The Limitations of Control-Energy KL Bounds for Noise-Reactive Safety Filters

Samuel Scheffler[*]
Purdue University

Roy Vaid[*]
University of California, Santa Barbara

Abstract: A safety filter in stochastic sampling may observe the current Gaussian increment before choosing its control. We show that the terminal-KL control-energy bound of Gadginmath et al. (arXiv:2602.21429v3, Theorem 4.3) does not generally hold for such noise-reactive controls. A one-step minimum-norm counterexample is safe in every realization and has finite expected control energy but infinite terminal KL; further examples cover bounded targets and a strict step-size inequality, including an absolutely continuous terminal law with finite KL exceeding the energy bound. We identify the conditional-law error and prove the bound for predictable controls chosen before the current noise, although these cannot guarantee bounded targets with probability one when the final Gaussian increment is nondegenerate and the pre-noise mean is finite. For noise-reactive maps satisfying the stated change-of-variables assumptions, KL also includes noise-correlation and log-Jacobian terms, while a conditioning identity gives a distribution-level benchmark for the cost of safety enforcement. These results separate constraint satisfaction from distributional guarantees in samplers of robot action sequences.

Keywords: Safe generative sampling, Barrier function, Relative entropy

1 The claim and the controller's information

Constraint-releasing barrier functions relax the target constraint at initialization and tighten it during sampling [1]. All source theorem, proof, and algorithm references concern arXiv:2602.21429v3 (31 July 2026).

The source sampler [1, Section 4.2] is

$$X_{k-1} = X_k - [f(X_k, t_k) + U_k]\Delta t + g_k\sqrt{\Delta t}\, Z_k, \qquad Z_k \sim \mathcal{N}(0, I), \tag{1}$$

Here $X_k \in \mathbb{R}^d$, $\Delta t = T/K > 0$, $t_k = k\Delta t$, and $g_k = g(t_k) > 0$. History $H_k = (X_K, \dots, X_k)$ precedes the independent noise Z_k. A noise-reactive filter chooses $U_k = u_k(H_k, Z_k, t_k)$ after observing Z_k; a predictable filter chooses $U_k = v_k(H_k, t_k)$ beforehand. Including initialization X_K in H_k makes explicit the constriction's dependence hidden in the source notation $u_k(X_k, Z_k, t_k)$.

The uncontrolled sampler has the same initialization and zero control. For guided and uncontrolled terminal laws Q_0 and P_0, source Theorem 4.3 claims

$$\mathrm{D}_{\mathrm{KL}}(Q_0 \| P_0) \le \mathcal{E}, \qquad \mathcal{E} := \frac{1}{2}\mathbb{E}_Q\left[\sum_{k=1}^{K} \frac{\Delta t}{g_k^2}\|U_k\|^2\right]. \tag{2}$$

KL is extended-valued: it is infinite without absolute continuity of the first law relative to the second.

Robot-action interpretation. For a fixed observation, X_0 in (1) can represent a completed action sequence, with nominal and filtered laws P_0 and Q_0. Here k indexes sampling steps, not physical time. A correction chosen after observing the current Gaussian increment is noise-reactive. Our

[*]Equal contribution.

CoRL 2026 Workshop on the Science of Physical AI Safety (SPAIS).

counterexamples show that minimizing squared intervention, even while satisfying sample constraints, need not certify (2).

For fixed initialization, define $\widetilde{h}(x,t) = h(x) + \epsilon(X_K, t)$. The exact barrier condition, with $\alpha > 0$, is

$$\widetilde{h}(X_{k-1}, t_{k-1}) \geq (1 - \alpha\Delta t)\widetilde{h}(X_k, t_k). \tag{3}$$

Source Definition 1 requires continuous differentiability, initial feasibility $\epsilon(a,T) \geq \max\{0, -h(a)\}$, target recovery $\epsilon(a,0) = 0$, and monotonicity $\partial_t \epsilon(a,t) \geq 0$. Under $\alpha\Delta t \leq 1$, source Theorem 4.2 proves terminal constraint satisfaction by induction from exact enforcement of (3) at each step [1]. All examples below satisfy the exact constraint.

