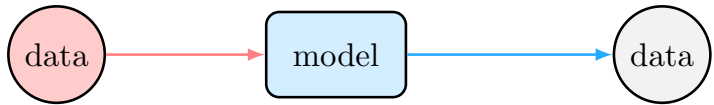
Controllable Generative Diffusion Models: A Constrained Optimization View

Dongsheng Ding
dongshed@utk.edu

THE UNIVERSITY OF TENNESSEE KNOXVILLE

UC San Diego; Mar. 12, 2026

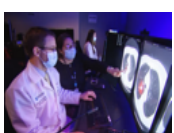
The reality of deep generative modeling



goal – likelihood

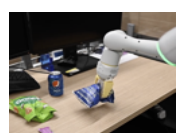
REQUIREMENTS

fairness



safety

robustness



harmlessness

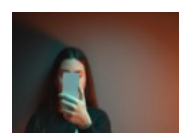


Image sources: NBC, Waymo, WIRED, Stanford HAI

The risk of generative models

Humans Are Biased. Generative AI Is Even Worse

Bloomberg, JUN 9, 2023

When AI-Powered Humanoid Robots Make Bad Choices

AI Business, JAN 5, 2026

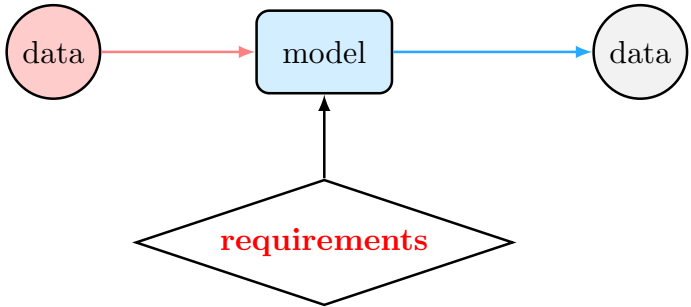
Their teenage sons died by suicide. Now, they are sounding an alarm about AI chatbots

NPR, SEP 19, 2025

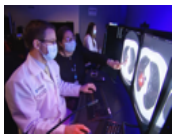
AI chatbots might be sabotaging women by advising them to ask for lower salaries, study says

New York Post, JUL 29, 2025

Requirement-driven deep generative modeling



fairness



safety

robustness

harmlessness

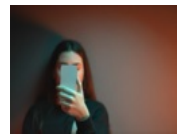


Image sources: NBC, Waymo, WIRED, Stanford HAI

Motivating application: Healthcare

- ONCOLOGY IMAGING

National Cancer Institute

$$\underset{\text{model}}{\text{minimize}} \quad \text{negative likelihood}$$

$$\text{subject to} \quad \text{distance to minority distribution} \leq \text{threshold}$$

REAL-WORLD CHALLENGE

Constraint satisfaction

OBJECTIVE

Find a deep generative model that minimizes a performance metric subject to a constraint on another performance metric

Outline

- CONSTRAINED GENERATIVE MODELS
- PART I PRE-TRAINING OF DIFFUSION MODELS
  - ⋆ constrained diffusion models
  - ⋆ dual method & optimality
- PART II POST-TRAINING OF DIFFUSION MODELS
  - ⋆ alignment with constraints
  - ⋆ dual method & optimality
- EMPIRICAL STUDY
- SUMMARY & OUTLOOK

CONSTRAINED GENERATIVE MODELS

constrained distribution optimization

Generative modeling

DATA GENERATION

$\Longrightarrow$

data $x \sim q(\cdot)$

new data $x \sim p(\cdot)$

⋆ distribution optimization

$$\underset{p}{\text{minimize}} \quad D_{\text{diff}}\left(p; q\right)$$

D_{diff} – distribution mismatch

p – trainable distribution model

Generation requirements

■ UNBIASEDNESS

⋆ stereotypes

dishwashers

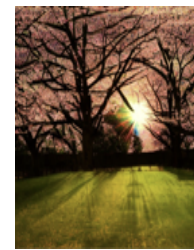
■ AESTHETICITY

⋆ artifacts

hand holding

■ COMPOSITIONALITY

⋆ compositional concepts



cherry tree AND sun dog

Generation has to be controllable satisfying requirements

Image sources: arXiv:2403.02726; arXiv: 2206.01714; arXiv: 2602.20951

Constrained distribution optimization

$$\underset{p}{\text{minimize}} \quad D_{\text{diff}}(p; q)$$

$$\text{subject to} \quad D_{\text{diff}}(p; q^i) \leq \epsilon_i \quad \text{for } i = 1, \ldots, m \quad \longrightarrow \quad \textbf{proximity}$$

$$\mathbb{E}_{x \sim p(\cdot)}[\, r_j(x) \,] \geq b_j \quad \text{for } j = 1, \ldots, n \quad \longrightarrow \quad \textbf{reward}$$

