

Requirement-Driven Machine Decision-Making

(Reinforcement Learning
Generative Artificial Intelligence)

Dongsheng Ding



University of Tennessee, Knoxville; March 6, 2025

Goal-oriented machine decision-making

■ HUMAN-LEVEL PERFORMANCE



AlphaGo '16



GT Sophy '22

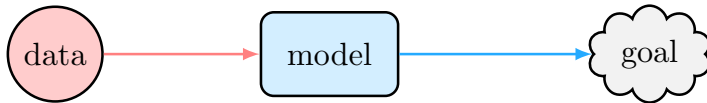


DALL-E 3 '23



ChatGPT '24

■ MACHINE DECISION MODEL TRAINING



model – decision-maker / policy / controller / agent

goal – loss / reward / likelihood

The challenge of imposing requirements

■ SHORTCOMINGS



Waymo robotaxi collision '24



DALL-E 3 biased images '24

■ REQUIREMENTS

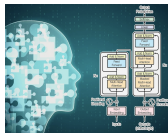
Safety



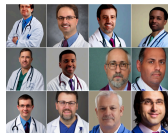
Robustness



Harmlessness



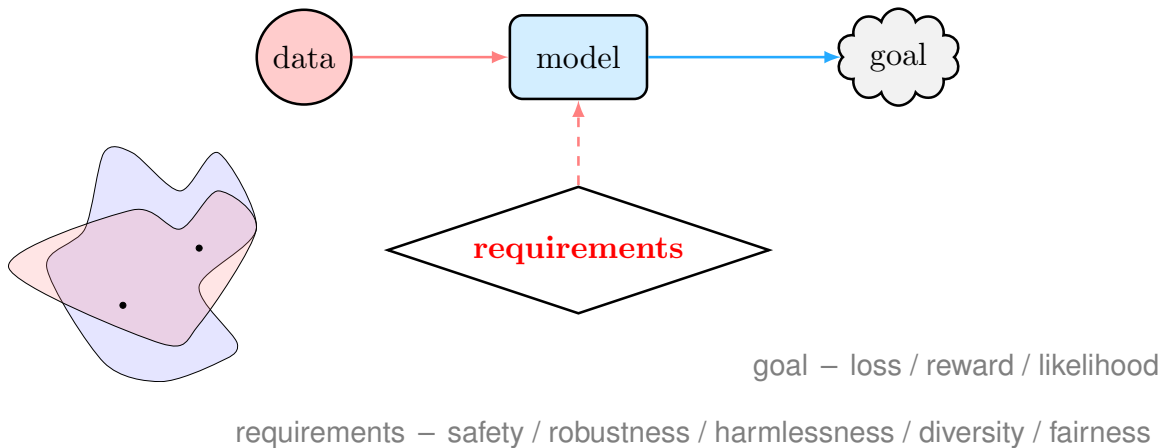
Diversity



Fairness

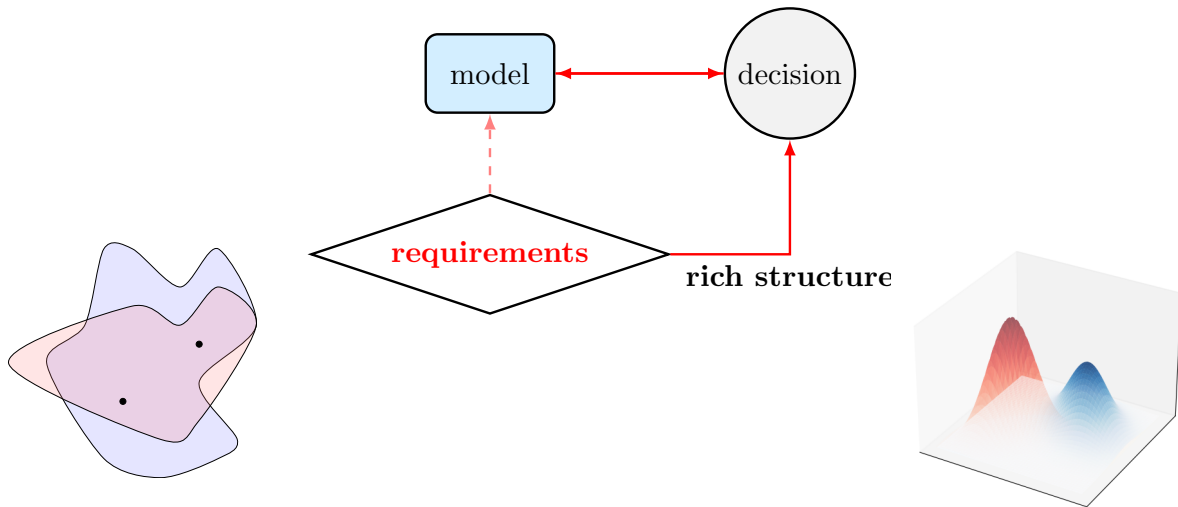


Requirement-driven machine decision-making



CHALLENGE: Nonconvex constraints

My research: Laying down the **principled methodologies**



From reinforcement learning to generative modeling

■ REQUIREMENT-DRIVEN REINFORCEMENT LEARNING

QUESTION

Can direct policy search methods find
an optimal constrained policy?