2 Counterexamples with boundary atoms

All examples are scalar and one-step: $K = T = \Delta t = g_1 = 1$, $f(x,1) = x$, with independent $X_1, Z \sim \mathcal{N}(0,1)$. The drift cancels the state, giving nominal output Z and guided output $Y = Z - U$.

Proposition 1 (Infinite KL for the stated QP). *For $h(x) = x$, $C = [0,\infty)$, $\epsilon(a,t) = t\sqrt{1+a^2}$, and $\alpha = 1$, source Algorithm 1 [1], including its linearized constraint, has a unique minimum-norm control with safe output:*

$$\begin{gathered} U^*(Z) = \min(Z, 0), \qquad Y = \max(Z, 0), \qquad \tfrac{1}{2}\mathbb{E}[(U^*)^2] = \tfrac{1}{4}, \\ \mathrm{D}_{\mathrm{KL}}(\mathcal{L}(Y) \| \mathcal{N}(0,1)) = +\infty. \end{gathered} \tag{4}$$

Proof. The smooth constriction vanishes at terminal time, has nonnegative time derivative, and is initially feasible because $\sqrt{1+a^2} \geq \max\{0, -a\}$. Set $s = \sqrt{1+X_1^2}$. The linearized constraint is $X_1 + s + (-X_1 - u + Z) - s = Z - u \geq 0$. At fixed initialization the barrier is affine in state and time, so this is also the exact constraint. Minimizing u^2 gives $\min(Z, 0)$. The safe output places mass $1/2$ at zero, a Gaussian-null set, yielding infinite KL. Symmetry gives $\mathbb{E}[Z^2 \mathbf{1}_{\{Z<0\}}]/2 = 1/4$.

By comparison, conditioning the Gaussian to be nonnegative gives a safe law with finite KL $\log 2$.

Bounded target and strict step-size inequality. Set

$$C = [-1, 1], \qquad h(x) = 1 - x^2, \qquad \epsilon(a,t) = ta^2, \qquad \alpha = \tfrac{1}{4}. \tag{5}$$

Here $h'(\pm 1) \neq 0$, Definition 1 holds, the tube is bounded, and $\widetilde{h}(X_1, 1) = 1$. For this nonlinear barrier we solve the exact constraint, rather than Algorithm 1's linearized QP constraint:

$$1 - Y^2 \geq \tfrac{3}{4}, \qquad \text{equivalently} \qquad |Y| \leq r := \tfrac{1}{2}. \tag{6}$$

The minimum-norm control projects the Gaussian increment onto this interval:

$$U^*(Z) = Z - \mathrm{clip}(Z, -r, r), \qquad Y = \mathrm{clip}(Z, -r, r). \tag{7}$$

Each boundary atom has mass $1 - \Phi(r) > 0$, giving infinite KL but finite control energy. With standard normal density ϕ and CDF Φ, the latter is

$$\frac{1}{2}\mathbb{E}[(U^*)^2] = \int_r^\infty (z-r)^2 \phi(z)\, dz = (1+r^2)[1 - \Phi(r)] - r\phi(r) < \infty. \tag{8}$$

Each atom has mass approximately 0.30854; expected control energy is approximately 0.20964. Both controls are globally Lipschitz and L^2, though unbounded in Gaussian noise; source Theorem 4.3 imposes no uniform control bound. See Everink et al. [2] for projection-induced atoms.

3 Finite terminal KL can exceed control energy

Retaining Algorithm 1 with $\alpha\Delta t < 1$, take $h(x) = x$ and the smooth initialization-dependent constriction

$$\epsilon(a,t) = t\big(\sqrt{1+a^2} - a\big), \qquad \alpha = \tfrac{1}{2}, \qquad B = \tfrac{1}{2}\sqrt{1+X_1^2}. \tag{9}$$

Definition 1 holds. At fixed initialization the barrier is affine in state and time, so exact and linearized constraints coincide. Since $\widetilde{h}(X_1, 1) = \sqrt{1 + X_1^2}$, the constraint is $Z - U \geq B$, yielding

$$U^* = \min(Z - B, 0), \qquad Y = \max(Z, B). \tag{10}$$

The control is globally Lipschitz in initialization and noise and safe for every realization.