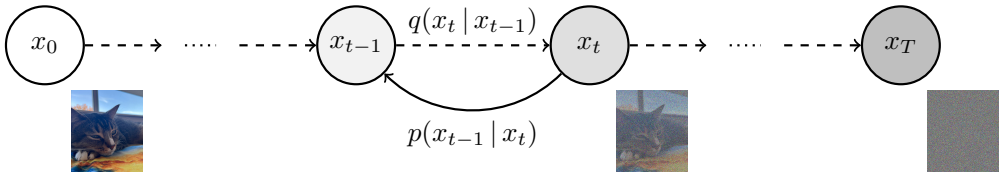
↘

functional constraints

⋆ limit the distribution space to inequality constraints

e.g., unbiased model, aestheticity, compositional model

Generative diffusion models



$q(x_t \mid x_{t-1})$ – forward process

$p(x_{t-1} \mid x_t)$ – backward process

■ DISTRIBUTION OPTIMIZATION

$$\underset{p}{\text{minimize}} \quad D_{\mathsf{KL}}\left(q(x_{0:T}) \,\|\, p(x_{0:T})\right)$$

Pre-training: Constrained diffusion models

KD†R, NeurIPS '24

Original data: $x_0 \sim q(\cdot)$

Desirable data: $x_0 \sim q^i(\cdot)$ for $i = 1, \ldots, m$

CONSTRAINED DISTRIBUTION OPTIMIZATION

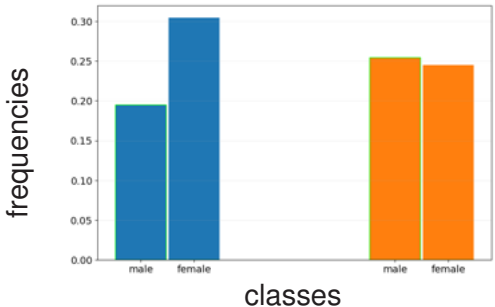
$$\underset{p}{\text{minimize}} \quad D_{\mathsf{KL}}\left(q(x_{0:T}) \,\|\, p(x_{0:T})\right)$$

$$\text{subject to} \quad D_{\mathsf{KL}}\left(q^i(x_{0:T}) \,\|\, p(x_{0:T})\right) \;\leq\; \epsilon_i \quad \text{for } i = 1, \ldots, m$$

- ⋆ minority-promoting distribution q^i

Constraints balance representativeness

KD†R, NeurIPS '24



unconstrained

constrained

- unconstrained model (▬)
- constrained model (▬)

Post-training: Alignment with constraints

KHD†R, NeurIPS '25

Pretrained model: $q(x_{0:T})$

Reward models: $r_i(x_0)$ for $i = 1, \ldots, m$

CONSTRAINED DISTRIBUTION OPTIMIZATION

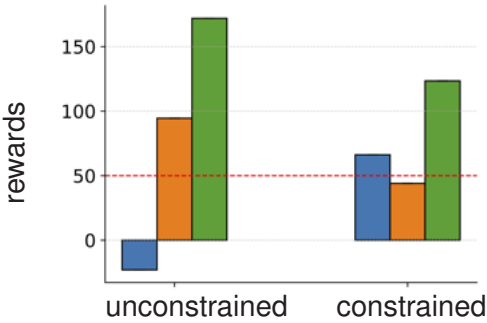
$$\underset{p}{\text{minimize}} \quad D_{\mathsf{KL}}\left(p(x_{0:T}) \,\|\, q(x_{0:T})\right)$$

$$\text{subject to} \quad \mathbb{E}_{x_0 \sim p(\cdot)}\left[\, r_i(x_0) \,\right] \;\geq\; b_i \;\; \text{for } i = 1, \ldots, m$$

- ⋆ aesthetic-constrained reward r_i

Constraints balance aesthetics

KH**D**†R, NeurIPS '25



- preference reward (▬)
- saturation reward (▬)
- local contrast reward (▬)

Lagrangian approach

LAGRANGIAN

$$L(p;\lambda) \;=\; D_{\text{diff}}(p;q) \;+\; \sum_{i=1}^{m} \lambda_i \left(D_{\text{diff}}(p;q^i) - \epsilon_i\right) \;+\; \sum_{j=1}^{n} \lambda_{m+j} \left(b_j - \mathbb{E}_{x\sim p(\cdot)}[\, r_j(x)\,]\right)$$

- penalize violation via dual variable $\lambda \geq 0$

DUAL FUNCTION

$$D(\lambda) \quad := \quad \underset{p}{\text{minimize}} \quad L(p;\lambda)$$

- concave function, regardless of the convexity in p

Dual problem

DUAL MAXIMIZATION

$$\underset{\lambda \geq 0}{\text{maximize}} \quad D(\lambda)$$

convex optimization

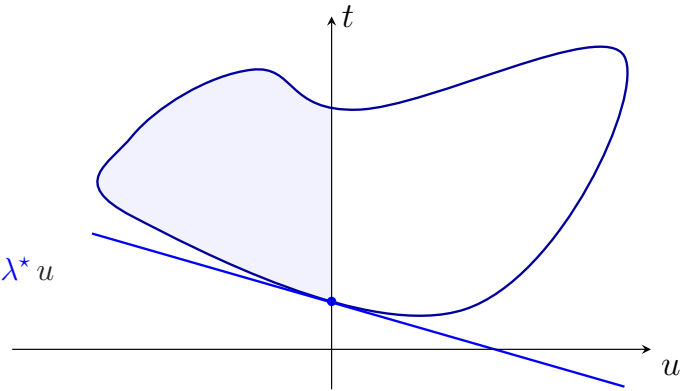
- ⋆ subgradient ascent finds an optimal dual variable $\lambda^\star$

