Some contributions on the adaptive importance sampling scheme 'cross-entropy' in high dimension

Jason Beh[1,2,3] Florian Simatos[1,3] Jérôme Morio[2,3]

[1]ISAE-SUPAERO
[2]ONERA
[3]Fédération ENAC ISAE-SUPAERO ONERA, Université de Toulouse, France

22 April 2025, Grenoble

Context: Rare event analysis

$Y \sim f$ initial density $\qquad \varphi : \mathbb{R}^d \to \mathbb{R} \qquad \varphi(Y)$

- $Y \sim f = N(0, I)$ standard Gaussian in dimension d
- $A = \{y \in \mathbb{R}^d : \varphi(y) \geq 0\}$, measurable $\varphi : \mathbb{R}^d \mapsto \mathbb{R}$
- **Estimate**

$$p = \mathbb{P}_f \left(Y \in A\right) = \int \mathbb{1}\left(y \in A\right) f(y) \, \mathrm{d}y$$

in high dimension: $d \to +\infty$

Monte Carlo

- Monte Carlo (MC) estimator: N samples
  $(Y_i) \overset{\text{iid}}{\sim} f$ initial density

$$\hat{p}_{\mathsf{MC}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left(Y_i \in A\right), \quad \mathbb{E}(\hat{p}_{\mathsf{MC}}) = p$$

- Choice of N : usually based on coefficient of variation (fluctuation around the mean)

$$\mathrm{cov}(\hat{p}_{\mathsf{MC}}) := \frac{\sqrt{\mathbb{V}\mathsf{ar}_f(\hat{p}_{\mathsf{MC}})}}{\mathbb{E}_f(\hat{p}_{\mathsf{MC}})} = \frac{\sqrt{1-p}}{\sqrt{Np}} \approx \frac{1}{\sqrt{Np}}$$

cov of $10\% \Rightarrow$ $N = 10^9$ if $p \sim 10^{-7}$

1 ms per sample $\Rightarrow$ 11 days to generate 10^9 samples

Importance sampling

- **Auxiliary density** g s.t. $g(x) = 0 \Rightarrow \mathbb{1}(x \in A) f(x) = 0$

$$p = \int_{\mathbb{R}^d} \mathbb{1}(x \in A) \frac{f(x)}{g(x)} g(x) \, \mathrm{d}x = \mathbb{E}_g \left(\mathbb{1}(X \in A) \frac{f(X)}{g(X)} \right)$$

- Importance sampling (IS): N samples $X_i \overset{\text{iid}}{\sim} g$

$$\hat{p}_g = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(X_i \in A) \frac{f(X_i)}{g(X_i)}, \ \mathbb{E}(\hat{p}_g) = p$$

Ideal density s.t. $\hat{p}_g = p$ constant estimator :

$$f|_A = \frac{\mathbb{1}(. \in A) f}{p} \quad \text{(conditional law } Y|\varphi(Y) \geq 0)$$

p unknown $\Rightarrow$ $f|_A$ unusable as auxiliary density!
But a good choice of g: Variance reduction compared to $\hat{p}_{\text{MC}}$

Importance sampling

- **Auxiliary density** g s.t. $g(x) = 0 \Rightarrow \mathbb{1}(x \in A) f(x) = 0$

$$p = \int_{\mathbb{R}^d} \mathbb{1}(x \in A) \frac{f(x)}{g(x)} g(x) \, \mathrm{d}x = \mathbb{E}_g \left(\mathbb{1}(X \in A) \frac{f(X)}{g(X)} \right)$$

- Importance sampling (IS): N samples $X_i \overset{\text{iid}}{\sim} g$

$$\hat{p}_g = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}(X_i \in A) \frac{f(X_i)}{g(X_i)}, \ \mathbb{E}(\hat{p}_g) = p$$

Ideal density s.t. $\hat{p}_g = p$ constant estimator :

$$f|_A = \frac{\mathbb{1}(. \in A) f}{p} \quad \text{(conditional law } Y|\varphi(Y) \geq 0)$$

p unknown $\Rightarrow$ $f|_A$ unusable as auxiliary density!
But a good choice of g: Variance reduction compared to $\hat{p}_{\text{MC}}$

Gaussian setting

Search g_A among **Gaussians** which minimizes Kullback-Leibler (KL) divergence with $f|_A$, $D(f|_A||\cdot)$:

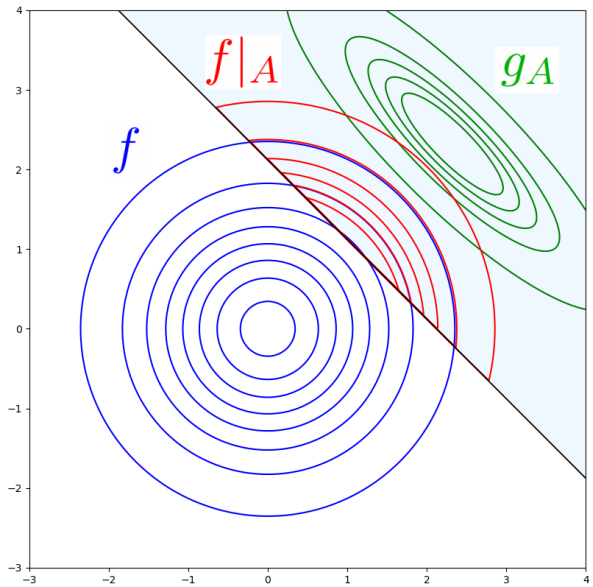
$$\min_{g=N(\mu,\Sigma)} D(f|_A||g) = \min_{g=N(\mu,\Sigma)} \mathbb{E}_{f|_A}\left(\log\left(\frac{f|_A(X)}{g(X)}\right)\right)$$

- Solution: g_A "Optimal Gaussian": Same mean and covariance matrix as $f|_A$

$$g_A = N(\mu_A, \Sigma_A);$$

$$\mu_A = \mathbb{E}_{f|_A}(X),\ \Sigma_A = \mathbb{V}\text{ar}_{f|_A}(X)$$

- μ_A et Σ_A unknown $\Rightarrow$ Estimation by Adaptive Importance Sampling
- Problem becomes vector and matrix estimation in high dimension

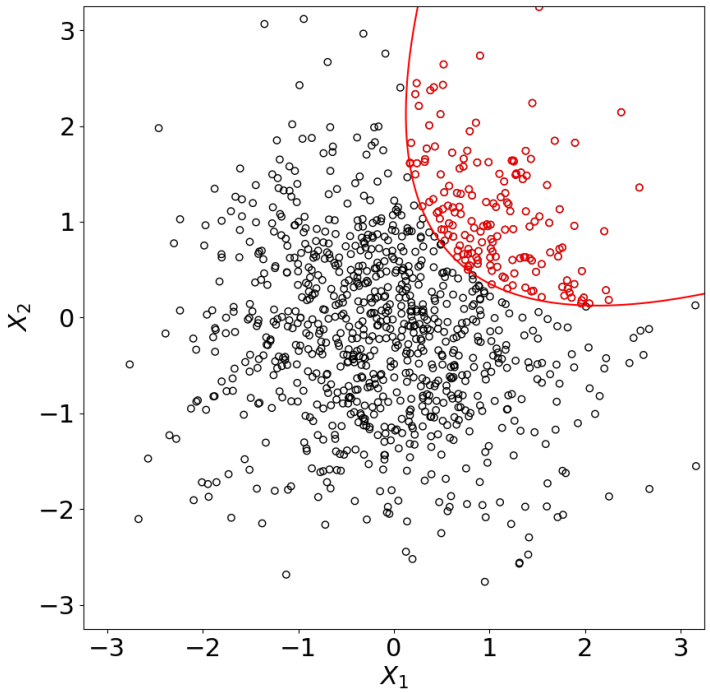


Cross-Entropy scheme (single Gaussian) [RK04]

μ_A and Σ_A unknown $\Rightarrow$ Estimation by Adaptive Importance Sampling

Cross-Entropy (CE): adaptive scheme which learns g_A iteratively.