Proposition 2 (Finite terminal KL can exceed control energy). *The terminal law $Q_0 = \mathcal{L}(Y)$ is absolutely continuous with finite KL relative to $P_0 = \mathcal{N}(0, 1)$, but*

$$\mathrm{D}_{\mathrm{KL}}(Q_0 \| P_0) \geq -\log[1 - \Phi(1/2)] > 1.17, \qquad \tfrac{1}{2}\mathbb{E}[(U^*)^2] \leq \tfrac{3}{4}. \tag{11}$$

Proof. The output is at least one-half. Let $S = [1/2, \infty)$. The conditioning formula $\mathrm{D}_{\mathrm{KL}}(Q_0 \| P_0) = \mathrm{D}_{\mathrm{KL}}(Q_0 \| P_0(\cdot \mid S)) - \log P_0(S)$ for the KL divergence gives the lower bound, while independence together with the second moment of the threshold, $\mathbb{E}B^2 = 1/2$, leads to the upper bound on control energy, $\mathbb{E}[(U^*)^2]/2 \leq \mathbb{E}[(Z - B)^2]/2 = (1 + 1/2)/2 = 3/4$. Absolute continuity and finiteness of KL: we first consider the distribution function of the threshold on the domain where it is larger than one-half. This is $G(y) = 2\Phi(s(y)) - 1$, where $s(y) = \sqrt{4y^2 - 1}$ and $y > 1/2$. By independence, we get the distribution function of the output, $\Pr(Y \leq y) = \Phi(y)G(y)$, and its density.

$$q(y) = \phi(y)G(y) + \Phi(y)\frac{8y\phi(s(y))}{s(y)}. \tag{12}$$

Its distribution function is zero until one-half and continuous at one-half which rules out atom at one-half. Close to this point, the density exhibits an inverse-square-root singularity $q(y) = O((y - 1/2)^{-1/2})$ which renders the integrand of the KL integrable. In the tail, the ratio of the densities $q(y)/\phi(y)$ converges to one and remains bounded. Here $\phi(s(y))/\phi(y) = \exp((1 - 3y^2)/2)$.

Path measure still has the singular component on the graph, $\Gamma = \{(a, y) : y = \frac{1}{2}\sqrt{1 + a^2}\}$: the guided state/output pair falls on the graph with positive probability $\mathbb{E}[\Phi(B)] > 0$ (that equals zero under the uncontrolled law).

4 What constitutes a valid distributional certificate

Conditional-law error. The source proof conditions on state and noise yet treats both transition laws as Gaussians with equal nonzero covariance. With initialization explicit, conditioning on (H_k, Z_k) instead makes the next state deterministic: the laws are Dirac measures, with infinite divergence when their locations differ. Conditioning only on history gives a Gaussian nominal transition, but not necessarily a Gaussian controlled transition.

Proposition 3 (Predictable-control repair). *Assume identical initial laws, $g_k > 0$, and finite measurable controls $U_k = v_k(H_k, t_k)$ chosen before independent Z_k. For finite expected energy and guided/nominal laws $Q_{\mathrm{path}}, P_{\mathrm{path}}$ of $(X_K, \ldots, X_0)$,*

$$\mathrm{D}_{\mathrm{KL}}(Q_{\mathrm{path}} \| P_{\mathrm{path}}) = \mathcal{E}, \qquad \mathrm{D}_{\mathrm{KL}}(Q_0 \| P_0) \leq \mathcal{E}. \tag{13}$$

Proof. At the same history, both kernels are Gaussian with covariance $g_k^2 \Delta t I$ and mean difference $-\Delta t\, v_k(H_k, t_k)$. Their KL is $\Delta t \|v_k(H_k, t_k)\|^2/(2g_k^2)$. The entropy chain rule [3], averaged under the guided law with zero initial-law contribution, gives equality. Terminal projection and data processing give the inequality.

Continuous-time entropy-energy relationships rely upon the adaptation assumption together with change-of-measure and invertibility assumptions for the equality case in [4].

Safety limitation of predictability. With a nondegenerate final Gaussian increment and finite pre-noise mean, the final conditional law has full support on $\mathbb{R}^d$. Every bounded target therefore has positive conditional violation probability, which remains positive after averaging. This rules out probability-one satisfaction under these assumptions; a noiseless final step is exempt.