Recovery of an optimal model $p^\star$

$$p^\star \quad \in \quad \underset{p}{\text{argmin}} \quad L(p; \lambda^\star)$$

Strong duality

- GEOMETRIC INTERPRETATION



$D^\star = t + \lambda^\star u$

image $\mathcal{E} = \{ (D_{\text{diff}}(p; q^i) - \epsilon_i, D_{\text{diff}}(p; q)) \mid p \}$

Optimal hyperplane touches $\mathcal{E}$ at an optimal model: $D^\star = D_{\text{diff}}(p^\star; q)$

Constrained parameter optimization

$$\underset{\theta}{\text{minimize}} \quad D_{\text{div}}\left(p_\theta; q\right)$$

$$\text{subject to} \quad D_{\text{div}}\left(p_\theta; q^i\right) \;\leq\; \epsilon_i \quad \text{for } i = 1, \ldots, m \qquad \longrightarrow \quad \textbf{proximity}$$

$$\mathbb{E}_{x \sim p_\theta(\cdot)}\left[\, r_j(x) \,\right] \;\geq\; b_j \quad \text{for } j = 1, \ldots, n \qquad \longrightarrow \quad \textbf{reward}$$

⋆ diffusion model parameter θ

CHALLENGE

Nonconvex constrained optimization → Lack of strong duality

Overview of our results

- PRE-TRAINING: CONSTRAINED DIFFUSION MODELS — KD†R, NeurIPS '24
  - ⋆ optimal model is a mixture distribution
  - ⋆ dual training & optimality gap
- POST-TRAINING: ALIGNMENT WITH CONSTRAINTS — KHD†R, NeurIPS '25
  - ⋆ optimal model is a reward weighted distribution
  - ⋆ dual training & optimality gap — objective & constraint

Dual training finds an optimal diffusion model,
up to an approximation error

PART I PRE-TRAINING OF DIFFUSION MODELS

constrained diffusion models

Constrained diffusion models

Original data: $x_0 \sim q(\cdot)$

Desirable data: $x_0 \sim q^i(\cdot)$ for $i = 1, \ldots, m$

CONSTRAINED DISTRIBUTION OPTIMIZATION

$$\begin{aligned} &\underset{p}{\text{minimize}} && D_{\mathsf{KL}}\left(q(x_{0:T}) \,\|\, p(x_{0:T})\right) \\ &\text{subject to} && D_{\mathsf{KL}}\left(q^i(x_{0:T}) \,\|\, p(x_{0:T})\right) \;\leq\; \epsilon_i \quad \text{for } i = 1, \ldots, m \end{aligned}$$

⋆ convex forward KL objective and convex forward KL constraints

Convex constrained distribution optimization $\rightarrow$ **Strong duality**

One pillar

- EVIDENCE LOWER BOUND (ELBO)

$$E(p; q) \;=\; \mathbb{E}_{q(x_0)} \mathbb{E}_{q(x_{1:T} \,|\, x_0)} \left[\log \frac{p(x_{0:T})}{q(x_{1:T} \,|\, x_0)} \right]$$

lower bound on log-likelihood

- ⋆ KL divergence decomposition

$$D_{\mathsf{KL}} \left(q(x_{0:T}) \,\|\, p(x_{0:T}) \right) \;=\; -\; E(p; q) \;+\; \mathbb{E}_{q(x_0)} \left[q(x_0) \right]$$

ELBO is linear in data distribution $q(x_0)$

KL divergence equals negative ELBO (up to a constant)

Lagrangian approach

LAGRANGIAN

$$L(p;\lambda) \;=\; D_{\mathsf{KL}}\left(q(x_{0:T}) \,\|\, p(x_{0:T})\right) \;+\; \sum_{i=1}^{m} \lambda_i \left(D_{\mathsf{KL}}\left(q^i(x_{0:T}) \,\|\, p(x_{0:T})\right) - \epsilon_i\right)$$

- ⋆ existence of an optimal dual variable $\lambda^\star$

OPTIMAL MODEL RECOVERY

$$\underset{p}{\text{minimize}} \;\; L(p;\lambda^\star) \qquad \Longleftrightarrow \qquad \underset{p}{\text{minimize}} \;\; -E(p;q) \;-\; \sum_{i=1}^{m} \lambda_i^\star \, E(p;q^i)$$

Optimal model

ELBO LINEARITY

$$-E(p;q) - \sum_{i=1}^{m} \lambda_i^\star E(p;q^i) \;=\; -E\left(p; q + \sum_{i=1}^{m} \lambda_i^\star q^i\right)$$

⋆ a mixture data distribution $q_{\text{mix}}^{(\lambda^\star)}(x_0) = \frac{1}{1+(\lambda^\star)^\top \mathbf{1}} \left(q(x_0) + \sum_{i=1}^{m} \lambda_i^\star q^i(x_0)\right)$

