\right),$$

and

$$h_3(\mathbf{x}) = 2.8\exp\left(-\frac{(x_1-3.3)^2 + (x_2-3.3)^2}{1.6}\right) - 0.9\exp\left(-\frac{x_1^2}{0.7}\right) - 0.9\exp\left(-\frac{x_2^2}{0.7}\right).$$

This construction ensures that, in a neighborhood of the origin, the gradient points toward either $\boldsymbol{\mu}_1$ or $\boldsymbol{\mu}_2$, making $\boldsymbol{\mu}_3$ difficult to reach. Figure 4 illustrates the geometry of this example. The gradient field directs trajectories toward $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, while $\boldsymbol{\mu}_3$ has a very small basin of attraction. We perform

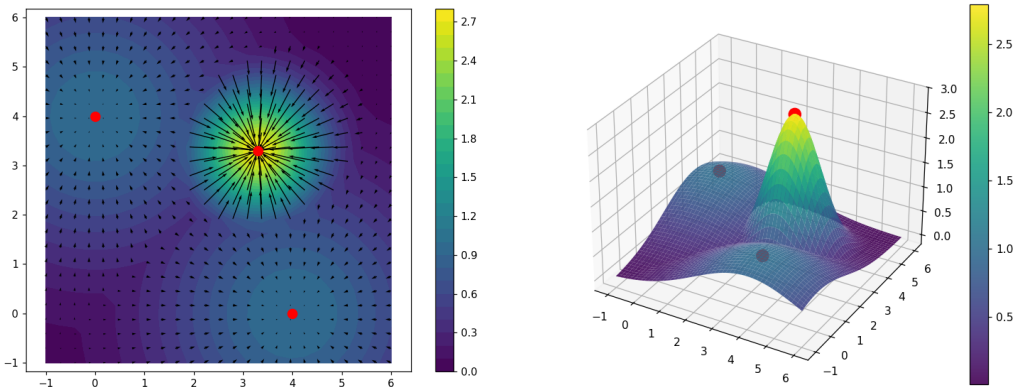

Figure 4: Multimodal example: contour plot with gradient field (left) and surface plot (right). Gradient ascent trajectories initialized near the origin are attracted to $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_2$, while $\boldsymbol{\mu}_3$ has a very small basin of attraction.

1000 gradient ascent runs initialized from $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and assign each final point to the nearest mode. We compare this with CFI-c using Gaussian mixture proposals. For gradient ascent, the entries below count final points among the 1000 restarts; for CFI-c, they count fitted mixture components assigned to each mode.

Method	$\boldsymbol{\mu}_1$	$\boldsymbol{\mu}_2$	$\boldsymbol{\mu}_3$
Gradient ascent (1000 restarts)	499	501	0
CFI-c (GMM6)	1	1	4
CFI-c (GMM16)	4	3	9
CFI-c (GMM18)	8	5	5

Gradient ascent never discovers $\boldsymbol{\mu}_3$, while CFI-c identifies all three modes in each reported configuration. This highlights a fundamental limitation of gradient ascent: even with multiple restarts, it produces a collection of local optima rather than a proposal distribution covering the relevant regions.

(ii) Unbounded ("runaway") functions (global failure). We now consider the function

$$g(x_1, x_2) = \max(x_1 - 4,\ -x_1 x_2 - 4,\ x_1 x_2 - 4).$$

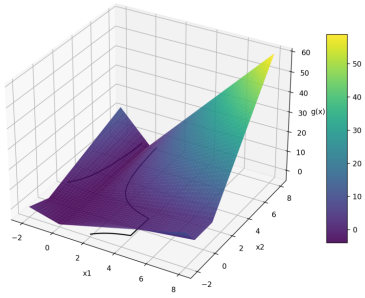

Figure 5: Runaway example: the function grows unbounded in certain directions, causing gradient ascent trajectories to diverge instead of approaching relevant level sets.

This function grows unbounded in multiple directions, making direct maximization ill-posed. Gradient ascent trajectories initialized from $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ diverge rapidly, with $\|\mathbf{x}_t\| \to \infty$, rather than converging to regions near the level set $g(\mathbf{x}) \approx \gamma$.

Figure 5 illustrates the geometry of this example. The absence of a well-defined maximizer prevents gradient ascent from producing a meaningful proposal distribution or notion of coverage.

In contrast, CFI-c targets successive level sets rather than a maximizer, so unboundedness of g alone does not make its objective ill-posed.

