Algorithm 1 Adversarial Negative Mining (ANM) for IRED

Input: Problem dist. $p_D(x, y)$, EBM $E_\theta(\cdot)$, noise schedules $\{\sigma_k\}$, step sizes $\{\eta_k\}$, adversarial strength ε, # adversarial steps T_{adv}, landscapes $k = 1, \ldots, K$.

Output: Adversarial negatives $\tilde{y}_i^-$ used in $\mathcal{L}_{\text{Contrast}}$.

while not converged **do**

▷ **Sample batch and landscape**

$x_i, y_i \sim p_D, \quad \epsilon \sim \mathcal{N}(0, I), \quad k \sim \{1, \ldots, K\}$

▷ **Initialize diffusion-corrupted label**

$\tilde{y}_i^{(0)} \leftarrow \sqrt{1 - \sigma_k^2}\, y_i + \sigma_k \epsilon$

$y_i^{\text{orig}} \leftarrow \tilde{y}_i^{(0)}$ ▷ anchor for distance penalty

▷ **Adversarial refinement by energy descent**

for $s = 0$ to $T_{\text{adv}} - 1$ **do**

(1) Energy gradient at current point

$g^{(s)} \leftarrow \nabla_y E_\theta(x_i, \tilde{y}_i^{(s)}, k)$

(2) Distance penalty from start

$d^{(s)} \leftarrow \|\tilde{y}_i^{(s)} - y_i^{\text{orig}}\|_2^2$

$\lambda_{\text{pen}}^{(s)} \leftarrow \varepsilon \cdot \text{clip}\left(\frac{1}{d^{(s)} + 10^{-6}},\, 0.1,\, 2.0\right)$

$h^{(s)} \leftarrow \nabla_y d^{(s)}$

(3) Modified gradient (energy − distance penalty)

$\hat{g}^{(s)} \leftarrow g^{(s)} - \lambda_{\text{pen}}^{(s)} h^{(s)}$

(4) Decayed step size within landscape k

$\eta^{(s)} \leftarrow \eta_k \alpha^s$ ▷ $0 < \alpha < 1$ (e.g. $\alpha = 0.7$)

(5) Gradient step and projection back to feasible set

$\tilde{y}_i^{(s+1/2)} \leftarrow \tilde{y}_i^{(s)} - \eta^{(s)} \hat{g}^{(s)}$

$\tilde{y}_i^{(s+1)} \leftarrow \text{Proj}_{\mathcal{C}(x_i, k)}\left(\tilde{y}_i^{(s+1/2)}\right)$

▷ **Use adversarial negative in contrastive loss**

$\tilde{y}_i^- \leftarrow \tilde{y}_i^{(T_{\text{adv}})}$

$E_i^+ \leftarrow E_\theta(x_i, \tilde{y}_i, k)$

$E_i^- \leftarrow E_\theta(x_i, \tilde{y}_i^-, k)$

$\mathcal{L}_{\text{Contrast}} \leftarrow -\log \frac{e^{-E_i^+}}{e^{-E_i^+} + e^{-E_i^-}}$