ELBO-KL DIVERGENCE EQUIVALENCE

$$\underset{p}{\text{minimize}} \;\; L(p;\lambda^\star) \quad \Longleftrightarrow \quad \underset{p}{\text{minimize}} \;\; D_{\text{KL}}\left(q_{\text{mix}}^{(\lambda^\star)}(x_{0:T}) \,\|\, p(x_{0:T})\right)$$

Optimal model is a mixture distribution $p^\star(x_{0:T}) = q_{\text{mix}}^{(\lambda^\star)}(x_{0:T})$

PART I PRE-TRAINING OF DIFFUSION MODELS

method & theory

Constrained diffusion models

Original data: $x_0 \sim q(\cdot)$

Desirable data: $x_0 \sim q^i(\cdot)$ for $i = 1, \dots, m$

- CONSTRAINED PARAMETER OPTIMIZATION

$$\begin{aligned} \underset{\theta}{\text{minimize}} \quad & -E(p_\theta; q) \\ \text{subject to} \quad & -E(p_\theta; q^i) \;\le\; b_i \quad \text{for } i = 1, \dots, m \end{aligned}$$

$$b_i = \epsilon_i - \mathbb{E}_{q^i(x_0)}\left[q^i(x_0)\right]$$

- ELBO serves as a score-matching loss

$$E(p_\theta; q) \;\approxeq\; -\mathbb{E}_{q(x_0),\, t,\, x_t}\left[\, \|\hat{s}_\theta(x_t, t) - \nabla \log q(x_t)\|^2 \,\right]$$

Nonconvex constrained parameter optimization $\rightarrow$ **Lack of strong duality**

Parametrized dual problem

- PARAMETRIZED DUAL FUNCTION

$$D_{\mathsf{p}}(\lambda) \;:=\; \underset{\theta}{\text{minimize}}\; L(p_\theta;\lambda)$$

  - ⋆ concave, and nondifferentiable function

- PARAMETRIZED DUAL MAXIMIZATION

$$\underset{\lambda \geq 0}{\text{maximize}} \quad D_{\mathsf{p}}(\lambda)$$

Existence of an optimal parametrized dual variable $\lambda^\star_{\mathsf{p}}$

Search for an optimal parametrized dual variable

Lagrangian minimizer: $\theta^\star(\lambda) \in \operatorname{argmin}_\theta L(p_\theta; \lambda)$

Subgradient: $u(\lambda) = \nabla_\lambda L(p_\theta; \lambda) \mid_{\theta = \theta^\star(\lambda)}$

- PROJECTED SUBGRADIENT ASCENT

$$\lambda^+ \leftarrow [\, \lambda + \eta\, u(\lambda) \,]_+$$

- ⋆ explicit subgradient $u(\lambda) = -E(p_{\theta^\star(\lambda)}; q^i) - b_i$

Subgradient ascent converges to an optimal parametrized dual variable $\lambda_{\mathrm{p}}^\star$

Dual training algorithm

- ITERATION #1: COMPUTE A LAGRANGIAN MINIMIZER

$$\theta^\star(\lambda) \quad \in \quad \operatorname*{argmin}_{\theta} \; L(p_\theta; \lambda)$$

- ITERATION #2: PERFORM A SUBGRADIENT ASCENT STEP

$$\lambda^+ \quad \leftarrow \quad \left[\, \lambda \, - \, \eta \, \left(\, E(p_{\theta^\star(\lambda)}; q^i) \, + \, b_i \, \right) \, \right]_+$$

QUESTION: Optimality of $\lambda^\star_{\mathsf{p}}$-recovered model $p_{\theta^\star(\lambda^\star_{\mathsf{p}})}$?

Optimality

KD†R, NeurIPS '24

Optimality gap: $\mathsf{TV}\left(q_{\mathsf{mix}}^{(\lambda^\star)}, p_{\theta^\star(\lambda_{\mathsf{p}}^\star)}\right)$

Implication (informal)

★ **Optimality gap** is dominated by

$$\frac{\mathsf{poly}(d)}{\sqrt{T}} + \sqrt{d}\,\epsilon_{\mathsf{score}} + \sqrt{\nu}$$

Sublinear in time T & polynomial in dimension d

Linear in score-matching error $\epsilon_{\mathsf{score}}$ & root-scaling of param. gap ν

$$\nu := \max_{\hat{s}} \min_{\theta} \mathsf{dist}(\hat{s}, \hat{s}_\theta)$$

$$\left|E(p_\theta; q_{\mathsf{mix}}^{(\lambda)})\right| \le \epsilon_{\mathsf{score}}^2$$

Optimality analysis

Step #1: ELBO linearity

$$\begin{aligned}\theta^\star(\lambda_{\mathsf{p}}^\star) &\in \operatorname*{argmin}_\theta \; L(p_\theta; \lambda_{\mathsf{p}}^\star) \\ &= \operatorname*{argmin}_\theta \; -E(p_\theta; q) - \sum_{i=1}^m \lambda_{i,\mathsf{p}}^\star E(p_\theta; q^i) \\ &= \operatorname*{argmin}_\theta \; -E\left(p_\theta; q_{\mathsf{mix}}^{(\lambda_{\mathsf{p}}^\star)}\right)\end{aligned}$$

