Alignment of Large Language Models with Constrained Learning

Botong Zhang[1] Shuo Li[1] Ignacio Hounie[1]
Osbert Bastani[1] Dongsheng Ding[2] Alejandro Ribeiro[1]

[1]University of Pennsylvania [2]University of Tennessee, Knoxville

Motivation

Problem: Align LLMs to optimize a primary reward while satisfying constraints on secondary objectives.

Challenges:

- **Single-reward RLHF**: cannot capture multidimensional and conflicting human preferences
- **Multi-objective alignment**: manual tuning and lack constraint guarantees
- **Primal-dual methods**: high computational effort
- **One-shot methods**: optimize distributions, not actual LLM policies
- **Lack of theory** on policy optimality in the LLM parameter space

Approach: Propose an iterative dual-based method that alternates between:

- Lagrangian maximization to update the LLM policy
- Dual descent to update the Lagrange multipliers

Key Insight: Dual-based methods can find **optimal constrained LLM policies**, up to a parameterization gap.

Problem Formulation

Constrained alignment problem:

Language policy: $\pi_\theta(\cdot \mid \mathbf{x})$: $\mathcal{X} \to \Delta(\mathcal{Y})$

- Maps prompt $\mathbf{x}$ to a distribution over responses
- θ: LLM parameters (e.g., transformer weights)

Goal: Maximize reward $r(\mathbf{x}, \mathbf{y})$ with constraints on each utility $g_i(\mathbf{x}, \mathbf{y})$

Optimization Problem (P-CA):

$$\begin{aligned} \max_{\theta \in \Theta} \quad & \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{\mathbf{y}\sim\pi_\theta}[\, r(\mathbf{x},\mathbf{y})\,] \;-\; \beta\, D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}};\mathbf{x})] \\ \text{s.t.} \quad & \mathbb{E}_{\mathbf{x}}[\mathbb{E}_{\mathbf{y}\sim\pi_\theta}[g_i(\mathbf{x},\mathbf{y})] \;-\; \mathbb{E}_{\mathbf{y}\sim\pi_{\mathrm{ref}}}[g_i(\mathbf{x},\mathbf{y})]\,] \;\geq\; b_i, \qquad i = 1,\dots,m \end{aligned}$$

- $\mathbb{E}_{\mathbf{x}}$: Expectation over prompt distribution $\mathcal{D}$
- D_{KL}: KL divergence at prompt $\mathbf{x}$
- β: KL regularization coefficient
- b_i: Targeted improvement in utility g_i

One-Shot Optimal Dualization:

$$\begin{aligned} L(\pi_\theta, \lambda) \;:=\; \mathbb{E}_{\mathbf{x}}\,[\,\mathbb{E}_{\mathbf{y}\sim\pi_\theta}[\,r(\mathbf{x},\mathbf{y}) + \lambda_i \cdot (g_i(\mathbf{x},\mathbf{y}) - \mathbb{E}_{\mathbf{y}\sim\pi_{\mathrm{ref}}}[\,g_i(\mathbf{x},\mathbf{y})\,] - b_i)\,] \\ -\beta\, D_{\mathrm{KL}}(\pi_\theta \,\|\, \pi_{\mathrm{ref}};\mathbf{x})\,] \end{aligned}$$

Dual Problem: We define the dual function and solve

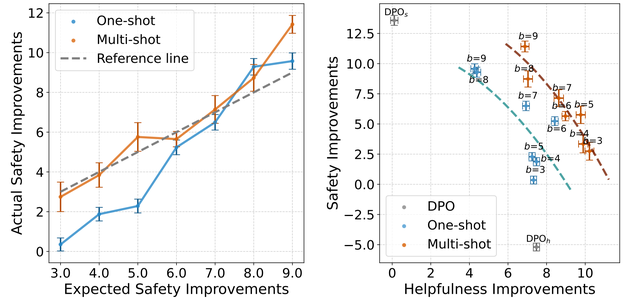
$$\min_{\lambda \geq 0} D_{\mathrm{p}}(\lambda), \quad \text{where } D_{\mathrm{p}}(\lambda) = \max_{\theta \in \Theta} L(\pi_\theta, \lambda)$$

Safety Alignment Algorithm:

Our **multi-shot method** closely approximates an **optimal constrained LLM policy**, outperforming the **one-shot baseline**.

