Canonical joint energy-based model on CIFAR-10: failure modes and practical indistinguishability of predictor-corrector and SGLD samplers

D. O. Knopov
National University of Kyiv-Mohyla Academy, Kyiv, Ukraine
d.knopov@ukma.edu.ua

We address the problem of CIFAR-10 image classification and out-of-distribution (OOD) detection using a Joint Energy-Based Model (JEM) [1], in which a classifier $f_\theta\colon \mathbb{R}^D \to \mathbb{R}^K$ is reinterpreted as a parameterisation of an energy function and an associated Gibbs density

$$E_\theta(x) = -\log \sum_y \exp f_\theta(x)[y], \qquad p_\theta(x) \propto \exp(-E_\theta(x)), \tag{1}$$

so that the resulting model simultaneously serves as a classifier and a generative model. Canonical JEM training employs stochastic gradient Langevin dynamics (SGLD) [2] as a numerical discretisation of the equation

$$dx = \nabla \log p_\theta(x)\, dt + \sqrt{2}\, dW_t \tag{2}$$

with stationary distribution p_θ. A theoretically more general alternative is the predictor-corrector (PC) sampler [3], which integrates the time-reversed stochastic differential equation (SDE)

$$dx = -g_t^2 \nabla \log p_t(x)\, dt + g_t\, d\bar{W}_t \tag{3}$$

along an annealing schedule for σ_t, alternating a predictor step (Euler–Maruyama) with a Langevin corrector step. Here σ_t denotes the marginal noise scale and g_t the diffusion coefficient. For the variance-exploding SDE [3], $g_t^2 = d\sigma_t^2/dt$. In canonical JEM the noise scale is fixed ($\sigma_{\mathrm{SGLD}} = 0.01$), which, as shown below, formally reduces PC to SGLD. This is consistent with a known limitation of the original PC formalism: on the static density (1), Langevin dynamics without noise annealing does not improve the mixing time [4].

Remark 1. Under a constant diffusion coefficient $g_t \equiv g_0$, a fixed noise scale $\sigma_t \equiv \sigma_{\mathrm{SGLD}}$ and a static target $\nabla \log p_t \equiv \nabla \log p_\theta$, the reverse-time SDE (3) reduces — via a time change and a constant rescaling of the noise term — to (2), and therefore belongs to the same Langevin family, with a stationary distribution of the form $\exp(-\beta E_\theta)$ for some $\beta > 0$. Consequently, the Euler–Maruyama predictor step for the reverse-time SDE (3) coincides, up to a constant factor, with an SGLD step (2), and the Langevin corrector at the same σ_{SGLD} contributes no additional dynamics. This reduction is formulated at the level of continuous-time SDEs; whether it carries over to discrete training with a replay buffer and gradient updates of the model parameters is a separate empirical question, addressed below.

Across two pairs of runs (SGLD and PC, with seeds 42 and 123), WideResNet-28-10 attains an accuracy of 92.88% on CIFAR-10 (canonical value 92.9% in [1]), while the between-sampler difference in AUROC (area under the ROC curve) for static OOD detection does not exceed 0.007 across five standard datasets (SVHN, CIFAR-100, DTD, LSUN-R, iSUN). All four runs independently terminate in catastrophic divergence between epochs 115 and 131 via two related mechanisms: a high-energy outlier in the replay buffer [1, Appendix H.3] and a concomitant inversion of the OOD energy landscape ($E_\theta(\mathrm{SVHN}) < E_\theta(\mathrm{CIFAR\text{-}10})$). Eliminating these failure modes requires structural changes to the training dynamics [5, 6].

[1] Grathwohl W., Wang K.-C., Jacobsen J.-H., Duvenaud D., Norouzi M., Swersky K., Your classifier is secretly an energy-based model and you should treat it like one, *ICLR* (2020), arXiv:1912.03263.

[2] Welling M., Teh Y. W., Bayesian learning via stochastic gradient Langevin dynamics, *ICML* (2011), 681–688.

[3] Song Y., Sohl-Dickstein J., Kingma D. P., Kumar A., Ermon S., Poole B., Score-based generative modeling through stochastic differential equations, *ICLR* (2021), arXiv:2011.13456.

[4] Bradley A., Nakkiran P., Classifier-free guidance is a predictor-corrector, *Trans. Mach. Learn. Res.* (2025), arXiv:2408.09000.

[5] Yang X., Su Q., Ji S., Towards bridging the performance gaps of joint energy-based models, *CVPR* (2023), 15732–15741.

[6] Yin X., Zhang C., Steele J., Shavit N., Wang T. T., Scalable energy-based models via adversarial training: unifying discrimination and generation, *ICLR* (2026), arXiv:2510.13872.