Training a diffusion model with a data mixture $q_{\mathsf{mix}}^{(\lambda_{\mathsf{p}}^\star)}$

⋆ non-asymptotic convergence — LWCC, ICLR '24

$$\mathsf{TV}\left(q_{\mathsf{mix}}^{(\lambda_{\mathsf{p}}^\star)}, p_{\theta^\star(\lambda_{\mathsf{p}}^\star)}\right) \lesssim \frac{\mathsf{poly}(d)}{\sqrt{T}} + \sqrt{d}\,\epsilon_{\mathsf{score}}$$

QUESTION: Gap between $q_{\mathsf{mix}}^{(\lambda_{\mathsf{p}}^\star)}$ and $q_{\mathsf{mix}}^{(\lambda^\star)}$?

Duality gap

- GEOMETRIC INTERPRETATION

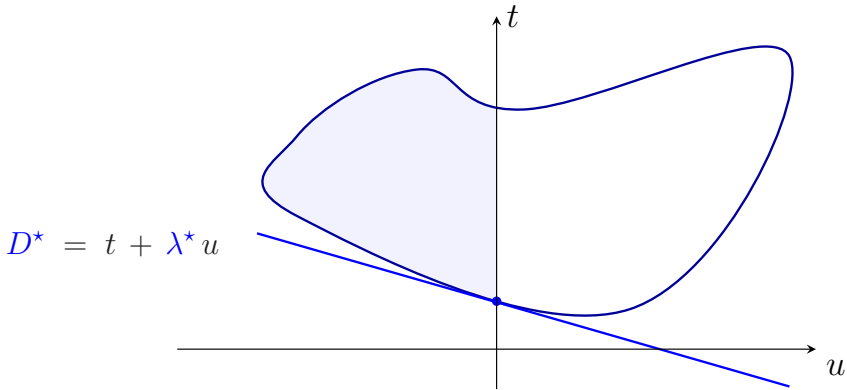


image $\mathcal{E} = \left\{ \left(-E^i(p) - b_i, -E(p) \right) \mid p \right\}$

Duality gap

GEOMETRIC INTERPRETATION

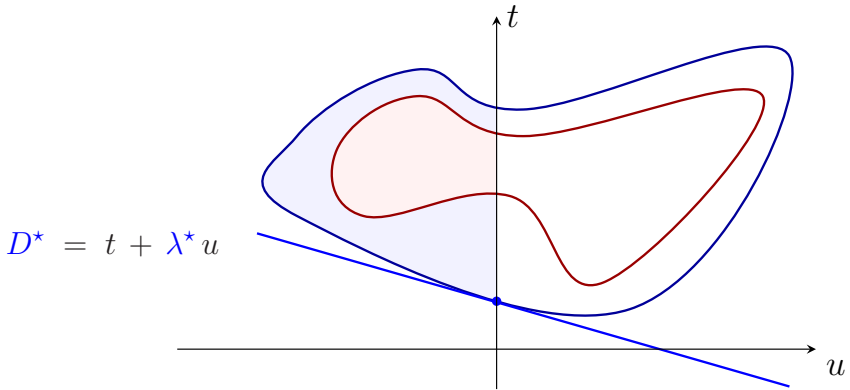


image $\mathcal{E} = \left\{ \left(-E^i(p) - b_i, -E(p) \right) \mid p \right\}$ image $\mathcal{E}_\mathsf{p} = \left\{ \left(-E^i(p_\theta) - b_i, -E(p_\theta) \right) \mid \theta \right\}$

Duality gap

GEOMETRIC INTERPRETATION

$$D_{\mathsf{p}}^{\star} = t + \lambda_{\mathsf{p}}^{\star} u$$

$$D^{\star} = t + \lambda^{\star} u$$

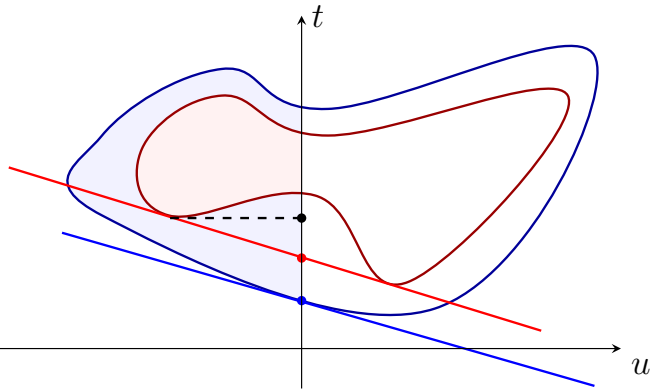


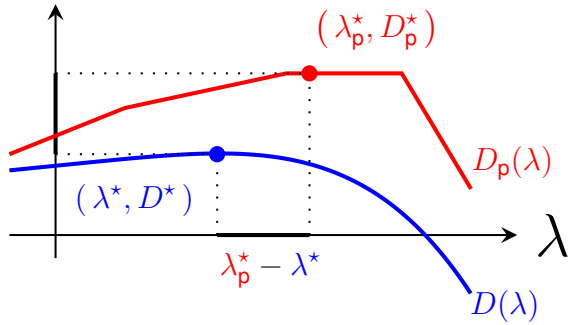
image $\mathcal{E} = \left\{ \left(-E^i(p) - b_i, -E(p) \right) \mid p \right\}$ image $\mathcal{E}_{\mathsf{p}} = \left\{ \left(-E^i(p_\theta) - b_i, -E(p_\theta) \right) \mid \theta \right\}$