Learning step: uses n samples per iteration



At iteration t:

1. n samples $(X_i) \overset{\text{iid}}{\sim} \hat{g}_t$
2. ρn highest $(\varphi(X_i)) \Rightarrow \hat{q}_{(\lfloor(1-\rho)n\rfloor)}$
3. $\hat{A}_t = \{y \in \mathbb{R}^d : \varphi(y) \geq \hat{q}_{(\lfloor(1-\rho)n\rfloor)}\}$
4. Estimate $\mathbb{E}_f(Y|Y \in \hat{A}_t)$ and $\mathbb{V}\text{ar}_f(Y|Y \in \hat{A}_t) \Rightarrow \hat{\mu}_{t+1}, \hat{\Sigma}_{t+1}$
5. $\hat{g}_{t+1} = N(\hat{\mu}_{t+1}, \hat{\Sigma}_{t+1})$

Stopping criterion $\hat{q}_{(\lfloor(1-\rho)n\rfloor)} > 0$ attained at some t^*: estimate p

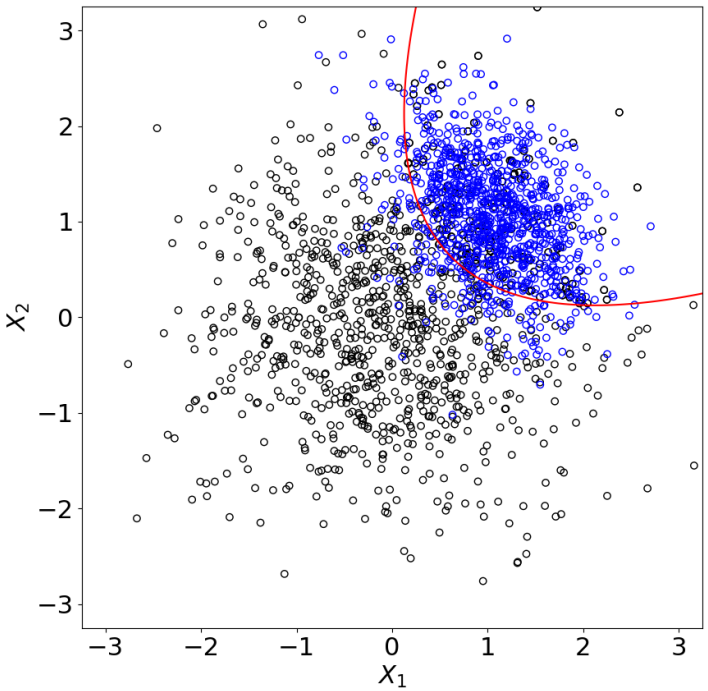
N samples $(X_i) \overset{\text{iid}}{\sim} \hat{g}_{t^*}$, $\hat{p}_{\text{CE}} = \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \mathbb{1}(X_i \in A)\, \dfrac{f(X_i)}{\hat{g}_{t^*}(X_i)}$

Cross-Entropy scheme (single Gaussian) [RK04]

μ_A and Σ_A unknown $\Rightarrow$ Estimation by Adaptive Importance Sampling

Cross-Entropy (CE): adaptive scheme which learns g_A iteratively.

Learning step: uses n samples per iteration



At iteration t:

1. n samples $(X_i) \overset{\text{iid}}{\sim} \hat{g}_t$
2. ρn highest $(\varphi(X_i)) \Rightarrow \hat{q}_{(\lfloor(1-\rho)n\rfloor)}$
3. $\hat{A}_t = \{y \in \mathbb{R}^d : \varphi(y) \geq \hat{q}_{(\lfloor(1-\rho)n\rfloor)}\}$
4. Estimate $\mathbb{E}_f(Y|Y \in \hat{A}_t)$ and $\mathbb{V}\text{ar}_f(Y|Y \in \hat{A}_t) \Rightarrow \hat{\mu}_{t+1}, \hat{\Sigma}_{t+1}$
5. $\hat{g}_{t+1} = N(\hat{\mu}_{t+1}, \hat{\Sigma}_{t+1})$

Stopping criterion $\hat{q}_{(\lfloor(1-\rho)n\rfloor)} > 0$ attained at some t^*: estimate p

$$N \text{ samples } \quad (X_i) \overset{\text{iid}}{\sim} \hat{g}_{t^*}, \quad \hat{p}_{\text{CE}} = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\left(X_i \in A\right) \frac{f(X_i)}{\hat{g}_{t^*}(X_i)}$$

Covariance matrix estimators in CE and CE with projection

- In CE at step t, estimation of $\mathbb{V}\mathrm{ar}_f(Y|Y \in \hat{A}_t)$ with $(X_i)_{i=1\dots n} \sim \hat{g}_t = N(\hat{\mu}_t, \hat{\Sigma}_t)$,

$$\hat{\Sigma}_{t+1} = \frac{1}{n\ \hat{p}_t} \sum_{i=1}^{n} \mathbb{1}\left(X_i \in \hat{A}_t\right) \frac{f(X_i)}{\hat{g}_t(X_i)} (X_i - \hat{\mu}_{t+1})(X_i - \hat{\mu}_{t+1})^\top$$

- This is an **importance sampling estimator of covariance matrix**
- CE with **projection** [Uri+21; EMS21; EMS24]: Fix $r \geq 1$, $(v_k, k = 1, \dots, r)$ orthonormal family, and use instead

$$\hat{\Sigma}_{t+1}^{\mathrm{proj}} = \sum_{k=1}^{r} (\lambda_k - 1) v_k v_k^\top + I, \lambda_k = v_k^\top \hat{\Sigma}_{t+1} v_k$$

