Generative Diffusion Model Unlearning: A Constrained Optimization View

Dongsheng Ding
dongshed@utk.edu

THE UNIVERSITY OF TENNESSEE KNOXVILLE

2026 INFORMS Annual Meeting; Nov. 2, 2026

Motivating application

- FACE GENERATION: RIGHT TO BE FORGOTTEN



forget retain retain

CelebA-HQ

CHALLENGE: forget vs. retain – conflicting objectives

Context

- SUCCESS STORIES OF GENAI

  Gemini, Claude, GPT Image, Nano Banana

- LESSONS LEARNED
  - learning complex distributions

    Multimodal distributions; high-dimensional
  - models reproduce training data; generate copyrighted content

    Security; copyright

- WHAT NOW ?
  - learning under requirements; constrained optimization
  - GenAI unlearning

OBJECTIVE

Steer a model that
stays close to the pretrained model
subject to separation from the unlearning distributions

Unlearning in diffusion models

Pretrained model: q – to retain

Unlearning distributions: q_{u}^i for $i = 1, \ldots, m$ – to forget

setting	goal	retain q	forget q_{u}^i
CONCEPT UNLEARNING	remove a concept, e.g., 'horse' from 'cowboy'	$p(\cdot \mid \text{'cowboy'})$	$p(\cdot \mid \text{'horse'})$
DATA UNLEARNING	remove specific (memorized) training samples	CelebA-HQ faces to keep	ith face image to forget

Gandikota et al., ICCV '23; Kumari et al., ICCV '23 (concept unlearning)

Alberti et al., ICLR '25; Wu et al., CVPR '25 (data unlearning)

Prior art

- WEIGHTED-SUM UNLEARNING

$$\underset{\theta}{\text{minimize}} \quad \underbrace{\ell_{\mathsf{r}}(\theta)}_{\text{retain}} + w \underbrace{\ell_{\mathsf{f}}(\theta)}_{\text{forget}}$$

Alberti et al., ICLR '25; Feng et al., ICLR '25

- CONSTRAINED UNLEARNING

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & \ell_{\mathsf{r}}(\theta) \\ \text{subject to} \quad & \ell_{\mathsf{f}}(\theta) \leq \epsilon \end{aligned}$$

Feng et al., ICLR '25; Zhang et al., NeurIPS '24; Wu et al., CVPR '25

This talk: optimality in distribution space

A CONSTRAINED OPTIMIZATION VIEW

Three constrained unlearning problems

KRD†, ICML '26

problem	concept	data	likelihood
NEEDS	models of q, q_{u}^i 'cowboy', 'horse'	samples of q_{u}^i memorized images	models of q, q_{u}^i likelihood of 'horse'
MINIMIZE SUBJECT TO $\forall i$	$D_{\mathsf{KL}}\left(p \,\|\, q\right)$ $D_{\mathsf{KL}}\left(p \,\|\, q_{\mathsf{u}}^i\right) \geq b_i$	$D_{\mathsf{KL}}\left(q \,\|\, p\right)$ $D_{\mathsf{KL}}\left(q_{\mathsf{u}}^i \,\|\, p\right) \geq b_i$	$D_{\mathsf{KL}}\left(p \,\|\, q\right)$ $\mathbb{E}_{x\sim p}\left[\, q_{\mathsf{u}}^i(x) \,\right] \leq \epsilon_i$
GENERALIZES	concept erasing Gandikota et al., '23	gradient ascent Alberti et al., '25	new

Thresholds: b_i**: separation level;** ϵ_i**: likelihood level**

Challenge: **nonconvex constraints in concept and data unlearning**

DUALITY THROUGH CONCEPT UNLEARNING

Lagrangian duality

Lagrangian: $L(p, \lambda) \;=\; D_{\mathsf{KL}}\left(p \,\|\, q\right) \;+\; \sum_{i=1}^{m} \lambda_i \left(b_i - D_{\mathsf{KL}}\left(p \,\|\, q_{\mathsf{u}}^{i}\right)\right)$

Dual problem: $D^{\star} \;=\; \underset{\lambda \geq 0}{\text{maximize}} \left\{ D(\lambda) \;:=\; \underset{p}{\text{minimize}} \; L(p, \lambda) \right\}$