Optimal hyperplane touches $\mathcal{E}_{\mathsf{p}}$ w/ t-intercept $D_{\mathsf{p}}^{\star}$: $D_{\mathsf{p}}^{\star} - D^{\star} \lesssim \nu$

Optimality analysis

Step #2: Strong concavity of dual function

$D_{\mathsf{p}}^\star - D^\star \lesssim \nu$



Optimal dual variables: $\lambda^\star$, $\lambda_{\mathsf{p}}^\star$ are close: $\|\lambda_{\mathsf{p}}^\star - \lambda^\star\| \lesssim \sqrt{\nu}$

$$\implies \mathsf{TV}\left(q_{\mathsf{mix}}^{(\lambda^\star)}, q_{\mathsf{mix}}^{(\lambda_{\mathsf{p}}^\star)}\right) \lesssim \sqrt{\nu}$$

PART II POST-TRAINING OF DIFFUSION MODELS

alignment with constraints

Alignment with constraints

Pretrained model: $q(x_{0:T})$

Reward models: $r_i(x_0)$ for $i = 1, \ldots, m$

■ CONSTRAINED DISTRIBUTION OPTIMIZATION

$$\begin{aligned} &\underset{p}{\text{minimize}} && D_{\mathsf{KL}}\left(p(x_{0:T}) \,\|\, q(x_{0:T})\right) \\ &\text{subject to} && \mathbb{E}_{x_0 \sim p(\cdot)}\left[\, r_i(x_0) \,\right] \;\geq\; b_i \;\; \text{for } i = 1, \ldots, m \end{aligned}$$

⋆ convex reverse KL objective and linear constraints

Convex constrained distribution optimization $\rightarrow$ **Strong duality**

Lagrangian approach

LAGRANGIAN

$$L(p;\lambda) \;=\; D_{\mathsf{KL}}\left(p(x_{0:T}) \,\|\, q(x_{0:T})\right) + \sum_{i=1}^{m} \lambda_i \left(b_i - \mathbb{E}_{x_0 \sim p(\cdot)}\left[\, r_i(x_0) \,\right]\right)$$

- ⋆ existence of an optimal dual variable $\lambda^\star$

OPTIMAL MODEL RECOVERY

$$\underset{p}{\text{minimize}} \;\; L(p;\lambda^\star) \qquad \Longleftrightarrow \qquad \underset{p}{\text{minimize}} \;\; D_{\mathsf{KL}}\left(p(x_{0:T}) \,\|\, q_{\mathsf{rw}}^{(\lambda^\star)}(x_{0:T})\right)$$

- ⋆ a reward weighted distribution $q_{\mathsf{rw}}^{(\lambda^\star)}(\cdot) \propto q(\cdot)\mathrm{e}^{(\lambda^\star)^\top r(\cdot)}$

Optimal model is a reward weighted distribution $p^\star(x_{0:T}) = q_{\mathsf{rw}}^{(\lambda^\star)}(x_{0:T})$

PART II POST-TRAINING OF DIFFUSION MODELS

method & theory

Alignment with constraints

Pretrained model: $q(x_{0:T})$

Reward models: $r_i(x_0)$ for $i = 1, \ldots, m$

■ CONSTRAINED PARAMETER OPTIMIZATION

$$\begin{aligned} &\underset{\theta}{\text{minimize}} && D_{\mathsf{KL}}\left(p_\theta(x_{0:T}) \,\|\, q(x_{0:T})\right) \\ &\text{subject to} && \mathbb{E}_{x_0 \sim p_\theta(\cdot)}\left[r_i(x_0)\right] \;\geq\; b_i \;\text{ for } i = 1, \ldots, m \end{aligned}$$

⋆ reverse KL serves as an on-policy score-matching loss

$$D_{\mathsf{KL}}\left(p_\theta(x_{0:T}) \,\|\, q(x_{0:T})\right) \;\;\approxeq\;\; \mathbb{E}_{x_t \sim p_\theta(\cdot),\, t}\left[\, \|s_\theta(x_t, t) - \nabla \log q(x_t)\|^2 \,\right]$$

Nonconvex constrained parameter optimization $\rightarrow$ **Lack of strong duality**

Parametrized dual problem

- PARAMETRIZED DUAL FUNCTION

$$D_{\mathsf{p}}(\lambda) \;:=\; \underset{\theta}{\text{minimize}}\; L(p_\theta;\lambda)$$