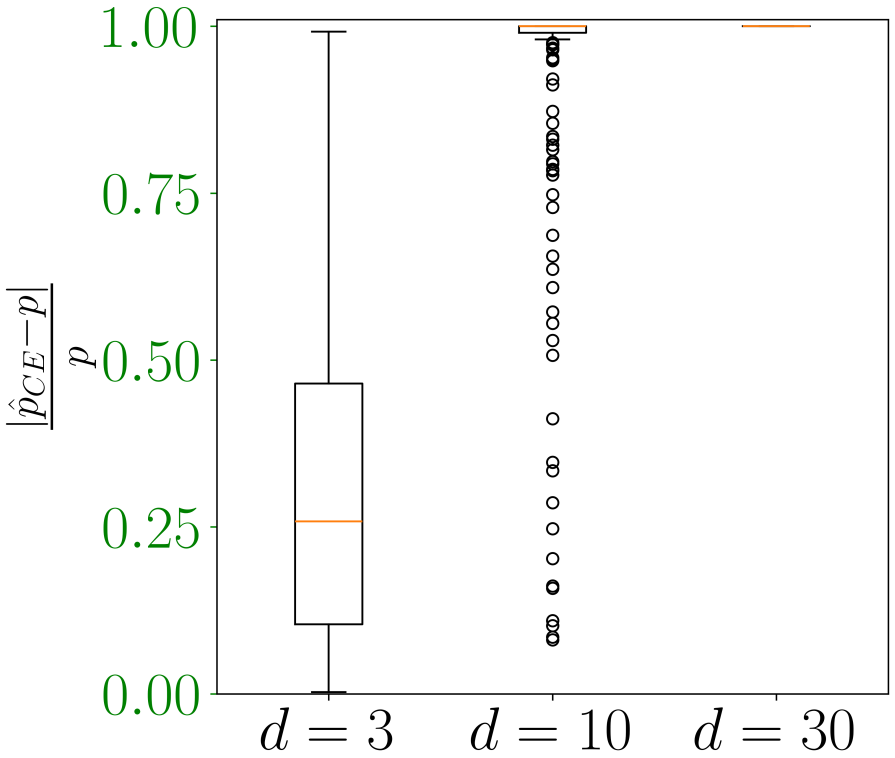
These matrices = culprit for bad performance of CE in high dimension ?

Bad performance of CE in high dimension

In low dimensions, CE = popular algorithm for rare event estimation

In high dimensions, CE does not converge

Example: $\varphi(y) = \sum_{j=1}^{d} y(j) - 5\sqrt{d},\ p \sim 10^{-7}$, plot $|\hat{p}_{\mathsf{CE}} - p|/p$



- To estimate p of order 10^{-7}, even $d = 10$ does not work well
- This inspires theoretical analysis of CE in high dimension ($d \to +\infty$)

Consistency of CE in high dimension

Joint work with F. Simatos, Yonatan Shadmi (Imperial College)
(to appear in *the Annals of Applied Probability* [BSSar])
Setting: $d \to +\infty$. n: no of samples per iteration of CE
N: no of IS samples to estimate p
Central assumption: $\inf_d p > 0$ + Technical assumptions

Theorem

$$\exists \kappa > 0 : \; n \gg d^{\kappa} \Longrightarrow \frac{\hat{p}_{\mathsf{CE}}}{p} \Rightarrow 1 \; \forall N \to \infty$$

n that scales polynomially with d suffices for the $\hat{p}_{\mathsf{CE}}$ to be consistent for any $N \to \infty$ (no minimal growth rate with d).

Why is it interesting? Popular folklore for IS:
$N \gg \exp(d)$ samples are needed to have a consistent estimator [BBL08]
What we have proven:
by learning g_A with CE beforehand using $n \gg d^{\kappa}$, $\hat{p}_{\mathsf{CE}}$ is consistent without needing $N \gg \exp(d)$.

Consistency of CE in high dimension

Joint work with F. Simatos, Yonatan Shadmi (Imperial College)
(to appear in *the Annals of Applied Probability* [BSSar])
Setting: $d \to +\infty$. n: no of samples per iteration of CE
N: no of IS samples to estimate p
Central assumption: $\inf_d p > 0$ + Technical assumptions

Theorem

$$\exists \kappa > 0 : \; n \gg d^{\kappa} \Longrightarrow \frac{\hat{p}_{\mathsf{CE}}}{p} \Rightarrow 1 \; \forall N \to \infty$$

n that scales polynomially with d suffices for the $\hat{p}_{\mathsf{CE}}$ to be consistent for any $N \to \infty$ (no minimal growth rate with d).

Why is it interesting? Popular folklore for IS:
$N \gg \exp(d)$ samples are needed to have a consistent estimator [BBL08]
What we have proven:
by learning g_A with CE beforehand using $n \gg d^{\kappa}$, $\hat{p}_{\mathsf{CE}}$ is consistent without needing $N \gg \exp(d)$.

Some remarks

1. In our knowledge, first result of the consistency of CE when $d \to +\infty$
2. if d is fixed and $n \to \infty$, CLT available in [PD18]
3. for $\hat{g}_{t+1} = N(\hat{\mu}_{t+1}, \hat{\Sigma}_{t+1})$, $n \gg d^{\kappa}$, we conjecture that κ is linked to $1/\lambda_{\min}(\hat{\Sigma}_t)$: cf. the rest of the talk
4. Similar results on more recent variants of CE: CE with projection [Uri+21; EMS22; EMS24] and improved CE [PGS19] can be obtained
5. The assumption $\inf_d p > 0$ is already considered for theoretical analysis in [AB03; CD18; CHR22]. Example of applications in [Uri+21; EPS24].

The importance sampling covariance matrix estimators

- Recall in CE with projection,

$$\hat{\Sigma}_{t+1}^{\mathrm{proj}} = \sum_{k=1}^{r} (\lambda_k - 1) v_k v_k^\top + I, \lambda_k = v_k^\top \hat{\Sigma}_{t+1} v_k$$

- Aim: confirm the conjecture that in $n \gg d^\kappa$, κ is linked to $1/\lambda_{\min}(\hat{\Sigma}_t^{\mathrm{proj}})$. Check also the necessity of $n \gg d^\kappa$.

Strategy: model $\hat{\Sigma}_{t+1}$ as a random matrix to use recent concentration inequalities from [BH24]

Joint work with F. Simatos, J. Morio (Preprint soon™)

Simplest matrix model: identity estimate

Covariance matrix estimation (Wishart)

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^\top := \frac{1}{n}\boldsymbol{Y}\boldsymbol{Y}^\top,$$

$Y_i \overset{\text{iid}}{\sim} f = N(0, I)$ or $\boldsymbol{Y}_{ij} \overset{\text{iid}}{\sim} N(0,1)$

Classical result $d \to +\infty$

[BY93] If $d/n \to c \in]0, 1]$,

- $\lambda_{\max}(\hat{I}) \to (1+\sqrt{c})^2$ a.s.
- $\lambda_{\min}(\hat{I}) \to (1-\sqrt{c})^2$ a.s.

Universality [BY93]: same result holds for iid $\boldsymbol{Y}_{ij}$ if

$$\mathbb{E}(\boldsymbol{Y}_{ij}) = 0,\ \mathbb{V}\text{ar}(\boldsymbol{Y}_{ij}) = 1 \text{ and } \mathbb{E}(\boldsymbol{Y}_{ij}^4) < +\infty.$$

Covariance matrix estimation by importance sampling

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)} X_i X_i^\top,$$

$X_i \sim g = N(0, \Sigma)$ iid

Universality inapplicable if $\frac{f(X)}{g(X)}$ does not have finite variance typically the case in importance sampling

Simplest matrix model: identity estimate

Covariance matrix estimation (Wishart)

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} Y_i Y_i^\top := \frac{1}{n}\boldsymbol{Y}\boldsymbol{Y}^\top,$$

$Y_i \overset{\text{iid}}{\sim} f = N(0, I)$ or $\boldsymbol{Y}_{ij} \overset{\text{iid}}{\sim} N(0,1)$

Classical result $d \to +\infty$

[BY93] If $d/n \to c \in]0, 1]$,

- $\lambda_{\max}(\hat{I}) \to (1+\sqrt{c})^2$ a.s.
- $\lambda_{\min}(\hat{I}) \to (1-\sqrt{c})^2$ a.s.

