Constrained Policy Optimization for Large Language Model Alignment

Dongsheng Ding

THE UNIVERSITY OF
TENNESSEE
KNOXVILLE

2025 SIAM-NNP, State College, PA; Nov. 1, 2025

Motivating application

- AI THERAPY CHATBOTS

Stanford HAI

$$\underset{\text{LLM policy}}{\text{maximize}} \quad \text{helpfulness}$$

$$\text{subject to} \quad \text{harmlessness} \geq \text{margin}$$

CHALLENGE: Constraint satisfaction

Context

- POWERFUL CAPABILITIES

  question-answering, summarization, translation, etc.

- LESSONS LEARNED

  - ⋆ importance of policy optimization in alignment

    reward/preference-based

  - ⋆ multidimensionality of human preference; multi-objective alignment

  - ⋆ convex in distribution space; nonconvex in parameter space

- WHAT NOW ?

  - ⋆ safety-critical domains: healthcare, robotics, finance

  - ⋆ constrained policy optimization: tremendous advances

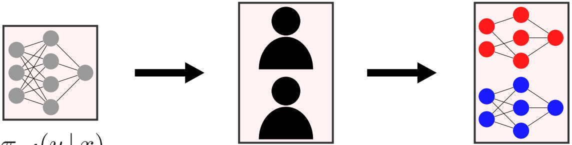
OBJECTIVE

Find a Large Language Model (LLM) that maximizes a performance metric subject to a constraint on another performance metric

Alignment framework

■ REINFORCEMENT LEARNING FROM HUMAN FEEDBACK

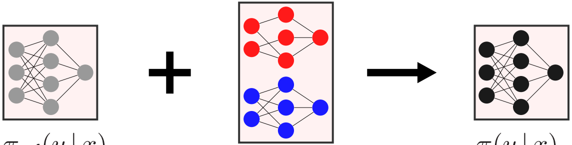
- ⋆ reward modeling



$\pi_{\text{ref}}(y \mid x)$

$r(x, y)$, $g(x, y)$ – reward/utility models

- ⋆ policy optimization



$\pi_{\text{ref}}(y \mid x)$ $\qquad$ $\pi(y \mid x)$

π_{ref}, π: $\mathcal{X}$ (prompts) $\to$ $\mathcal{Y}$ (responses) – LLM policies

Constrained alignment problem

$$\begin{array}{ll} \underset{\pi}{\text{maximize}} & \mathbb{E}_x\left[\,\mathbb{E}_{y\,\sim\,\pi}\left[\,r(x,y)\,\right] \,-\, \beta\, D_{\mathsf{KL}}(\pi(\cdot\,|\,x)\,\|\,\pi_{\mathsf{ref}}(\cdot\,|\,x))\,\right] \\ \text{subject to} & \mathbb{E}_x\left[\,\mathbb{E}_{y\,\sim\,\pi}\left[\,g(x,y)\,\right]\right] \;\geq\; 0 \end{array}$$

$\searrow$

KL-regularized objective

$\downarrow$

policy constraint

$\star$ limit the policy space to an inequality constraint

Convex constrained policy optimization $\longrightarrow$ **Strong duality**

Lagrangian approach

LAGRANGIAN

$$L(\pi, \lambda) \;=\; \mathbb{E}_x \left[\mathbb{E}_{y \sim \pi} \left[r(x, y) + \lambda \, g(x, y) \right] - \beta D_{\mathsf{KL}}(\pi(\cdot \,|\, x) \,\|\, \pi_{\mathsf{ref}}(\cdot \,|\, x)) \right]$$

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

LAGRANGIAN RELAXATION

$$\underset{\pi}{\text{maximize}} \quad L(\pi, \lambda)$$

convex conjugate

- ⋆ exponentially tilted distribution $\pi^\star(\cdot \,|\, x; \lambda) \;\propto\; \pi_{\mathsf{ref}}(\cdot \,|\, x) \, \mathrm{e}^{(r(x,\cdot) \,+\, \lambda g(x,\cdot))/\beta}$
- ⋆ existence of an optimal dual variable $\lambda^\star$