Weak duality: $D^{\star} \;\leq\; P^{\star}$ always; gap possible if nonconvex

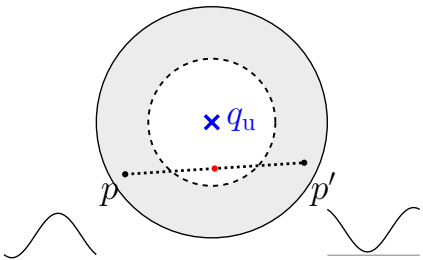
Theorem (informal)

★ Concept unlearning is **strongly dual**: $P^{\star} = D^{\star}$, despite nonconvexity

Solve the dual:	**concave** $D(\lambda)$ **in only** m **variables**
Recover a model:	$p^{\star} \in \operatorname{argmin}_p L(p; \lambda^{\star})$ **in closed form**
Weights:	**optimal** $\lambda_i^{\star}$

Why strong duality?

nonconvex superlevel set

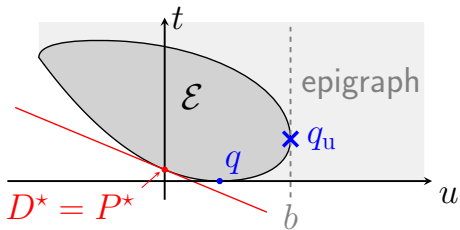


a 2-D slice of densities on $[0, 1]$
shaded: $D_{\mathrm{KL}}(p \,\|\, q_{\mathrm{u}}) \geq b$; midpoint not

$p \mapsto (u, t)$

$\longrightarrow$

convex epigraph of image $\mathcal{E}$



u: constraint slack; t: objective
q: pretrained model (infeasible)

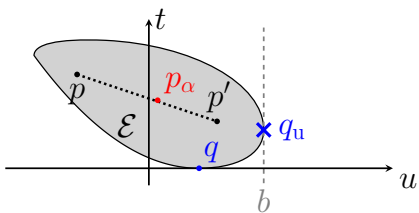
nonconvex in p, but convex epigraph of $\mathcal{E}$ $\implies$ strong duality

Lyapunov's convexity theorem

Lyapunov, 1940 (finite KLs)

range $\{v(A)\}_A$ of a finite, non-atomic vector measure v is convex

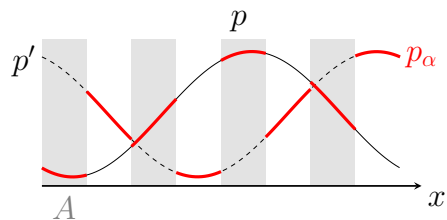
values average



$v(p_\alpha) = \alpha\, v(p) + (1-\alpha)\, v(p')$

$\longleftrightarrow$

splice two models



$p_\alpha = p$ on A (gray); p' elsewhere

$\mathcal{E}$ **convex** $\Longrightarrow$ **epigraph convex**

Optimal unlearning target

$$p^\star \propto \frac{(q)^{1/(1-\mathbf{1}^\top \lambda^\star)}}{\prod_i (q_{\mathsf{u}}^i)^{\hat{\lambda}_i^\star}}$$

$$\hat{\lambda}_i^\star = \lambda_i^\star/(1-\mathbf{1}^\top\lambda^\star), \quad 0 \le \mathbf{1}^\top\lambda^\star < 1$$

- ⋆ down-weight regions where q_{u}^i is large
- ⋆ generalize existing unlearning/negation targets

Gandikota et al., ICCV '23; Du et al., ICML '23

Optimal dual variables $\lambda^\star$ = optimal weights

From distributions to diffusion models

Score function: $s_p(x_t, t)$ induces $p_0(\cdot\,; s_p)$

CONSTRAINED SCORE OPTIMIZATION

$$\begin{aligned} &\underset{s_p \in \mathcal{S}}{\text{minimize}} && D_{\mathsf{KL}}\left(p_0(\cdot\,; s_p) \,\|\, q_0\right) \\ &\text{subject to} && D_{\mathsf{KL}}\left(p_0(\cdot\,; s_p) \,\|\, q^i_{\mathsf{u},0}\right) \;\geq\; b_i \quad \text{for } i = 1, \ldots, m \end{aligned}$$

Theorem (informal)

★ If score class $\mathcal{S}$ is expressive, strong duality holds in score space

Lyapunov's convexity theorem on path space

Primal-dual method

- ITERATION k

$$\textbf{primal:} \quad s_{k+1} \in \operatorname{argmin}_{s_p} L(s_p, \lambda_k)$$

$$\textbf{dual:} \quad \lambda_{i,k+1} = \left[\lambda_{i,k} + \eta \left(b_i - D_{\mathsf{KL}}(p_0(\cdot\,; s_{k+1}) \,\|\, q_{\mathsf{u}}^i) \right) \right]_+$$