Universality [BY93]: same result holds for iid $\boldsymbol{Y}_{ij}$ if

$$\mathbb{E}(\boldsymbol{Y}_{ij}) = 0,\ \mathbb{V}\text{ar}(\boldsymbol{Y}_{ij}) = 1 \text{ and } \mathbb{E}(\boldsymbol{Y}_{ij}^4) < +\infty.$$

Covariance matrix estimation by importance sampling

$$\hat{I} = \frac{1}{n}\sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)} X_i X_i^\top,$$

$X_i \sim g = N(0, \Sigma)$ iid

Universality inapplicable if $\frac{f(X)}{g(X)}$ does not have finite variance typically the case in importance sampling

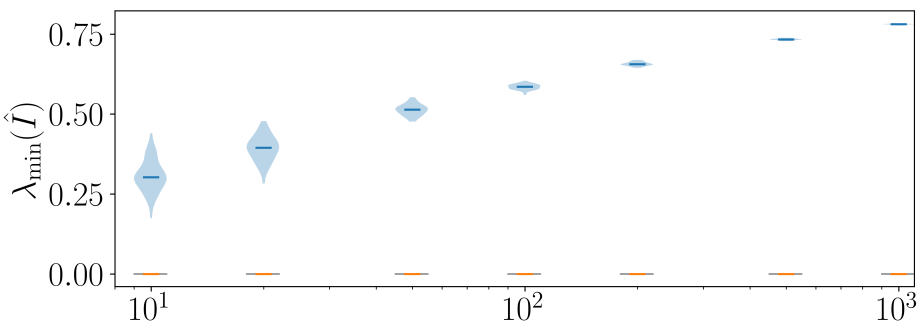
Identity estimate: result

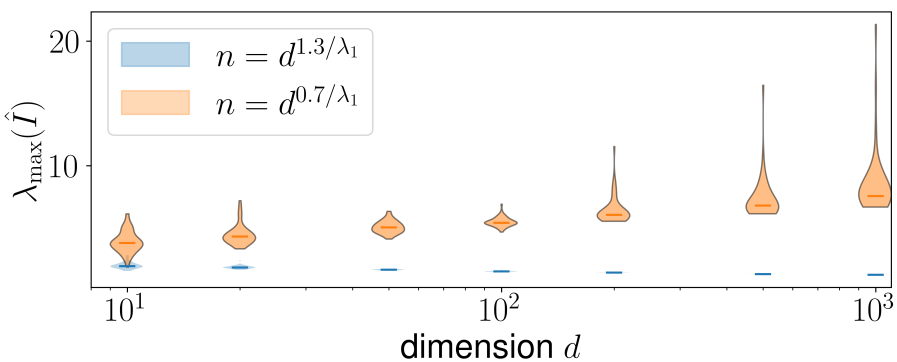
Fix $r \geq 1$, $\lambda_1 \leq \ldots \leq \lambda_r \in]0, +\infty[$,
$(v_k, k = 1, \ldots, r)$ orthonormal family, take $g = N(0, \Sigma)$ with

$$\Sigma = \sum_{k=1}^{r} (\lambda_k - 1) v_k v_k^\top + I. \quad \text{Consider} \quad \hat{I} = \frac{1}{n} \sum_{i=1}^{n} \frac{f(X_i)}{g(X_i)} X_i X_i^\top.$$

Theorem ($d \to +\infty$): If $n = d^\kappa$ for $\kappa > 0$, then

- if $\kappa > 1/\lambda_1$, then $\lambda_{\max}(\hat{I}), \lambda_{\min}(\hat{I}) \to 1$ in L_1
- if $\kappa < 1/\lambda_1$, then $\lambda_{\max}(\hat{I}) \Rightarrow +\infty$ $\quad (\lambda_1 = \lambda_{\min}(\Sigma)$ for $g)$





Model with limited dependency

Take $g = N(0, \Sigma)$, $\Sigma = \sum_{k=1}^{r} (\lambda_k - 1) e_k e_k^\top + I$ (canonical basis).

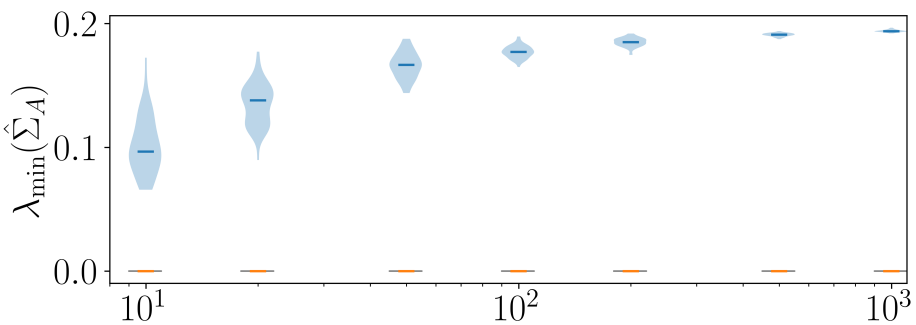
Consider $\hat{\Sigma}_A = \dfrac{1}{np} \sum_{i=1}^{n} \dfrac{f(X_i)}{g(X_i)} \mathbb{1}(X_i \in A) X_i X_i^\top$ (ignoring the mean)

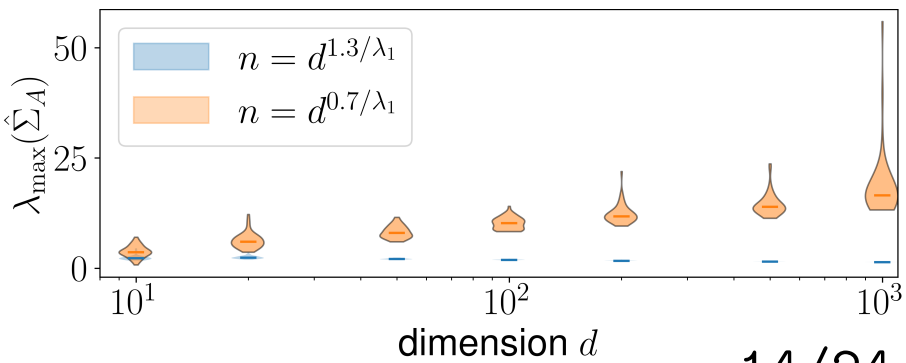
Theorem ($d \to +\infty$):

Assume that $\inf_d p > 0$. Then $\sup_d \lambda_{\max}(\Sigma_A) < +\infty$.