Dual problem

DUAL FUNCTION

$$\begin{aligned} D(\lambda) \;&:=\; \underset{\pi}{\text{maximize}}\;\; L(\pi,\lambda) \\ &=\; \beta\,\mathbb{E}_x\left[\log\left(\sum_y \pi_{\text{ref}}(y\,|\,x)\mathrm{e}^{(r(x,y)+\lambda g(x,y))/\beta}\right)\right] \end{aligned}$$

cumulant-generating function

⋆ convex, smooth, and local strongly convex function

DUAL PROBLEM

$$\underset{\lambda\,\geq\,0}{\text{minimize}}\;\; D(\lambda)$$

Gradient descent finds an optimal dual variable $\lambda^\star$

GEOMETRIC INTERPRETATION OF STRONG DUALITY

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

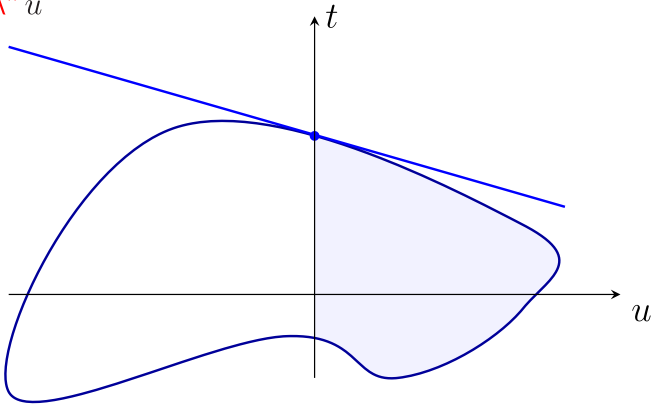


image $\mathcal{E} = \{ (g(\pi), r_{\mathsf{KL}}(\pi)) \mid \pi \}$

Optimal hyperplane touches $\mathcal{E}$ at an optimal policy: $D^\star = r_{\mathsf{KL}}(\pi^\star) := P^\star$

Dualization-based alignment

H†L†DBLHD†, NeurIPS ’24

- STAGE #1: FINDING AN OPTIMAL DUAL VARIABLE

$$\lambda^\star \;=\; \underset{\lambda \geq 0}{\operatorname{argmin}} \; D(\lambda)$$

smooth convex optimization

- STAGE #2: SEARCHING FOR AN LLM POLICY

$$\pi^\star \;=\; \underset{\pi}{\operatorname{argmax}} \; L(\pi, \lambda^\star)$$

unconstrained alignment

Computational efficiency

Constrained parameter optimization

$$\begin{array}{ll} \underset{\theta}{\text{maximize}} & \mathbb{E}_x\left[\,\mathbb{E}_{y\sim\pi_\theta}\left[\,r(x,y)\,\right] - \beta\, D_{\mathsf{KL}}(\pi_\theta(\cdot\,|\,x)\,\|\,\pi_{\mathsf{ref}}(\cdot\,|\,x))\,\right] \\ \text{subject to} & \mathbb{E}_x\left[\,\mathbb{E}_{y\sim\pi_\theta}\left[\,g(x,y)\,\right]\right] \;\geq\; 0 \end{array}$$

$\searrow$

KL-regularized objective

$\downarrow$

policy constraint

- search model parameter θ

CHALLENGE

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

Overview of our results

ZLHB**D**†R, NeurIPS ’25

■ ITERATIVE DUALIZATION-BASED ALIGNMENT

- ⋆ duality gap
- ⋆ optimality gap — objective and constraint

Dual methods find an optimal constrained LLM policy,
up to a parametrization gap

Iterative dualization-based alignment

Parametrized dual problem

■ DUAL FUNCTION

$$D_{\mathsf{p}}(\lambda) \quad := \quad \underset{\theta}{\text{maximize}} \;\; L(\pi_\theta, \lambda)$$

⋆ convex, and nondifferentiable function

■ DUAL PROBLEM

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

Subgradient descent finds an optimal parametrized dual variable $\lambda_{\mathsf{p}}^\star$

Iterative dualization-based alignment

Lagrangian maximizer: $\theta^\star(\lambda) \in \underset{\theta}{\operatorname{argmax}}\ L(\pi_\theta, \lambda)$

Subgradient: $u(\lambda) = \nabla_\lambda L(\pi_\theta, \lambda)\,|_{\theta\,=\,\theta^\star(\lambda)}$

- SUBGRADIENT DESCENT

$$\lambda(t+1) = [\,\lambda(t) - \eta\, u(\lambda(t))\,]_+$$

⋆ explicit subgradient $u(\lambda(t)) = \mathbb{E}_x[\,\mathbb{E}_{y\,\sim\,\pi_{\theta^\star(\lambda(t))}}[\,g(x,y)\,]\,]$

QUESTION: Optimality of $\lambda^\star_{\mathrm{p}}$-recovered model $\pi_{\theta^\star(\lambda^\star_{\mathrm{p}})} := \pi^\star_{\mathrm{p}}(\lambda^\star_{\mathrm{p}})$?