(i) Better aligns π_{ref} to satisfy **safety constraints**.

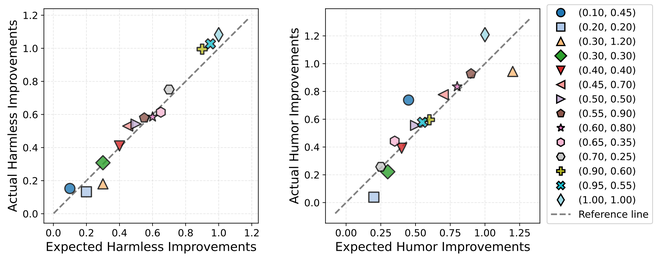
(ii) Improves the **helpfulness–safety trade-off**.



Multi-Constraint Alignment:

We extend our method to settings with more than one constraint.

(i) Achieves **multiple-constraint satisfaction**



Constrained Alignment via Iterative Dualization

Input: Reference model π_{ref}, initial dual λ_{init}, reward r, utilities $\{g_i\}_{i=1}^m$, step size η, total iterations T, regularization parameter β, targeted improvements in utilities $\{b_i\}_{i=1}^m$

Initialization: $\lambda(0) = \lambda_{\mathrm{init}}$, $\pi_{\theta^\star(0)} = \pi_{\mathrm{ref}}$

For $t = 0, 1, \dots, T-1$ do:

- Dual subgradient step:

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

where $u(\lambda(t))$ is a subgradient direction
$u(\lambda(t)) \;=\; \mathbb{E}_{\mathbf{x}}\left[\,\mathbb{E}_{\mathbf{y}\sim\bar{\pi}(t)}[\,g(\mathbf{x},\mathbf{y})\,] \;-\; \mathbb{E}_{\mathbf{y}\sim\pi_{\mathrm{ref}}}[\,g(\mathbf{x},\mathbf{y})\,]\right] \;-\; b,$
$\bar{\pi}(t) := \pi_{\theta^\star(t)}(\lambda(t))$

- LLM policy optimization step:

$$\theta^\star(t+1) \;\in\; \arg\max_{\theta \in \Theta} \; L(\pi_\theta, \lambda(t+1))$$

Output: $\{\theta^\star(t)\}_{t=1}^T$

Optimality Analysis

Primal-dual gap is provably small:

- $D_{\mathrm{p}}^\star := \mathrm{maximize}_{\theta\in\Theta}\, L_{\mathrm{p}}(\pi_\theta, \lambda_p^\star)$
- Let $P_{\mathrm{p}}^\star$ be the primal value of Problem (P-CA)
- Dual value approximates primal optimum up to a **parametrization gap** ν where $\nu := \max_\pi \min_\theta \|\pi - \pi_\theta\|_1$:

$$|D_{\mathrm{p}}^\star - P_{\mathrm{p}}^\star| \;=\; O(\nu)$$

Reward and utility optimality:

- The learned policy is near-optimal in both **reward** and **constraint**
- Optimality gaps scale with ν:

$$\text{Objective optimality, Constraint feasibility} \;=\; O(\sqrt{\nu})$$

Multi-shot vs. One-shot alignment:

- **Multi-shot** (ours): Iteratively improves policy; benefits from warm start and admits tighter optimality bounds (if dual is well-conditioned).
- **One-shot**: Simpler, closed-form dual; optimality holds under similar assumptions but lacks iterative refinement.

{bzhang16, lishuo1}@seas.upenn.edu dongshed@utk.edu

NeurIPS 2025, San Diego