If $\mathbb{1}(X_i \in A)$ does not depend on the first r coordinates of X_i, and if $n = d^\kappa$ for $\kappa > 0$, then

- if $\kappa > 1/\lambda_1$, then $\lambda_{\max}(\hat{\Sigma}_A)/\lambda_{\max}(\Sigma_A) \to 1$ in L_1
  $\lambda_{\min}(\hat{\Sigma}_A)/\lambda_{\min}(\Sigma_A) \to 1$ in L_1
- if $\kappa < 1/\lambda_1$, then $\lambda_{\max}(\hat{\Sigma}_A) \Rightarrow +\infty$





Takeaway from theoretical results

- $\lambda_{\min}(\Sigma)$ of $g = N(0, \Sigma)$ is crucial to assure that $\lambda_{\max}(\hat{\Sigma}_A)$ does not tend to infinity
- At iteration t of CE-proj, we expect similar phenomenon: $\lambda_{\min}(\hat{\Sigma}_t^{\text{proj}})$ is critical to assure a good estimation of $\hat{\Sigma}_{t+1}$

Expected phenomenon:

$\lambda_{\min}(\hat{\Sigma}_t^{\text{proj}})$ small $\longrightarrow$ $n \ll d^{1/\lambda_1(\hat{\Sigma}_t^{\text{proj}})}$ $\longrightarrow$ $\lambda_{\max}(\hat{\Sigma}_{t+1})$ large

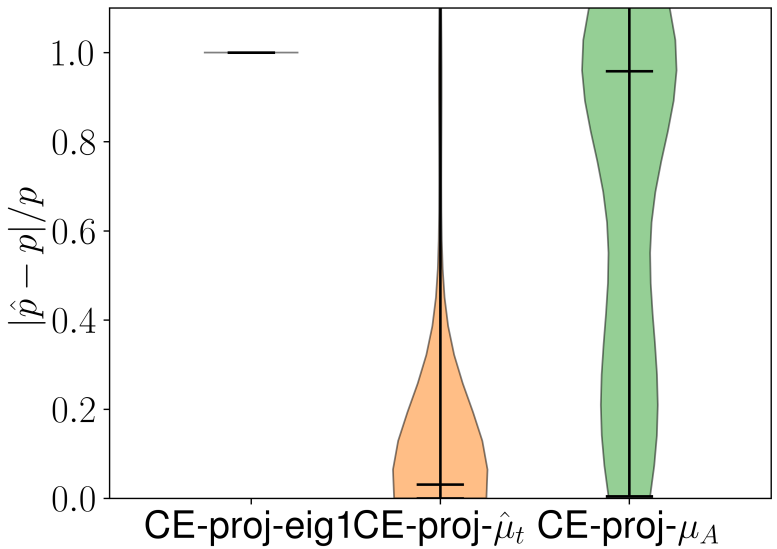
- Too large $\lambda_{\max}(\hat{\Sigma}_{t+1})$ degrades estimation of p since $\lambda_{\max}(\hat{\Sigma}_{t+1}) \Rightarrow +\infty \Longrightarrow D(f|_A||N(\mu_A, \hat{\Sigma}_{t+1})) \Rightarrow +\infty$ and [CD18]: required sample size to estimate p using g: $N \simeq e^{D(f|_A||g)}$
- Numerical verification on some test cases

Classical hyperplane test case

$$A = \left\{ x \in \mathbb{R}^d : \sum_{j=1}^{d} x(j) \geq 5\sqrt{d} \right\}, \; d = 100, \; p = 2.87 \cdot 10^{-7}$$

$\mu_A = \mathbb{E}_{f|_A}(X)$, $\Sigma_A = \mathbb{V}\text{ar}_{f|_A}(X)$
optimal parameters

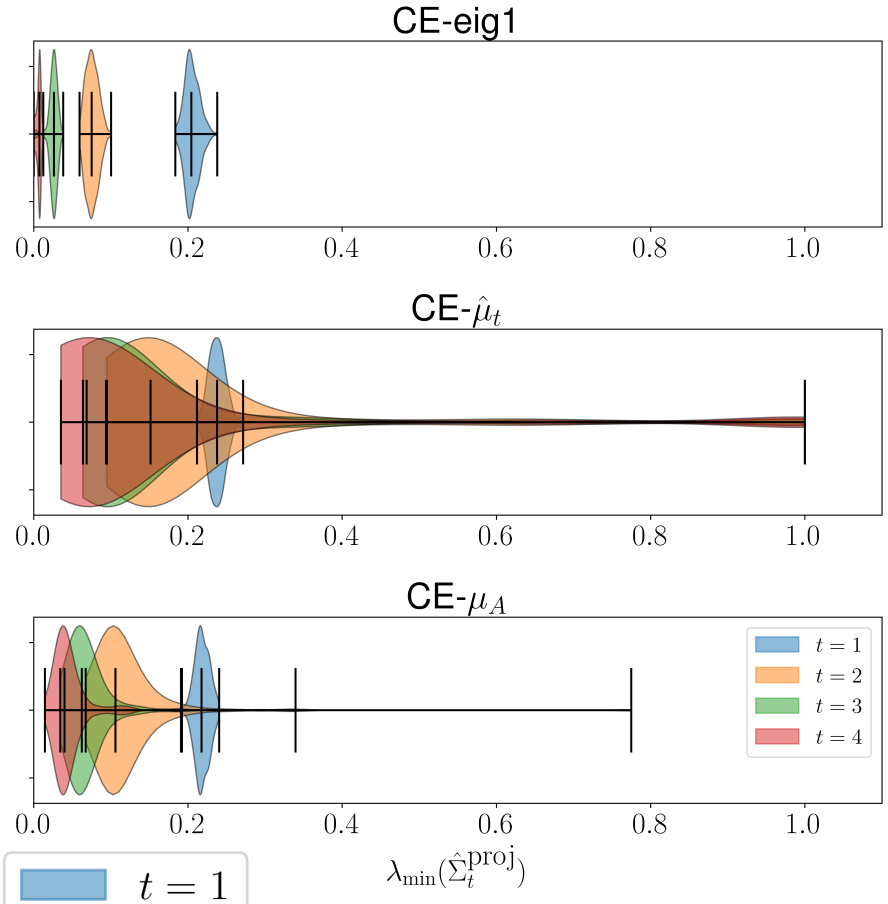
1. CE-eig1: eigenvector associated to the smallest eigenvalue of $\hat{\Sigma}_t$ [EMS24],
2. CE-$\hat{\mu}_t$: the direction of the estimated mean during each iteration t, $\hat{\mu}_t / \|\hat{\mu}_t\|$ [EMS21]
3. CE-μ_A: the theoretical mean $\mu_A / \|\mu_A\|$ [EMS21]



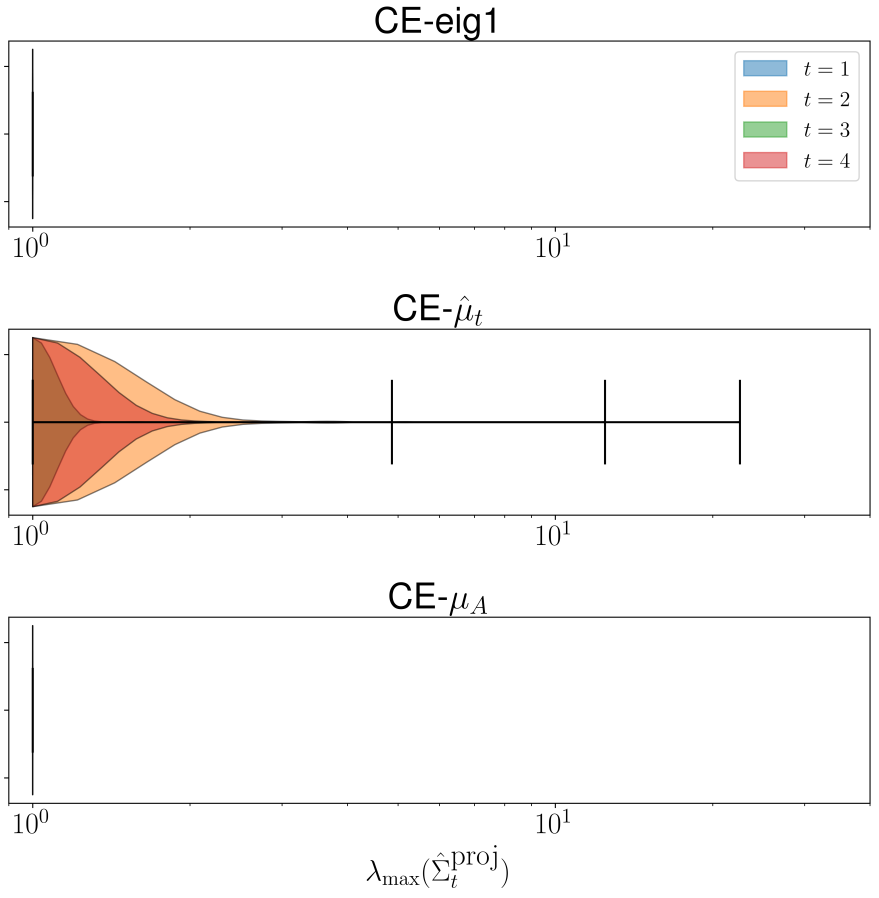
Fitted violin plot of $|\hat{p} - p|/p$ over 200 runs