Summary. These examples show that gradient ascent may fail either locally (missing modes) or globally (diverging), whereas CFI-c provides a stable distributional mechanism for covering the rare-event region.

K Comparison with Safe-ICE

We additionally compare our method with Safe-ICE, a recent safe cross-entropy-based importance sampling method for rare-event simulation [11]. We used the publicly available implementation from [26]. Since this implementation was computationally expensive in our setting, with some low-dimensional runs already requiring several hours for $K_0 = 16$ initial components, we restrict the comparison to representative benchmarks rather than running Safe-ICE across the full experimental suite.

The comparison is motivated by the close connection between Safe-ICE and cross-entropy-based adaptive importance sampling, as well as by related dimension-reduction-based CE approaches for rare-event simulation [33]. We consider three cases: a low-dimensional benchmark (g_2, $d = 2$), a moderate-dimensional version of g_5 with $d = 4$, and the higher-dimensional g_5 benchmark with $d = 40$.

For the two-dimensional benchmark g_2, both methods produce accurate estimates. The true value is $2.226 \cdot 10^{-3}$; Safe-ICE gives $2.18 \cdot 10^{-3}$, while our method gives $2.225 \cdot 10^{-3}$, with lower computational cost in our implementation. Thus, in low dimension, both methods are accurate.

For the moderate-dimensional case g_5 with $d = 4$, Safe-ICE also gives a reasonable estimate, $2.78 \cdot 10^{-5}$, compared with the reference value $3.15 \cdot 10^{-5}$. This indicates that Safe-ICE remains effective for moderately complex settings, although the error is slightly larger than for our method.

The behavior differs sharply in the higher-dimensional benchmark g_5 with $d = 40$. In this case, our method remains stable and accurate, giving $3.40 \cdot 10^{-5}$ compared with the reference value $3.15 \cdot 10^{-5}$. In contrast, Safe-ICE collapses and returns a degenerate estimate, $9.14 \cdot 10^{-24}$. This behavior is consistent with the high-dimensional concentration and covariance-collapse effects discussed in the main text: cross-entropy-based updates can become unstable when the learned proposal concentrates too strongly around a small part of the rare-event region.

Interpretation. These results suggest that Safe-ICE is effective in low-dimensional or moderately complex settings, but its stability may degrade in higher dimensions. By contrast, our constrained-GMM variants combine coverage-first initialization with covariance constraints that provide the finite-variance guarantee of Theorem 1; the robustness of the flow-refined variants is an empirical finding. From a methodological perspective, the difference is structural: Safe-ICE follows a CE/ICE-style proposal adaptation based on cross-entropy updates, whereas our constrained CFI-based pipeline explicitly separates coverage from fitting and enforces covariance floors that prevent the variance-collapse mechanism identified in Proposition 1 and Theorem 1.

L Benchmark function definitions

Here we provide explicit definitions of the benchmark functions used in the experimental evaluation. All benchmarks are standard in the rare–event simulation literature and were adopted from previous studies, as summarized in the main text.

For $\mathbf{x} = (x_1, \ldots, x_d)^\top$ and $\mathbf{c} = (3.8, \ldots, 3.8)^\top$, the functions are defined as follows:

$$g_1(\mathbf{x}) = 1 - \min\left\{\|\mathbf{x} + \mathbf{c}\|^2,\ \|\mathbf{x} - \mathbf{c}\|^2\right\},$$

$$g_2(\mathbf{x}) = \min\left\{0.1(x_1 - x_2)^2 - \tfrac{x_1+x_2}{\sqrt{2}} + 3,\ 0.1(x_1 - x_2)^2 + \tfrac{x_1+x_2}{\sqrt{2}} + 3, x_1 - x_2 + \tfrac{7}{\sqrt{2}},\ x_2 - x_1 + \tfrac{7}{\sqrt{2}}\right\},$$

$$g_3(\mathbf{x}) = -\left(\sum_{i\ \text{odd}}^{d-1} 0.1\,(x_i - x_{i+1})^2 - \tfrac{1}{\sqrt{d}}\sum_{i=1}^{d} x_i + 2.5\right),$$

$$g_5(\mathbf{x}) = \sum_{i=1}^{m}\left[(x_{4i-3} + 10x_{4i-2})^2 + 5(x_{4i-1} - x_{4i})^2 + (x_{4i-2} - 2x_{4i-1})^4 + 10(x_{4i-3} - x_{4i})^4\right],$$

$$g_6(\mathbf{x}) = -\frac{1}{\sqrt{d}}\sum_{i=1}^{d} x_i, \quad g_7(\mathbf{x}) = \mathrm{ResNet}(\mathbf{x}), \quad g_{10}(\mathbf{x}) = \sum_{i=1}^{10}\exp(x_i).$$