  - ⋆ concave, and nondifferentiable function

- PARAMETRIZED DUAL MAXIMIZATION

$$\underset{\lambda \geq 0}{\text{maximize}} \quad D_{\mathsf{p}}(\lambda)$$

Existence of an optimal parametrized dual variable $\lambda^\star_{\mathsf{p}}$

Dual training algorithm

- ITERATION #1: COMPUTE A LAGRANGIAN MINIMIZER

$$\theta^\star(\lambda) \quad \in \quad \operatorname*{argmin}_{\theta} \; L(p_\theta; \lambda)$$

- ITERATION #2: PERFORM A SUBGRADIENT ASCENT STEP

$$\lambda^+ \quad \leftarrow \quad \Big[\, \lambda + \eta \, \Big(b - \mathbb{E}_{p_{\theta^\star(\lambda)}(x_0)} \, [\, r(x_0) \,] \Big) \, \Big]_+$$

QUESTION: Optimality of $\lambda^\star_{\mathsf{p}}$-recovered model $p_{\theta^\star(\lambda^\star_{\mathsf{p}})}$?

Optimality

KH**D**†R, NeurIPS '25

ZLHB**D**†R, NeurIPS '25

Objective optimality: $\left| D_{\mathsf{KL}}\left(p_{\theta^\star(\lambda^\star_{\mathsf{p}})}\right) - D_{\mathsf{KL}}(p^\star) \right|$

Constraint feasibility: $\left| r_i\left(p_{\theta^\star(\lambda^\star_{\mathsf{p}})}\right) - r_i(p^\star) \right|$

Implication (informal)

★ **Objective optimality & Constraint feasibility** are dominated by

$$\sqrt{\nu}$$

Root-scaling of param. gap ν

$$\nu := \max_{p} \min_{\theta} \mathsf{dist}(p, p_\theta)$$

Duality gap

GEOMETRIC INTERPRETATION

$D_{\mathsf{p}}^{\star} = t + \lambda_{\mathsf{p}}^{\star} u$

$D^{\star} = t + \lambda^{\star} u$

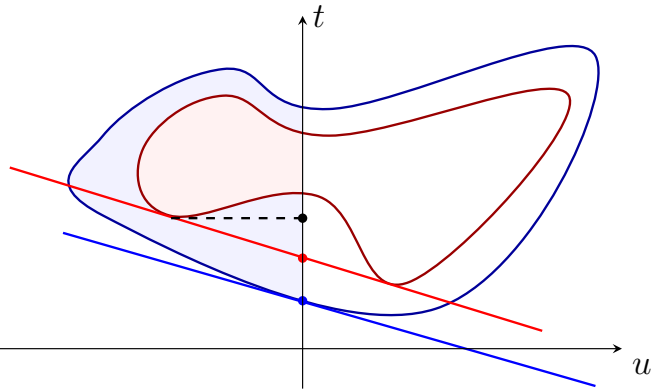


image $\mathcal{E} = \{ (b_i - r_i(p), D_{\mathsf{KL}}(p)) \mid p \}$ image $\mathcal{E}_{\mathsf{p}} = \{ (b_i - r_i(p_\theta), D_{\mathsf{KL}}(p_\theta)) \mid \theta \}$

Optimal hyperplane touches $\mathcal{E}_{\mathsf{p}}$ w/ t-intercept $D_{\mathsf{p}}^{\star}$: $D_{\mathsf{p}}^{\star} - D^{\star} \lesssim \nu$

Dual function

- LOWER ENVELOPE FUNCTION

$$
\begin{aligned}
D(\lambda) &:= \underset{p}{\text{minimize}} \quad L(p;\lambda) \\
&= \lambda^\top b - \log \mathbb{E}_{q(x_{0:T})}\left[\mathrm{e}^{\lambda^\top r(x_0)} \right]
\end{aligned}
$$

cumulant-generating function

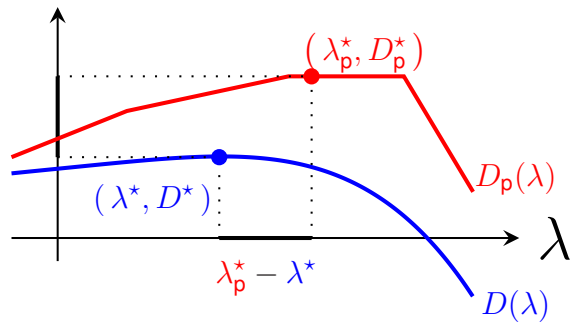
- ⋆ strictly concave, and locally strongly concave function

$$
\nabla^2 D(\lambda) \approx -\,\mathsf{Var}_{x_0 \sim p^\star(\cdot;\lambda)}\left[r(x_0) \right]
$$