Extreme eigenvalues across the iterations

Smallest eigs $\lambda_{\min}(\hat{\Sigma}_t^{\text{proj}})$



Largest eigs $\lambda_{\max}(\hat{\Sigma}_t^{\text{proj}})$



- It seems that larger smallest eigenvalues correlate to the better estimation of p

Plotting the extreme eigenvalues of two types of matrices

- In CE at step t, estimation of $\mathbb{V}\mathrm{ar}_f(Y|Y \in \hat{A}_t)$ with $(X_i)_{i=1\dots n} \sim \hat{g}_t = N(\hat{\mu}_t, \hat{\Sigma}_t)$,

$$\hat{\Sigma}_{t+1} = \frac{1}{n\ \hat{p}_t} \sum_{i=1}^{n} \mathbb{1}\left(X_i \in \hat{A}_t\right) \frac{f(X_i)}{\hat{g}_t(X_i)} (X_i - \hat{\mu}_{t+1})(X_i - \hat{\mu}_{t+1})^\top$$

We call the covariance matrix of CE before projection

- **Projection**: Given $(v_k, k = 1, \dots, r)$ orthonormal family of size r, construct

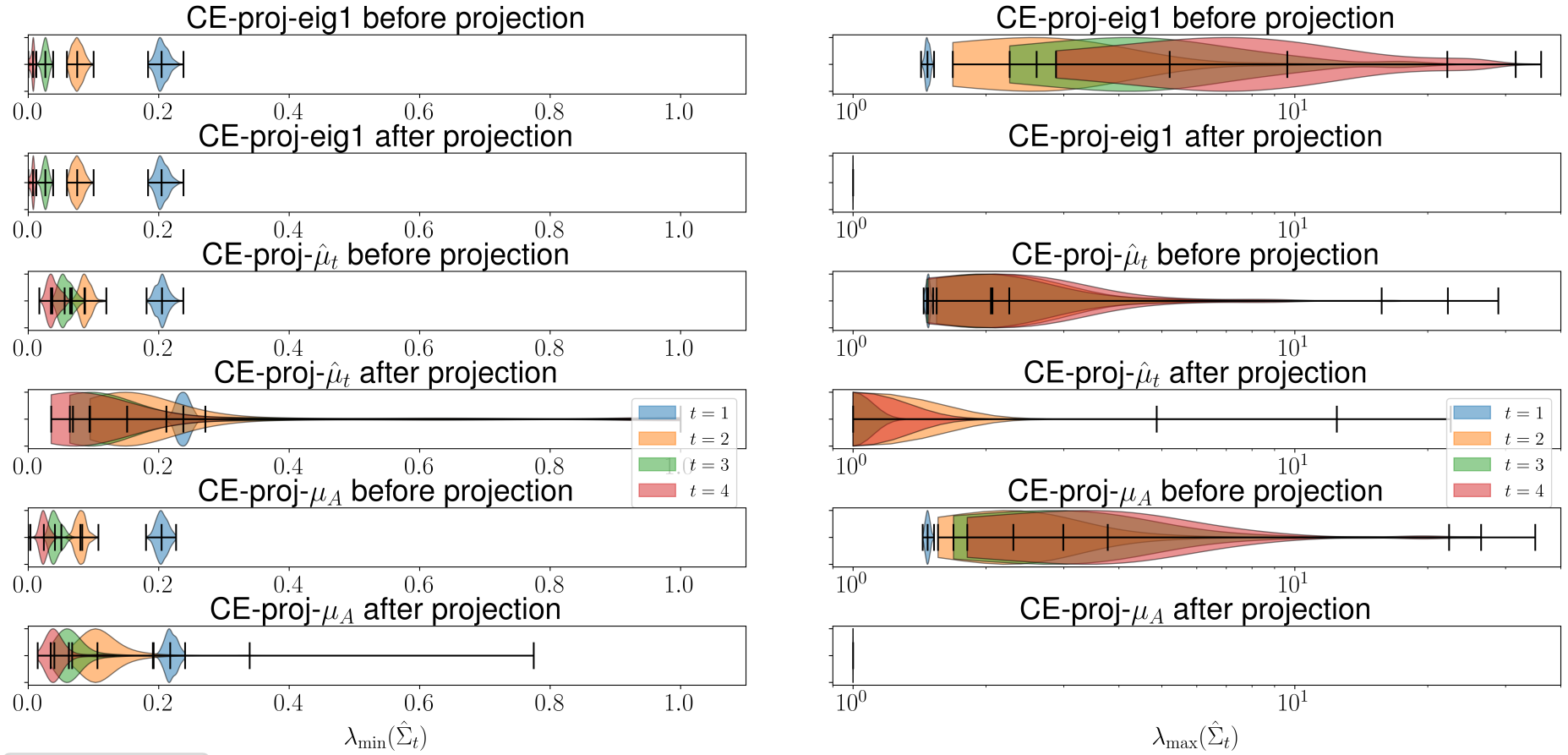
$$\hat{\Sigma}_{t+1}^{\mathrm{proj}} = \sum_{k=1}^{r} (\lambda_k - 1) v_k v_k^\top + I, \lambda_k = v_k^\top \hat{\Sigma}_{t+1} v_k$$

We call the covariance matrix of CE after projection

Effect of projection

Smallest eigs $\lambda_{\min}(\hat{\Sigma}_t), \lambda_{\min}(\hat{\Sigma}_t^{\text{proj}})$

Largest eigs $\lambda_{\min}(\hat{\Sigma}_t), \lambda_{\max}(\hat{\Sigma}_t^{\text{proj}})$



- If smallest eigenvalues collapse, the largest eigenvalues before projection become too large
- Projection usually sets largest eigenvalue to 1, and increases the smallest eigenvalues

Large portfolio losses example [BJZ08]

$$A = \left\{ x \in \mathbb{R}^d : \sum_{j=3}^{d} \mathbb{1}\left(\phi(x(1), x(2), x(j)) \geq 0.5\sqrt{d}\right) - 0.25d - 0.1 \right\},$$