The functions g_1, g_2, and g_3 are used in the two–dimensional setting ($d = 2$) and represent, respectively, a two–mode configuration ("Leaf"), a radial benchmark with piecewise boundary, and an anisotropic problem with a linear tilt. Functions g_5 and g_6 correspond to higher–dimensional test problems commonly used in rare–event estimation, while g_{10} models the tail probability of a sum of correlated lognormal random variables.

ResNet-based benchmark (g_7). We adopt the ResNet-based rare-event formulation introduced in [12], which studies rare but significant performance degradation of deep neural networks under stochastic parameter perturbations. Let $\mathbf{w} = \{w_i\}_{i=1}^d$ denote the parameters of a trained torchvision ResNet-18 model with `IMAGENET1K_V1` weights [32], grouped in a layerwise manner according to the network architecture, where $d = 62$ corresponds to the number of layerwise parameter groups. A perturbation vector $\mathbf{X} \in \mathbb{R}^d$ is drawn from a standard Gaussian distribution (i.e., $\mathbf{X} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$) and induces multiplicative noise applied independently to each parameter group.

In practice, the perturbed parameters are constructed as

```
for i, param in enumerate(model.parameters()):
    param.data = param.data * (1 + x[i] * noise_std)
```

which corresponds to the componentwise definition $w_i(\mathbf{x}) = w_i\big(1 + \kappa x_i\big), i = 1, \ldots, d$, where $\kappa = 0.2$ controls the perturbation magnitude. Let $\mathbf{w}$ denote the nominal (unperturbed) network parameters, while $\mathbf{w}(\mathbf{x})$ denotes their multiplicatively perturbed version corresponding to the realization $\mathbf{x}$ (thus $\mathbf{w} \equiv \mathbf{w}(\mathbf{0})$). Let $L(\mathbf{w})$ denote the average cross-entropy loss evaluated on a fixed validation set of 24 images from the `imagenetv2-matched-frequency-format` ImageNetV2 dataset [24]. The accepted g_7 evaluation uses the fixed prepared `imagenetv2-matched-frequency-format-val_first24.pt` tensor and torchvision ResNet-18 `IMAGENET1K_V1` weights. They are external, user-supplied local inputs and are neither bundled nor hosted by the release. The evaluator checks the documented fixed filename, size, and SHA-256 identities and runs offline without download or cache fallback. The legacy preparation script does not exactly reproduce the accepted tensor. The function is defined as

$$g_7(\mathbf{x}) = L(\mathbf{w}(\mathbf{x})) - L(\mathbf{w}). \tag{12}$$

Rare events correspond to $g_7(\mathbf{x}) > \gamma$, where $\gamma = 5.15715$ is chosen via extensive CMC calculations such that $\mathbb{P}(g_7(\mathbf{X}) > \gamma) \approx 6 \times 10^{-5}$, matching the reference value reported in [12].

Evaluating $g_7(\mathbf{x})$ is computationally expensive, as each function call requires inference with a perturbed network. To enable efficient evaluation while preserving the structure of the rare-event problem, a first-order approximation of the loss around the nominal parameters $\mathbf{w}$ is employed. A first-order Taylor expansion of (12) yields

$$g_7(\mathbf{x}) \approx \kappa \sum_{i=1}^{d} x_i \left\langle \nabla_{w_i} L(\mathbf{w}), w_i \right\rangle,$$

where w_i denotes the i-th layerwise parameter group and $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product. This linearized model enables rapid evaluation of $g_7(\mathbf{x})$ for a large number of perturbation samples and facilitates the application of importance-sampling-based rare-event estimation methods.

Summary of functions. References and experimental settings (thresholds γ, dimensions, and input distributions) follow the original works cited in the main text, they are summarized in Table 17. Input distribution A denotes $\mathcal{N}(\boldsymbol{\mu}^{(A)}, \boldsymbol{\Sigma}^{(A)})$ and B denotes $\mathcal{N}(\boldsymbol{\mu}^{g_{10}}, \boldsymbol{\Sigma}^{g_{10}})$.