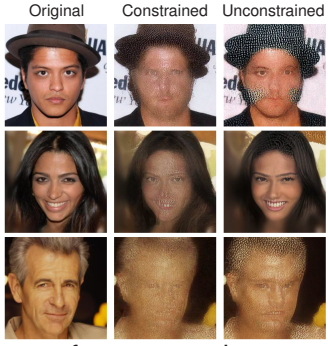
$\star$ $D_{\mathsf{KL}}(p_0(\cdot\,; s_p) \,\|\, q_0(\cdot\,; s_q)) \approx \sum_t \omega_t \, \mathbb{E} \, \|s_p(x_t, t) - s_q(x_t, t)\|^2$

Dual variables adapt weights to each unlearning target

EMPIRICAL RESULTS

Data unlearning: forward KL

DDPM on CelebA-HQ



forget samples

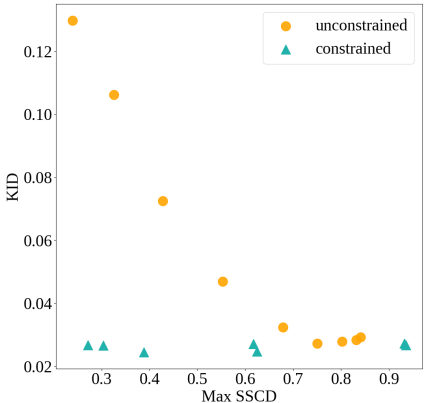


retain samples

- forget three randomly chosen training images
- unconstrained baseline: equal weights on all forget samples

Constraints forget the samples & keep other generations

Data unlearning: retention vs. unlearning



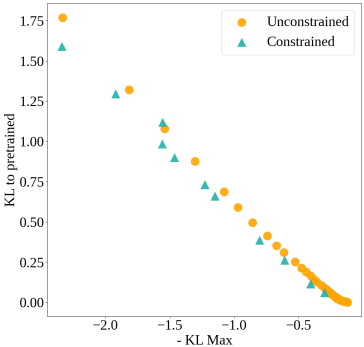
- KID: deviation from pretrained
  lower is better
- max SSCD: similarity to forget samples
  lower is better

Large weights on hard samples, small weights on easy ones

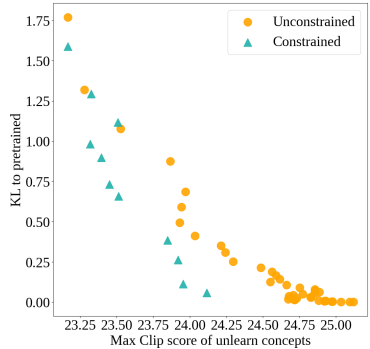
Constraints deviate less at the same level of unlearning

Concept unlearning: reverse KL

Stable Diffusion v1.4



worst-case KL constraint violation



max CLIP score of unlearning concepts

- ⋆ retain ‘dog’; unlearn ‘golden retriever’, ‘german shepherd’, ...

Constraints reduce model deviation at the same level of unlearning

Likelihood unlearning: cowboy without horse

Stable Diffusion v1.4



unlearn ‘horse’

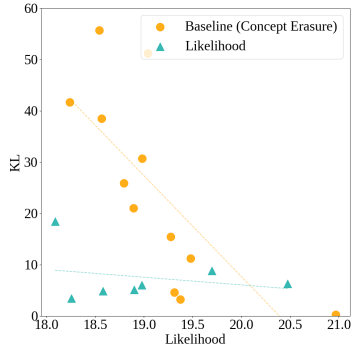


retain ‘cowboy’

- baseline: concept erasing (Gandikota et al., 2023)

Constraints keep ‘cowboy’ closer to pretrained

Likelihood unlearning: retention vs. unlearning



- KL to pretrained ‘cowboy’ vs. likelihood of ‘horse’

Unlearning with small deviation from what we retain

Summary & outlook

KR**D**†, ICML '26

- UNLEARNING AS CONSTRAINED OPTIMIZATION
  - reverse KL, forward KL, and likelihood constraints
  - strong duality despite nonconvexity
  - better retention-unlearning trade-offs
- OPEN CHALLENGES
  - primal-dual guarantees with score networks
  - robustness to relearning and fine-tuning
  - discrete outputs (e.g., LLMs, graphs), where non-atomicity fails

Thank you for your attention.