Noise-reactive maps. At fixed pre-noise history, write the nominal output as $m + \sigma Z$, where $Z \sim \mathcal{N}(0, I_d)$, $\sigma = g\sqrt{\Delta t}$, $b = \sqrt{\Delta t}/g$, and $F(z) = z - bu(z)$ suppresses fixed history/time. If F is a continuously differentiable diffeomorphism onto an open image and all three terms are integrable, change of variables gives

$$\begin{aligned} &\mathrm{D}_{\mathrm{KL}}\big(\mathcal{L}(m + \sigma F(Z)) \| \mathcal{N}(m, \sigma^2 I)\big) \\ &\quad = \mathbb{E}\left[\frac{\|F(Z)\|^2 - \|Z\|^2}{2} - \log|\det DF(Z)|\right] \\ &\quad = \frac{b^2}{2}\mathbb{E}\|u(Z)\|^2 - b\mathbb{E}[Z^\top u(Z)] - \mathbb{E}\log|\det(I - bDu(Z))|. \end{aligned} \tag{14}$$

Scalar scaling $F(z) = cz$, $c > 0$, normally makes KL different from the control energy, while constant shift makes it equal. The difference in the scaling case is $d(c - 1 - \log c)$. For jointly measurable transitions fulfilling these assumptions, with integrability over histories, one uses (14) conditionally and applies the path-space chain rule. The above projection maps do not meet the requirement of diffeomorphism.

Selecting a distributional certificate. First specify whether the correction precedes or follows current noise: Proposition 3 certifies energy only under its predictable-control assumptions. For regular noise-reactive maps, calculate or bound the full conditional KL in (14), including correlation and Jacobian terms; smoothness alone does not bound this cost. Check terminal absolute continuity: positive filtered mass on a nominal-null set gives infinite KL (Section 2). Proposition 2 shows both that finite terminal KL does not imply (2) and that singular conditional transitions need not give a singular terminal law.

5 Implications for safety-filter analysis

The counterexamples invalidate source Theorem 4.3's general noise-reactive energy certificate, including its per-step KL interpretation, without invalidating Theorem 4.2's exact-safety induction. In Propositions 1 and 2, linearized and exact constraints coincide, so approximation does not explain the failure. Proposition 3 and (14) supply certificates under their respective assumptions.

A KL benchmark for safety enforcement. For measurable C with $P_0(C) > 0$ and $Q_0(C) = 1$, the conditioning identity separates target restriction from additional distortion:

$$\mathrm{D}_{\mathrm{KL}}(Q_0 \| P_0) = \mathrm{D}_{\mathrm{KL}}\big(Q_0 \| P_0(\cdot \mid C)\big) - \log P_0(C).$$

Nonnegativity gives the lower bound $-\log P_0(C)$, attained by $P_0(\cdot \mid C)$. The remaining KL measures departure from that conditioned nominal law. This distribution-level benchmark implies neither efficient conditional sampling nor realization by the given filter; changing reference law cannot remove a failure of absolute continuity.

Evaluating filtered robot-action samplers. A small-correction objective alone does not establish (2). Report the information pattern, check constraints separately, and justify the distributional certificate's assumptions. Predictable controls retain Section 4's bounded-target limitation with nondegenerate final noise and finite pre-noise mean. These are sampling-distribution results, not guarantees of closed-loop physical safety, task performance, or software correctness; no empirical KL calculations are claimed.

References

[1] D. Gadginmath, A. Allibhoy, and F. Pasqualetti. Provably safe generative sampling with constricting barrier functions. arXiv:2602.21429v3, July 2026. URL https://arxiv.org/abs/2602.21429v3. Version dated 31 July 2026; Theorems 4.2–4.3 and Section 4.3.

[2] J. M. Everink, Y. Dong, and M. S. Andersen. Bayesian inference with projected densities. arXiv:2209.12481, 2022. URL https://arxiv.org/abs/2209.12481.

[3] T. M. Cover and J. A. Thomas. *Elements of Information Theory*. Wiley, 2nd edition, 2006. doi:10.1002/047174882X.

[4] A. S. Üstünel. Entropy, invertibility and variational calculus of the adapted shifts on Wiener space. arXiv:0903.3891, 2009. URL https://arxiv.org/abs/0903.3891.