M Normalizing flow architecture (RealNVP)

The normalizing flow used in the final refinement stage follows the RealNVP formulation introduced by [9]. The flow defines an invertible transformation $T_{\boldsymbol{\theta}} : \mathbb{R}^d \to \mathbb{R}^d$, which induces the density

$$p_{\boldsymbol{\theta}}^{\text{flow}}(\mathbf{X}) = p^{\text{KL}}\big(T_{\boldsymbol{\theta}}(\mathbf{X})\big) \left|\det J_{\boldsymbol{\theta}}(\mathbf{X})\right|,$$

where p^{KL} denotes the base distribution obtained after KL fitting (i.e., the constrained GMM), and $J_{\boldsymbol{\theta}}$ is the Jacobian of $T_{\boldsymbol{\theta}}$. The transformation $T_{\boldsymbol{\theta}}$ is constructed as a composition of K affine coupling layers,

$$T_{\boldsymbol{\theta}} = T_K \circ \cdots \circ T_1.$$

Table 17: Summary of benchmark settings used in the experiments.

Benchmark	Dimension d	Input	Threshold γ	Reference $\mathbb{P}\left(g(\mathbf{X}) > \gamma\right)$
g_1	2	A	0	4.7934e-6
g_2	2	A	0	2.22668e-3
g_3	2	A	0	4.2073e-3
g_5	40	A	-403.6636	3.1500e-5
g_6	100	A	3.5	2.3262e-4
g_7	62	A	5.15715	6.00e-5
g_{10}	10	B	20k	1.03526e-3
g_{10}	10	B	40k	4.683e-4
g_{10}	10	B	500k	1.77400e-5

In each coupling layer, the input vector is partitioned into two subvectors $(\mathbf{x}_a, \mathbf{x}_b)$, and the transformation is defined as

$$\mathbf{x}_a \mapsto \mathbf{x}_a,$$
$$\mathbf{x}_b \mapsto \mathbf{x}_b \odot \exp\left(s(\mathbf{x}_a)\right) + t(\mathbf{x}_a),$$

where $s(\cdot)$ and $t(\cdot)$ are neural networks. The partition alternates across successive layers, ensuring that all input dimensions are transformed and that the overall mapping is sufficiently expressive. This structure yields a triangular Jacobian, making the determinant $\det J_{\boldsymbol{\theta}}(\mathbf{X})$ computationally tractable.

Throughout training, the base distribution p^{KL} remains fixed, and the flow acts purely as a refinement layer on top of the KL-fitted proposal.

N Additional visualizations for g_1

This section provides additional graphical illustrations for the task of estimating $\mathbb{P}\left(g_1(\mathbf{X}) > 0\right)$ with $\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(A)}, \boldsymbol{\Sigma}^{g_{10}}\right)$ in the two–dimensional setting. The visualizations complement the numerical results reported in the main body and illustrate the qualitative behavior of the proposed pipeline across iterations, model classes, and refinement strategies.

Figure 6 shows representative snapshots of the learned proposal distributions for $K = 6$ mixture components. From top to bottom, the rows correspond to: (i) the constrained coverage-first initialization (CFI-c), (ii) constrained forward-KL refinement (FKL-c) initialized from CFI-c, and (iii) the subsequent flow-based refinement (+Flow) built on top of FKL-c. Each row displays selected iterations or epochs from the corresponding training procedure.

Comparison with fully adaptive CE. For comparison, Fig. 7 shows selected snapshots from the fully adaptive constrained CE method (fFKL-c), in which the proposal is re-optimized at each intermediate level using importance-weighted likelihood. On g_6 and g_7, all fFKL-c evaluations produce positive finite estimates. The method nevertheless typically requires more iterations and remains sensitive to hyperparameter choices and finite-sample variability; the $K = 16$ estimates in particular vary substantially across evaluation seeds. Thus, a finite numerical output should not be interpreted as uniformly accurate or low-variance performance. These observations motivate the two-stage pipeline based on CFI-c followed by constrained KL refinement.