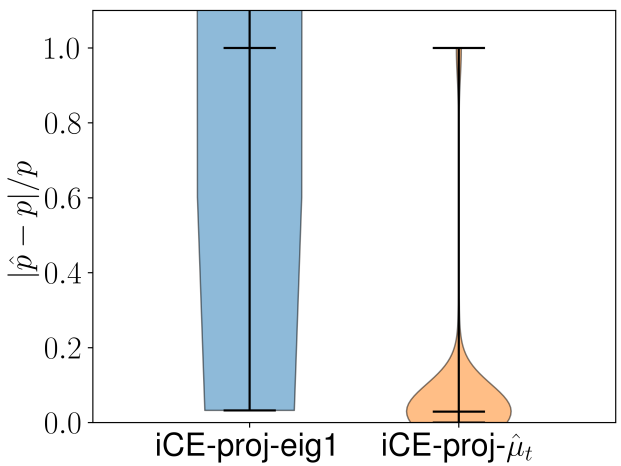
where for any $(x_1, x_2, x_3) \in \mathbb{R}^3$,

$$\phi(x_1, x_2, x_3) = \left(0.25\, x_1 + 3(1 - 0.25^2)^{1/2} x_3\right) \sqrt{F^{-1}_{\Gamma(6,6)}\left(F_N(x_2)\right)}$$

with $F_{\Gamma(6,6)}$ the cdf of the Gamma distribution.
$d = 334$, $p = 1.79 \cdot 10^{-6}$ (Monte Carlo with $5 \cdot 10^{10}$ samples)

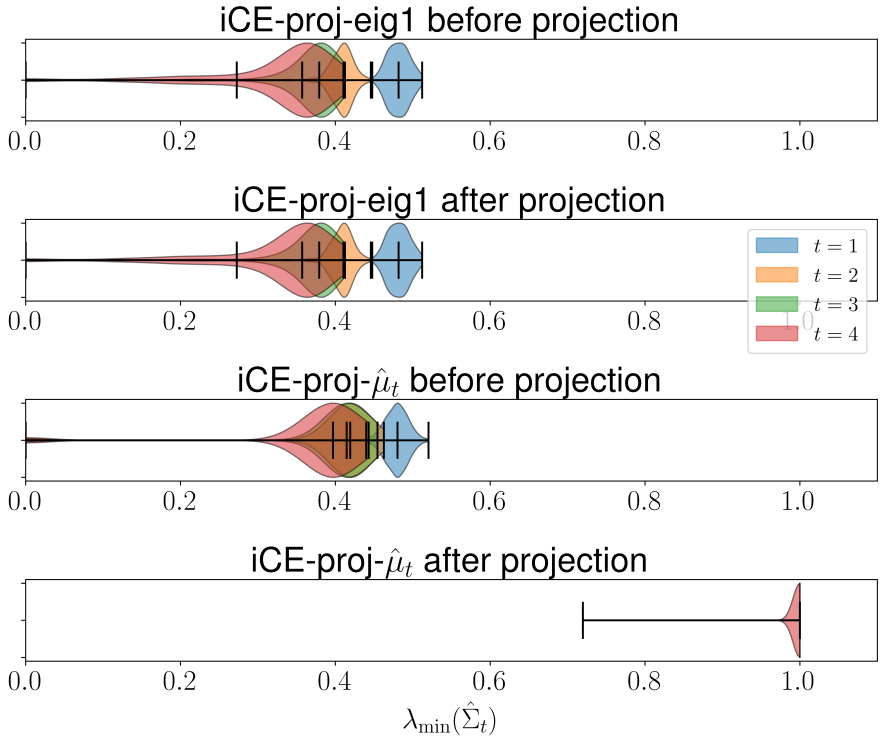
Improved cross-entropy (iCE) [PGS19]

1. iCE-proj-eig1: eigenvector associated to its smallest eigenvalue [EMS24],
2. iCE-proj-$\hat{\mu}_t$: the direction of the estimated mean during each iteration t, $\hat{\mu}_t / \|\hat{\mu}_t\|$ [EMS21]

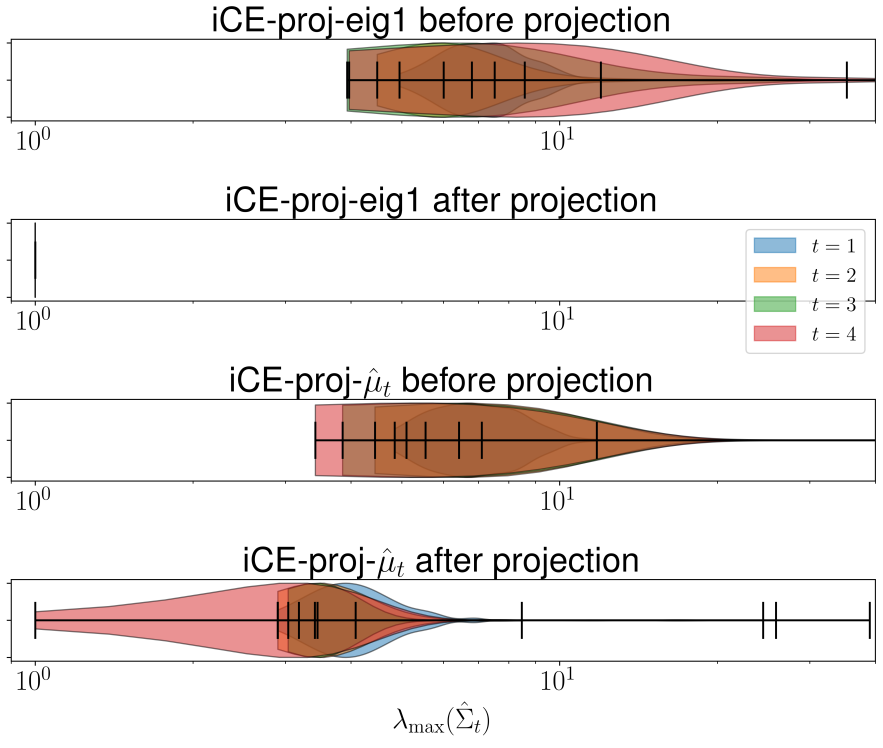


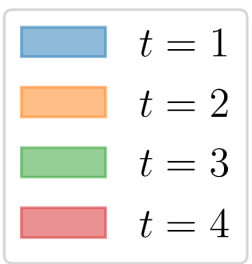
Extreme eigenvalues across the iterations

Smallest eigs $\lambda_{\min}(\hat{\Sigma}_t), \lambda_{\min}(\hat{\Sigma}_t^{\text{proj}})$



Largest eigs $\lambda_{\max}(\hat{\Sigma}_t), \lambda_{\max}(\hat{\Sigma}_t^{\text{proj}})$





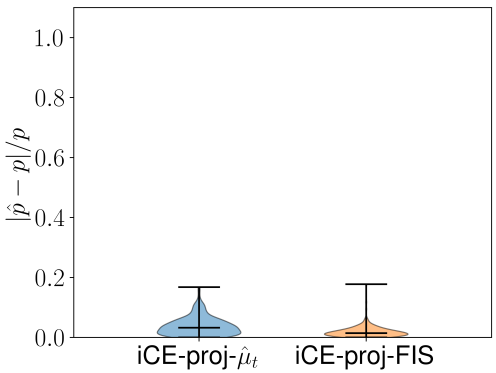
- Empirically, it seems that larger smallest eigenvalues translate into a better performance
- This is not always the case: cf. next example