Duality gap

Duality gap: $|\, P^\star - D_{\mathsf{p}}^\star \,|$

Theorem (informal)

★ **Duality gap** is dominated by

ν

parametrization gap $\nu := \max_{\pi} \min_{\theta} \mathsf{dist}_1(\pi, \pi_\theta)$

⋆ ν-parametrization gap yields ν-duality gap

linear independence

GEOMETRIC INTERPRETATION OF DUALITY GAP

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

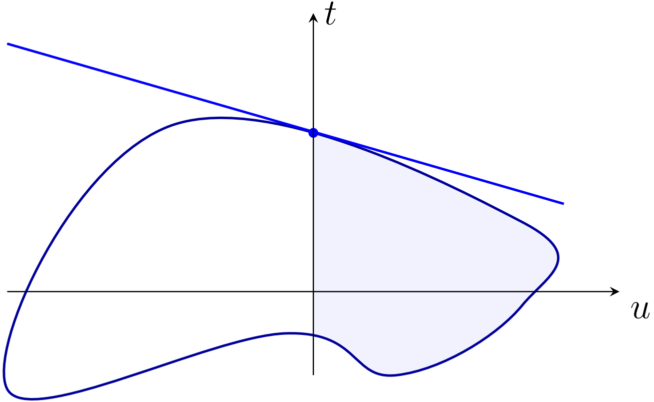


image $\mathcal{E} = \{ (g(\pi), r_{\mathsf{KL}}(\pi)) \mid \pi \}$

GEOMETRIC INTERPRETATION OF DUALITY GAP

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

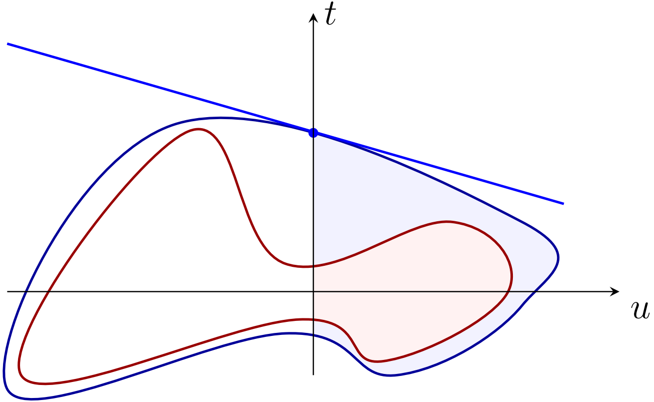


image $\mathcal{E} = \{\, (g(\pi), r_{\mathsf{KL}}(\pi)) \mid \pi \,\}$ image $\mathcal{E}_{\mathsf{p}} = \{\, (g(\pi_\theta), r_{\mathsf{KL}}(\pi_\theta)) \mid \theta \,\}$