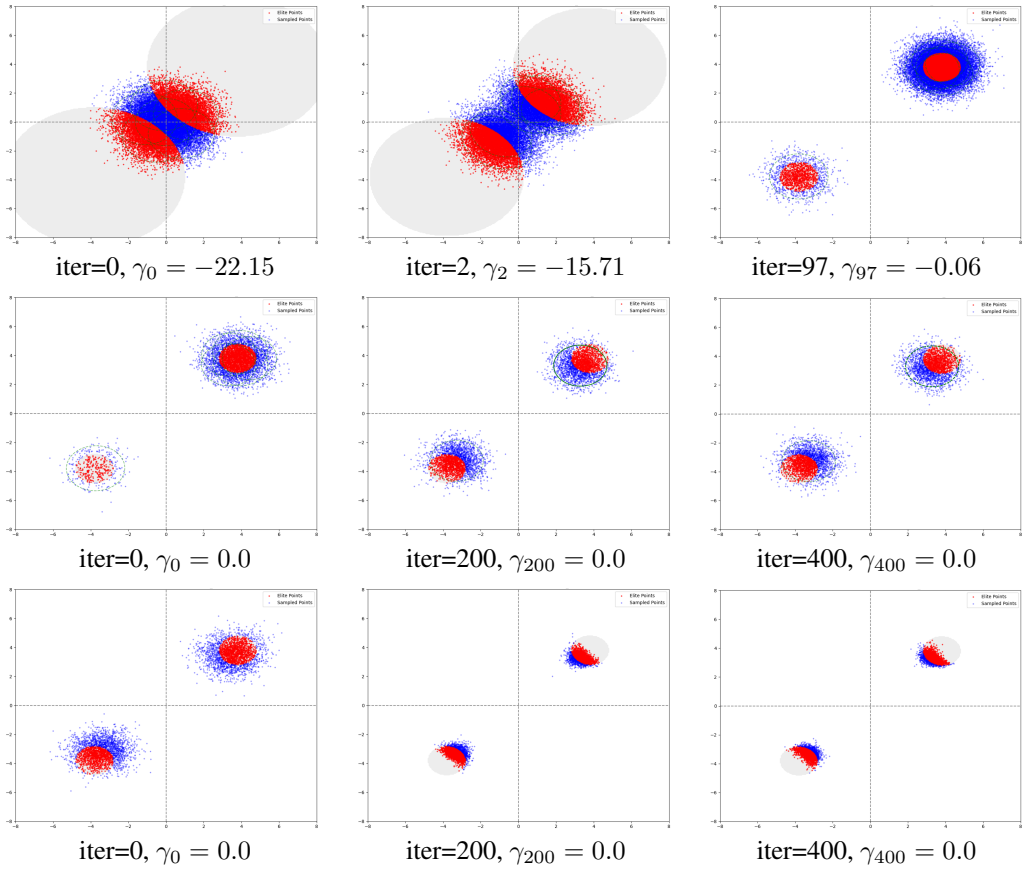

Figure 6: Evolution of the proposal distribution during training for estimating $\mathbb{P}\left(g_1(\mathbf{X}) > 0\right)$ with $\mathbf{X} \sim \mathcal{N}\left(\boldsymbol{\mu}^{(A)}, \boldsymbol{\Sigma}^{g_{10}}\right)$ and $K = 6$. Top to bottom: CFI-c, FKL-c initialized from CFI-c, and the subsequent flow-based refinement (+Flow).

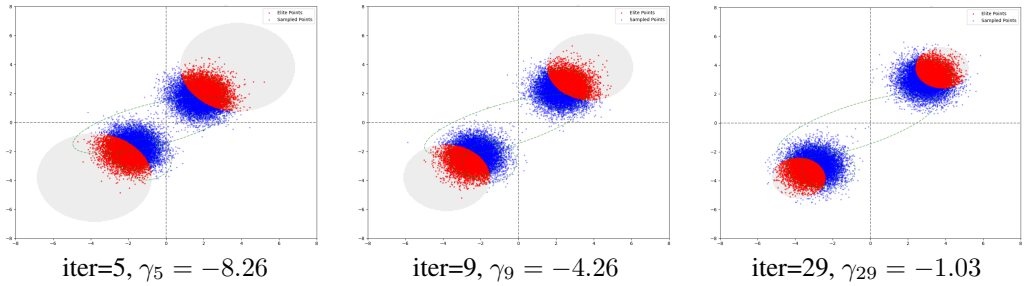

Figure 7: Selected snapshots from fFKL-c training for estimating $\mathbb{P}\left(g_1(\mathbf{X}) > 0\right)$ with $\mathbf{X} \sim \mathcal{N}\left(\mathbf{0}, I_2\right)$.