Quadratic example

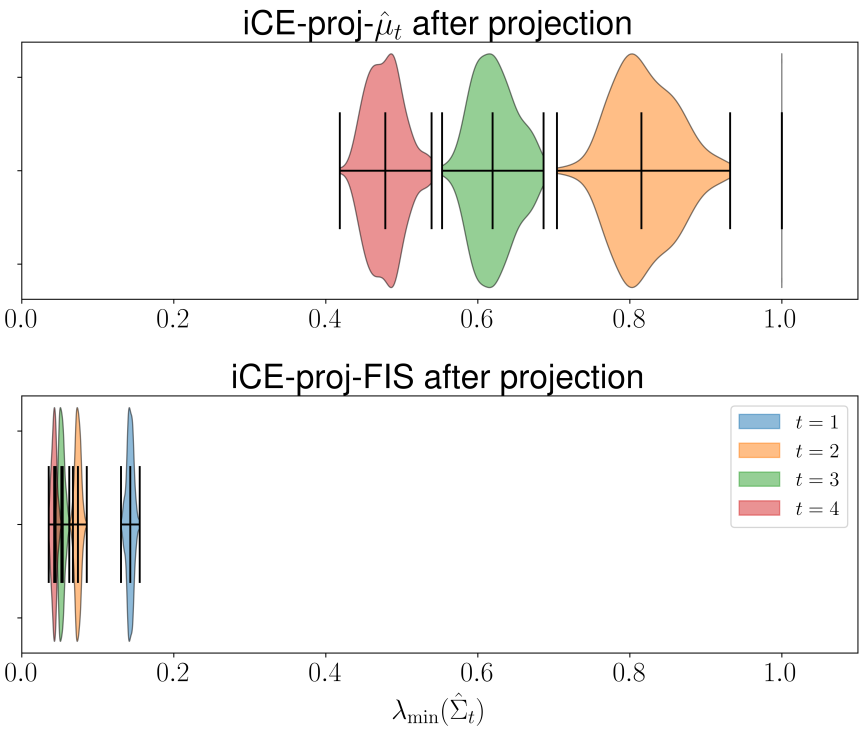
$$A = \left\{ x \in \mathbb{R}^d : -4 - \frac{5}{4}(x(1) - x(2))^2 + \frac{1}{\sqrt{d}} \sum_{j=1}^{d} x(j) \geq 0 \right\}$$

$d = 334$, $p = 6.62 \cdot 10^{-6}$

iCE-proj-FIS:
Failure-informed subspace constructed using the gradient $\nabla\varphi$ [Uri+21]



Smallest eigs $\lambda_{\min}(\hat{\Sigma}_t^{\text{proj}})$



Similar performance with iCE-proj-$\hat{\mu}_t$ but very small smallest eigenvalues: likely due to FIS = high quality projections

Numerical takeaways

1. We observed through numerical studies the phenomenon

$\lambda_{\min}(\hat{\Sigma}_t^{\text{proj}})$ small $\longrightarrow$ $n \ll d^{1/\lambda_1(\hat{\Sigma}_t^{\text{proj}})}$ $\longrightarrow$ $\lambda_{\max}(\hat{\Sigma}_{t+1})$ large

2. In most cases, for runs with a bad estimation of p, small $\lambda_{\min}(\hat{\Sigma}_t^{\text{proj}})$'s are observed

3. Our results suggest to regularize the eigenvalues of the covariance matrix estimator
   - Related to other work presented in SIAM-UQ24: Improved high-dimensional covariance matrix estimation in CE scheme (joint work with J. Morio, F. Simatos)

Perspective and advertisements

Perspective

- $\lambda_{\min}(\hat{\Sigma}_t)$ is not the entire story: iCE-proj-FIS can have small smallest eigenvalues but end up with a good estimation of p
- Influence on the choice of direction of projections not captured by the theoretical results

Advertisements

1. Upcoming work: interacting Langevin dynamics for rare event estimation (joint work with Simon Weissmann (University of Mannheim), F. Simatos, J. Morio) will be presented in ENUMATH 2025 @ Heidelberg (Preprint soon™)
2. Gladly appreciate post-doc opportunities

References I

[AB03] S. Au and J. Beck. "Important sampling in high dimensions". In: *Structural Safety* 25.2 (2003), pp. 139–163.

[BBL08] T. Bengtsson, P. Bickel, and B. Li. "Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems". In: *Institute of Mathematical Statistics Collections*. Beachwood, Ohio, USA: Institute of Mathematical Statistics, 2008, pp. 316–334.

[BH24] T. Brailovskaya and R. van Handel. "Universality and Sharp Matrix Concentration Inequalities". In: *Geometric and Functional Analysis* 34.6 (2024), pp. 1734–1838.

References II

[BJZ08] A. Bassamboo, S. Juneja, and A. Zeevi. "Portfolio Credit Risk with Extremal Dependence: Asymptotic Analysis and Efficient Simulation". In: *Operations Research* 56.3 (2008), pp. 593–606.

[BSSar] J. Beh, Y. Shadmi, and F. Simatos. "Insight from the Kullback–Leibler divergence into adaptive importance sampling schemes for rare event analysis in high dimension". In: *The Annals of Applied Probability* (to appear).

[BY93] Z. D. Bai and Y. Q. Yin. "Limit of the Smallest Eigenvalue of a Large Dimensional Sample Covariance Matrix". In: *The Annals of Probability* 21.3 (1993), pp. 1275–1294.

References III

[CD18] S. Chatterjee and P. Diaconis. "The sample size required in importance sampling". In: *The Annals of Applied Probability* 28.2 (2018).

[CHR22] F. Cérou, P. Héas, and M. Rousset. "Entropy minimizing distributions are worst-case optimal importance proposals". In: *arXiv preprint* (2022). arXiv: 2212.04292 [math.NA].

[EMS21] M. El Masri, J. Morio, and F. Simatos. "Improvement of the cross-entropy method in high dimension for failure probability estimation through a one-dimensional projection without gradient estimation". In: *Reliability Engineering & System Safety* 216 (2021), p. 107991.

References IV

[EMS22] M. El Masri, J. Morio, and F. Simatos. "Optimal projection to improve parametric importance sampling in high dimension". In: *arXiv preprint* (2022). arXiv: 2107.06091 [stat.CO].

[EMS24] M. El Masri, J. Morio, and F. Simatos. "Optimal Projection for Parametric Importance Sampling in High Dimensions". In: *Computo* (11, 2024).

[EPS24] M. Ehre, I. Papaioannou, and D. Straub. "Stein Variational Rare Event Simulation". In: *arXiv preprint* (2024). arXiv: 2308.04971 [stat.ME].

[PD18] F. Portier and B. Delyon. "Asymptotic optimality of adaptive importance sampling". In: *Advances in Neural Information Processing Systems*. Ed. by S. Bengio et al. Vol. 31. Curran Associates, Inc., 2018.

References V

[PGS19] I. Papaioannou, S. Geyer, and D. Straub. "Improved cross entropy-based importance sampling with a flexible mixture model". In: *Reliability Engineering & System Safety* 191 (2019), p. 106564.

[RK04] R. Y. Rubinstein and D. P. Kroese. *The cross entropy method: A unified approach to combinatorial optimization, monte-carlo simulation (information science and statistics)*. Berlin, Heidelberg: Springer-Verlag, 2004.

[Uri+21] F. Uribe et al. "Cross-Entropy-Based Importance Sampling with Failure-Informed Dimension Reduction for Rare Event Simulation". In: *SIAM/ASA Journal on Uncertainty Quantification* 9.2 (2021), pp. 818–847.