■ REQUIREMENT-DRIVEN GENERATIVE MODELING

QUESTION

How to efficiently impose constraints
on **large generative models?**

■ FUTURE DIRECTIONS

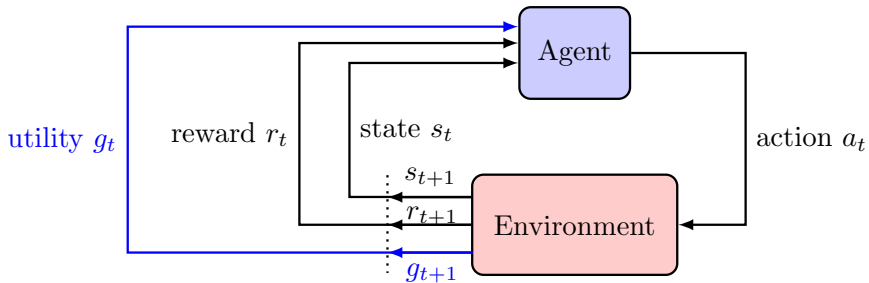
Part I

Requirement-Driven Reinforcement Learning

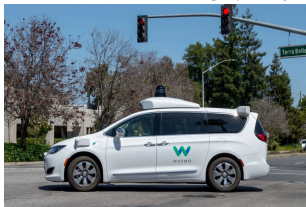
(control of constrained dynamical systems)

Reinforcement learning framework

■ CONSTRAINED MDP



Autonomous driving (Waymo)



Goal reaching a destination

Requirements collision avoidance
energy budget

⋮

Constrained policy optimization

policy: $\pi : S \text{ (states)} \rightarrow \Delta(A) \text{ (action distribution)}$

reward value function: $V_r^\pi(\rho) := \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \mid s_0 \sim \rho \right]$

utility value function: $V_g^\pi(\rho) := \mathbb{E} \left[\sum_{t=0}^{\infty} \gamma^t g(s_t, a_t) \mid s_0 \sim \rho \right]$

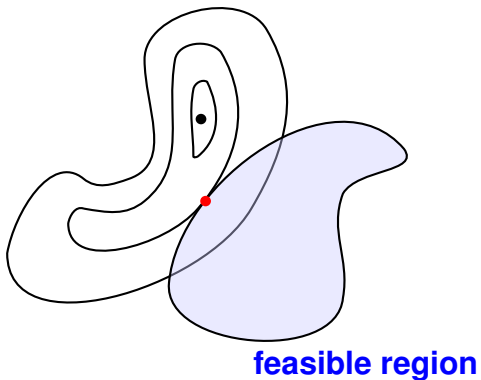
■ NONCONVEX POLICY OPTIMIZATION W/ FUNCTIONAL CONSTRAINT

maximize	$V_r^\pi(\rho)$	→	standard goal
subject to	$V_g^\pi(\rho) \geq 0$	→	policy constraint

★ Limit the policy space to **an inequality constraint**

■ 2-D LEVEL NONCONVEX CURVES

optimal, but in**feasible**



Optimality w/ **feasibility**

⇒ initialization dependent & stochastic

Lagrangian approach

Lagrangian: $L(\pi, \lambda) = V_r^\pi(\rho) + \lambda V_g^\pi(\rho)$

Lagrange multiplier: $\lambda (\geq 0)$

composite reward: $r + \lambda g$

■ NONCONVEX POLICY OPTIMIZATION

$$\underset{\pi}{\text{maximize}} \quad V_{r + \lambda g}^\pi(\rho)$$

Multiple solutions, but infeasible

Primal-dual method

dual problem: $\underset{\lambda}{\text{minimize}} \underset{\pi}{\text{maximize}} V_{r+\lambda g}^{\pi}(\rho)$

primal problem: $\underset{\pi}{\text{maximize}} \underset{\lambda}{\text{minimize}} V_{r+\lambda g}^{\pi}(\rho)$