GEOMETRIC INTERPRETATION OF DUALITY GAP

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

$D^\star_{\mathsf{p}} \;=\; t \,+\, \lambda^\star_{\mathsf{p}}\, u$

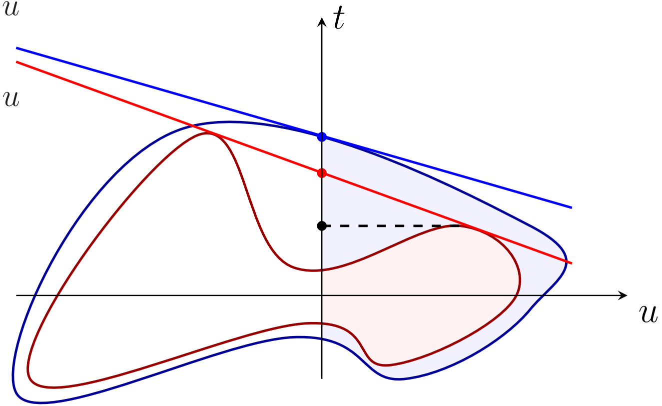


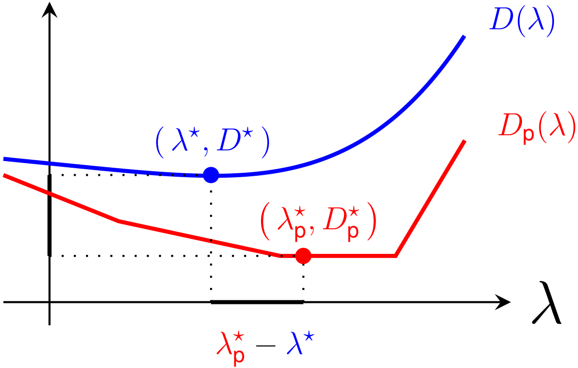
image $\mathcal{E} \;=\; \{\, (g(\pi), r_{\mathsf{KL}}(\pi)) \;|\; \pi \,\}$ image $\mathcal{E}_{\mathsf{p}} \;=\; \{\, (g(\pi_\theta), r_{\mathsf{KL}}(\pi_\theta)) \;|\; \theta \,\}$

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

$$D^\star - D^\star_{\mathsf{p}} \;=\; \mathcal{O}(\nu)$$

Gap between optimal dual variables

- GAP BETWEEN (UN)PARAMETRIZED DUAL FUNCTIONS



$$D^\star - D_{\mathsf{p}}^\star \quad = \quad \mathcal{O}(\nu)$$

Optimal dual variables: $\lambda^\star$, $\lambda_{\mathsf{p}}^\star$ are close:

$$\|\lambda^\star - \lambda_{\mathsf{p}}^\star\| \quad = \quad \mathcal{O}(\sqrt{\nu})$$

Optimality gap

Objective optimality: $\left| r_{\mathsf{KL}} \left(\pi_{\mathsf{p}}^{\star}(\lambda_{\mathsf{p}}^{\star}) \right) - r_{\mathsf{KL}}(\pi^{\star}) \right|$

Constraint feasibility: $\left| g \left(\pi_{\mathsf{p}}^{\star}(\lambda_{\mathsf{p}}^{\star}) \right) - g(\pi^{\star}) \right|$

Implication (informal)

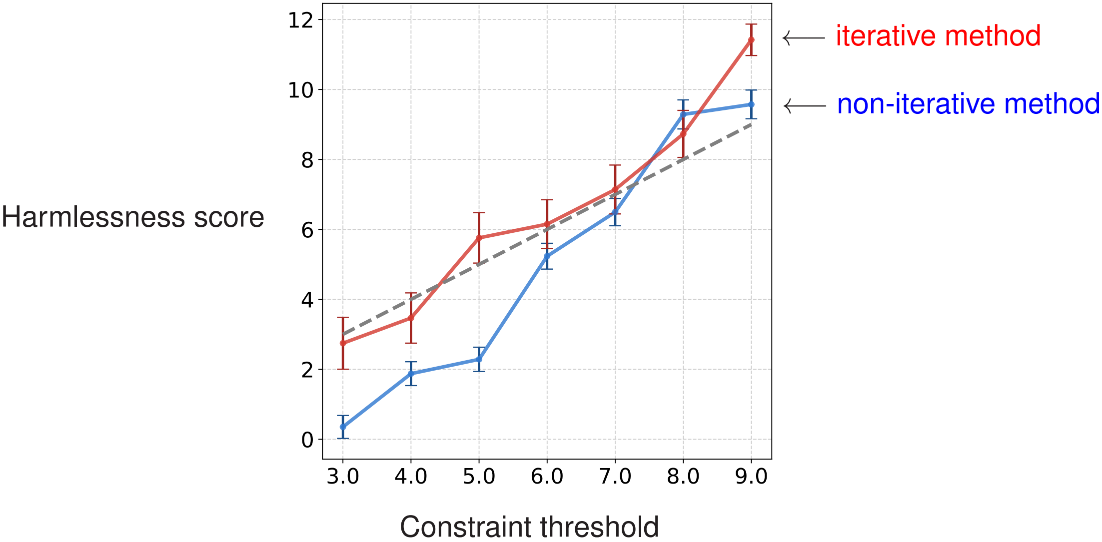
★ **Objective optimality & Constraint feasibility** are dominated by

$$\sqrt{\nu}$$

parametrization gap $\nu := \max_{\pi} \min_{\theta} \mathsf{dist}_1(\pi, \pi_\theta)$