Gap between optimal dual variables

- GAP BETWEEN (UN)PARAMETRIZED DUAL FUNCTIONS

$D_{\mathsf{p}}^\star - D^\star \lesssim \nu$

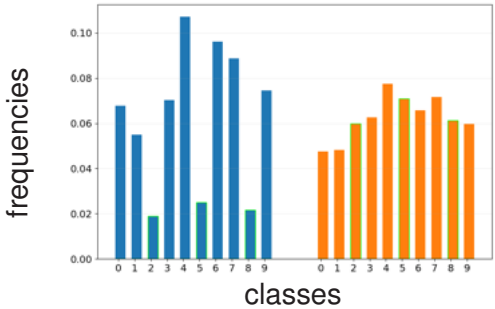


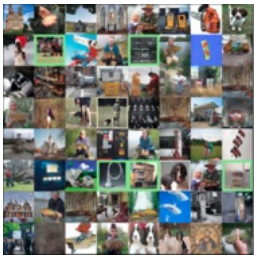
Optimal dual variables: $\lambda^\star$, $\lambda_{\mathsf{p}}^\star$ are close: $\|\lambda_{\mathsf{p}}^\star - \lambda^\star\| \lesssim \sqrt{\nu}$

EMPIRICAL STUDY

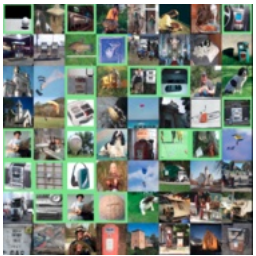
constrained diffusion models

Fairness in image generation





unconstrained

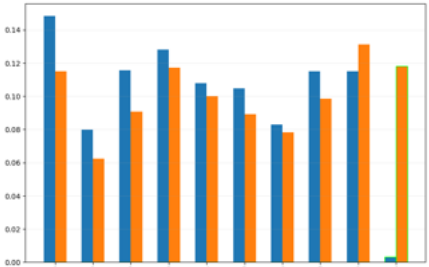


constrained

- unconstrained model (▬)
- constrained model (▬)

Constraints improve representativeness

Adaptation of pretrained model



unconstrained

constrained

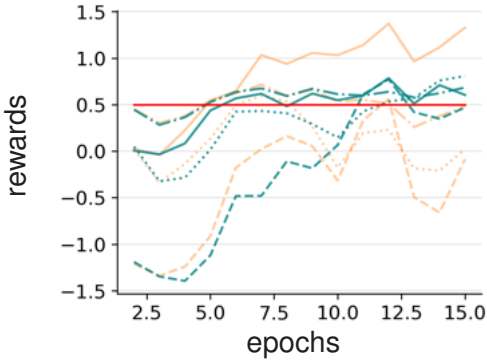
- unconstrained model ( )
- constrained model ( )

Constraints retain model utility

EMPIRICAL STUDY

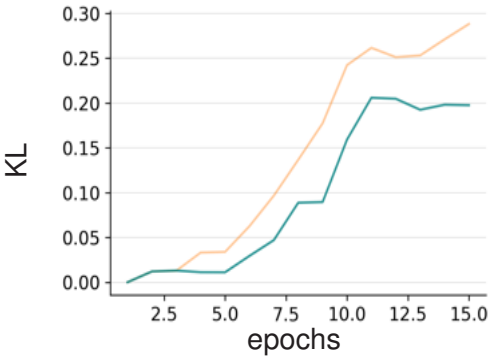
alignment with constraints

Reward alignment via constraints



- unconstrained model (—)
- constrained model (—)

Reward alignment via constraints



pretrained unconstrained constrained

- unconstrained model (—)
- constrained model (—)

Constraints reduce model deviation

Summary & outlook

KD†R, NeurIPS '24
KHD†R, NeurIPS '25
ZLHBD†R, NeurIPS '25

- TRAINING DIFFUSION MODELS WITH CONSTRAINTS
  - ⋆ constrained diffusion models
  - ⋆ alignment with constraints
  - ⋆ dual method & optimality
- OPEN CHALLENGES
  - ⋆ constrained solutions
  - ⋆ generalization

Composition of diffusion models

KHD†R, NeurIPS '25

Pretrained models: $q(x_0)$, $q^i(x_0)$ for $i = 1, \ldots, m$

CONSTRAINED DISTRIBUTION OPTIMIZATION

$$\begin{aligned} \underset{p}{\text{minimize}} \quad & D_{\mathsf{KL}}\left(p(x_0) \,\|\, q(x_0)\right) \\ \text{subject to} \quad & D_{\mathsf{KL}}\left(p(x_0) \,\|\, q^i(x_0)\right) \;\leq\; \epsilon_i \;\; \text{for } i = 1, \ldots, m \end{aligned}$$

- $\star$ concept-aligned pretrained model q^i

Convex constrained distribution optimization $\rightarrow$ **Strong duality**

Unlearning in diffusion models

KRD†, Ongoing

Pretrained model: $q(x_{0:T})$

Unlearning concepts: $q^i(x_{0:T})$ for $i = 1, \dots, m$

CONSTRAINED DISTRIBUTION OPTIMIZATION

$$\begin{aligned} \underset{p}{\text{minimize}} \quad & D_{\mathsf{KL}}\left(q(x_{0:T}) \,\|\, p(x_{0:T})\right) \\ \text{subject to} \quad & D_{\mathsf{KL}}\left(q^i(x_{0:T}) \,\|\, p(x_{0:T})\right) \;\geq\; b_i \quad \text{for } i = 1, \dots, m \end{aligned}$$

- $\star$ concept-aligned pretrained model q^i

Nonconvex constrained distribution optimization $\rightarrow$ **Strong duality**

Thank you for your attention.