■ NONCONVEX MIN-MAX POLICY OPTIMIZATION

Strong duality $\implies$ **exchange** of min and max

$$\underset{\lambda \in \Lambda}{\text{minimize}} \underset{\pi}{\text{maximize}} V_{r+\lambda g}^{\pi}(\rho) = \underset{\pi}{\text{maximize}} \underset{\lambda \in \Lambda}{\text{minimize}} V_{r+\lambda g}^{\pi}(\rho)$$

OBJECTIVE: find a minimax point (π^*, λ^*)

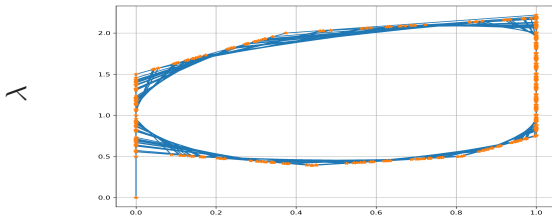
■ PRIMAL-DUAL UPDATE

policy ascent direction: $\nabla_{\pi} V_{r+\lambda g}^{\pi}(\rho)$

dual descent direction: $V_g^{\pi}(\rho)$

$$\pi^+ \leftarrow \pi + \eta_1 \nabla_{\pi} V_{r+\lambda g}^{\pi}(\rho)$$

$$\lambda^+ \leftarrow \lambda - \eta_2 V_g^{\pi}(\rho)$$



$$\pi \left(\pi^{\star} = 0.66 \right)$$

CHALLENGE

Stabilize primal-dual methods
in face of **nonconvex** Lagrangian

My first contribution: **Best-iterate convergent result**

DZBJ, NeurIPS '20; JMLR '23 (under review)

■ NATURAL POLICY GRADIENT PRIMAL-DUAL METHOD

primal: natural policy gradient
dual: subgradient

Theorem (informal)

★ **Average optimality gap and feasibility gap are bounded by $\frac{1}{\sqrt{T}} + \sqrt{\epsilon_{\text{fa}}}$.**

T – number of iterations

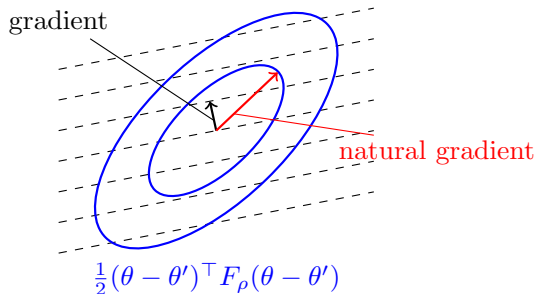
ϵ_{fa} – function approximation error

Natural (policy) gradient

parametrized policy: π_θ w/ parameter θ

natural gradient: $F_\rho(\theta)^\dagger \nabla_\theta V_{r+\lambda g}^{\pi_\theta}(\rho)$

Fisher information: $F_\rho(\theta) = \mathbb{E} \left[\nabla_\theta \log \pi_\theta (\nabla_\theta \log \pi_\theta)^\top \right]$



Steepest descent in Fisher information distance

Natural policy gradient primal-dual method

policy ascent direction: $F_{\rho}(\theta)^{\dagger} \nabla_{\theta} V_{r+\lambda g}^{\pi_{\theta}}(\rho)$

dual descent direction: $V_g^{\pi}(\rho)$

■ SIMPLE PRIMAL-DUAL UPDATE

$$\begin{aligned}\theta^+ &= \theta + \eta_1 F_{\rho}(\theta)^{\dagger} \nabla_{\theta} V_{r+\lambda g}^{\pi_{\theta}}(\rho) \\ \lambda^+ &= \operatorname{argmin}_{\lambda' \in \Lambda} \lambda' V_g^{\pi}(\rho) + \frac{1}{2\eta_2} (\lambda' - \lambda)^2\end{aligned}$$

Costly inverse of Fisher information matrix

Parallel b/w natural gradient and action value function

softmax policy: $\pi_{\theta}(a | s) = \frac{e^{\theta_{s,a}}}{\sum_{a'} e^{\theta_{s,a'}}}$ w/ parameter $\theta \in \mathbb{R}^{|S||A|}$

linear regression: $w^{\star} = \text{LR} \left(w; \nabla_{\theta} \log \pi_{\theta}, Q_{r+\lambda g}^{\pi_{\theta}} \right)$

■ MULTIPLE OPTIMAL SOLUTIONS

$$\begin{aligned} w^{\star} &= F_{\rho}(\theta)^{\dagger} \nabla_{\theta} V_{r+\lambda g}^{\pi_{\theta}}(\rho) \\ &\simeq Q_{r+\lambda g}^{\pi_{\theta}} \end{aligned}$$