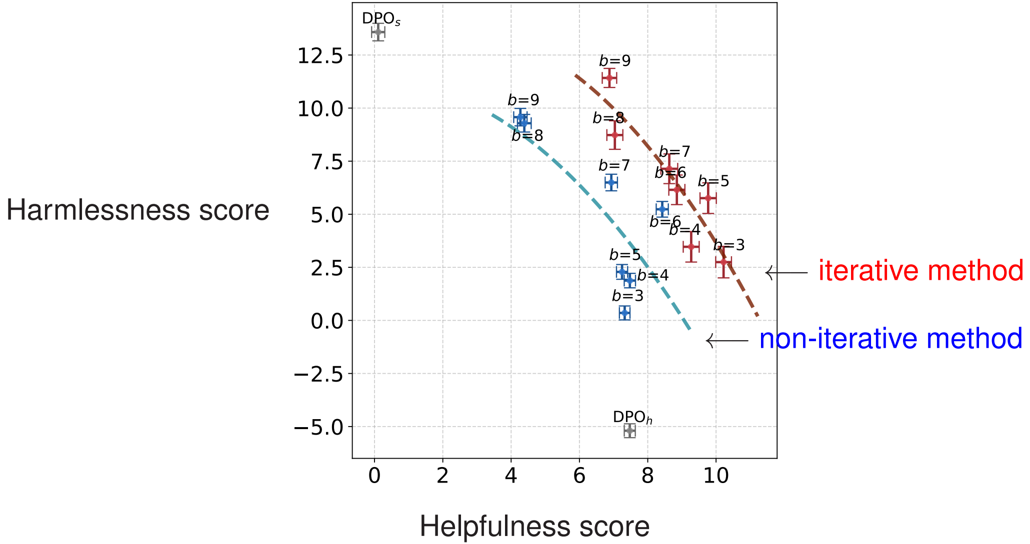
Root-scaling of parametrization gap

Constraint satisfaction



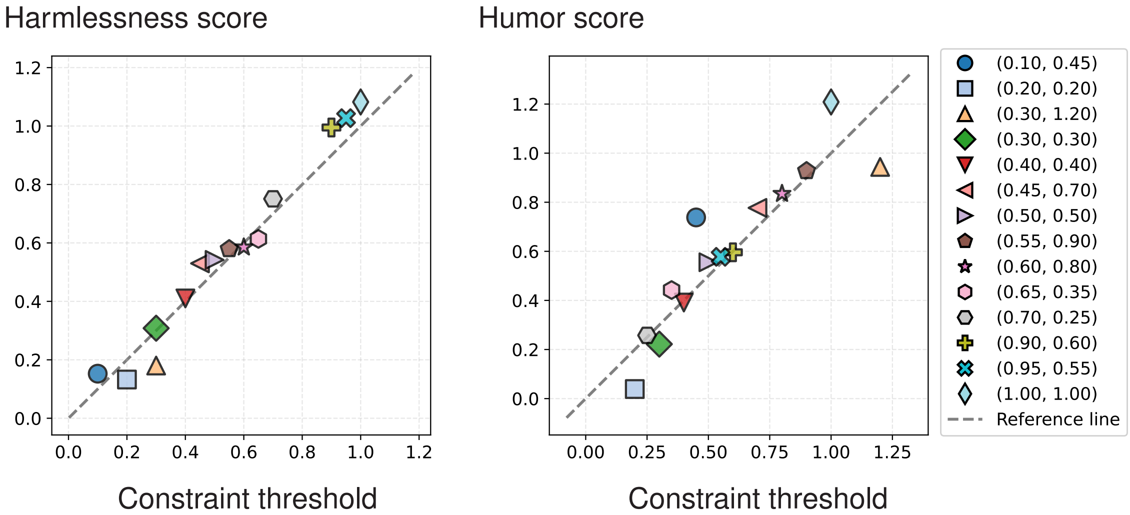
Better constraint satisfaction

Helpfulness and Harmlessness tradeoff



Higher Pareto frontier

Harmlessness and Humor constraints



Multi-constraint satisfaction

Summary

ZLHB**D**†R, NeurIPS ’25

H†L†DBLH**D**†, NeurIPS ’24

- DUALIZATION-BASED ALIGNMENT METHODS
  - ⋆ iterative dualization-based alignment
  - ⋆ duality gap and optimality gap
- ON-GOING EFFORTS
  - ⋆ sample complexity and convergence
  - ⋆ other complex constraints

Thank you for your attention.