$F_{\rho}(\theta)^{\dagger} \nabla_{\theta} V_{r+\lambda g}^{\pi_{\theta}}(\rho)$ is parallel with $Q_{r+\lambda g}^{\pi_{\theta}}$

Natural policy gradient primal-dual method

policy ascent direction: $Q_{r+\lambda g}^{\pi_\theta}$

dual descent direction: $V_g^\pi(\rho)$

■ POLICY PRIMAL-DUAL UPDATE

$$\begin{aligned}\pi^+(\cdot | s) &= \operatorname{argmax}_{\pi'} \langle \pi'(\cdot | s), Q_{r+\lambda g}^\pi(s, \cdot) \rangle - \frac{1}{\eta_1} D_{\text{KL}}(\pi'(\cdot | s) \| \pi(\cdot | s)) \\ \lambda^+ &= \operatorname{argmin}_{\lambda' \in \Lambda} \lambda' V_g^\pi(\rho) + \frac{1}{2\eta_2} (\lambda' - \lambda)^2\end{aligned}$$

★ Explicit policy update $\pi^+(\cdot | s) \propto \pi(\cdot | s) e^{\eta_1 Q_{r+\lambda g}^\pi(s, \cdot)}$

Convex policy update

Non-asymptotic time-average performance

average optimality gap: $\frac{1}{T} \sum_{t=1}^T (V_r^{\star}(\rho) - V_r^{\pi_t}(\rho))$

average feasibility gap: $\frac{1}{T} \sum_{t=1}^T (0 - V_g^{\pi_t}(\rho))$

Theorem (informal)

★ **Average optimality gap and feasibility gap** are bounded by

$$\frac{1}{\sqrt{T}}$$

Sublinear rate

Dimension-free rate w/ no dependence of $(|S|, |A|, \rho)$

Log-linear policy primal-dual update

log-linear policy: $\pi_{\theta}(a | s) = \frac{e^{\theta^{\top} \phi_{s,a}}}{\sum_{a'} e^{\theta^{\top} \phi_{s,a'}}}$ w/ feature $\phi_{s,a} \in \mathbb{R}^d$

policy ascent direction: $\hat{w}^{\star} \approx \text{LR} \left(w; \phi, Q_{r+\lambda g}^{\pi_{\theta}} \right)$

■ POLICY PRIMAL-DUAL UPDATE

$$\theta^+ = \theta + \eta_1 \hat{w}^{\star}$$

$$\lambda^+ = \underset{\lambda' \in \Lambda}{\operatorname{argmin}} \lambda' V_g^{\pi}(\rho) + \frac{1}{2\eta_2} (\lambda' - \lambda)^2$$

Non-asymptotic time-average performance

average optimality gap: $\frac{1}{T} \sum_{t=1}^T (V_r^*(\rho) - V_r^{\pi_t}(\rho))$

average feasibility gap: $\frac{1}{T} \sum_{t=1}^T (0 - V_g^{\pi_t}(\rho))$

Theorem (informal)

★ **Average optimality gap and feasibility gap** are bounded by

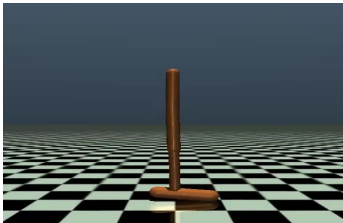
$$\frac{1}{\sqrt{T}} + \sqrt{\epsilon_{\text{fa}}}$$

Sublinear decay up to a function approximation error ϵ_{fa}

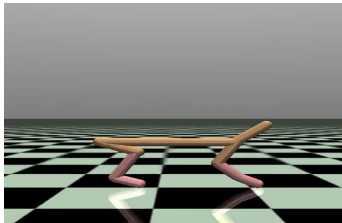
- GENERAL SMOOTH POLICY
- SAMPLE-BASED IMPLEMENTATION

MuJoCo robotics

Hopper

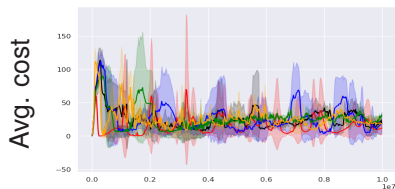
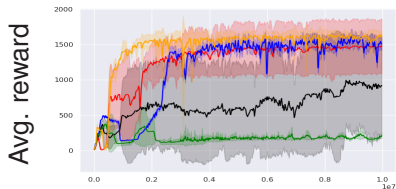


HalfCheetah

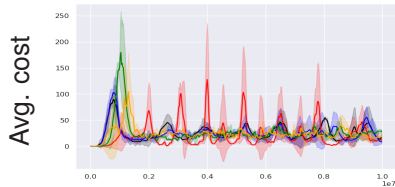
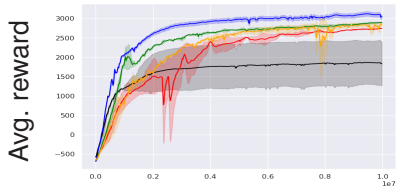


- ★ Energy budget = speed limit 25
- ★ **Cost** – **Reward** tradeoff: walk with energy efficiency

Hopper



HalfCheetah



horizontal axis – # sample count

★ Our method (—, blue)

★ CUP (—, red), FOCOPS (—, orange), TRPOLag (—, black), PPOLag (—, green)

My second contribution: Last-iterate convergent result

DWZR, NeurIPS '23

■ REGULARIZED POLICY GRADIENT PRIMAL-DUAL METHOD

natural policy primal-dual for **regularized** Lagrangian

■ OPTIMISTIC POLICY GRADIENT PRIMAL-DUAL METHOD

optimistic natural policy primal-dual for Lagrangian

Theorem (informal)

- ★ Last-iterate of regularized algorithm has ϵ -optimality gap and ϵ -feasibility gap after $\frac{1}{\epsilon^6}$ iterations.
- ★ Last-iterate of optimistic algorithm has ϵ -optimality gap and ϵ -feasibility gap after $\log \frac{1}{\epsilon^2}$ iterations.

Regularized method

Regularized Lagrangian approach

entropy-like term: $\mathcal{H}(\pi) = \mathbb{E} \left[\sum_{t=0}^{\infty} -\gamma^t \log \pi(a_t | s_t) \right]$

■ REGULARIZED LAGRANGIAN

$$L_{\tau}(\pi, \lambda) = V_{r+\lambda g}^{\pi}(\rho) + \tau \left(\mathcal{H}(\pi) + \frac{1}{2} \lambda^2 \right)$$



“convexify” Lagrangian $V_{r+\lambda g}^{\pi}(\rho)$

OBJECTIVE: find the regularized minimax point $(\pi_{\tau}^*, \lambda_{\tau}^*)$

★ **τ -near minimax point of** $V_{r+\lambda g}^{\pi}(\rho)$

Regularized policy gradient primal-dual method

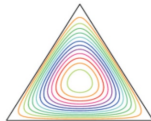
policy ascent direction: $Q_{L_\tau}^\pi(s, a) := Q_{r + \lambda g - \tau \log \pi}(s, a)$

dual descent direction: $V_g^\pi(\rho) + \tau \lambda$

■ SINGLE-TIME-SCALE PRIMAL-DUAL UPDATE

$$\begin{aligned}\pi^+(\cdot | s) &= \operatorname{argmax}_{\pi' \in \Delta_{\epsilon_0}} \langle \pi'(\cdot | s), Q_{L_\tau}^\pi(s, \cdot) \rangle - \frac{1}{\eta} D_{\text{KL}}(\pi'(\cdot | s) \| \pi(\cdot | s)) \\ \lambda^+ &= \operatorname{argmin}_{\lambda' \in \Lambda} \lambda' (V_g^\pi(\rho) + \tau \lambda) + \frac{1}{2\eta} (\lambda' - \lambda)^2\end{aligned}$$

★ Restricted policy set $\Delta_{\epsilon_0} = \{p_a \geq \epsilon_0, \forall a\}$



Non-asymptotic last-iterate performance

distance of (π_t, λ_t) to $(\pi_\tau^*, \lambda_\tau^*)$: $\mathbb{E} [D_{\text{KL}}(\pi_\tau^* \parallel \pi_t)] + \frac{1}{2}(\lambda_t - \lambda_\tau^*)^2$

Theorem (informal)

★ **Distance of (π_t, λ_t) to $(\pi_\tau^*, \lambda_\tau^*)$ is bounded by**

$$e^{-\eta \tau t} + \frac{\eta}{\tau}$$

Exponential decay up to a ratio

★ ϵ -near regularized minimax point requires

$$\eta = \epsilon \tau \quad \text{and} \quad t = \frac{1}{\epsilon \tau^2}$$

Sublinear rate

optimality gap: $V_r^*(\rho) - V_r^{\pi_t}(\rho)$

feasibility gap: $0 - V_g^{\pi_t}(\rho)$

Implication (informal)

★ **Achieving ϵ -optimality gap and ϵ -feasibility gap** requires

$\frac{1}{\epsilon^6}$ iterations

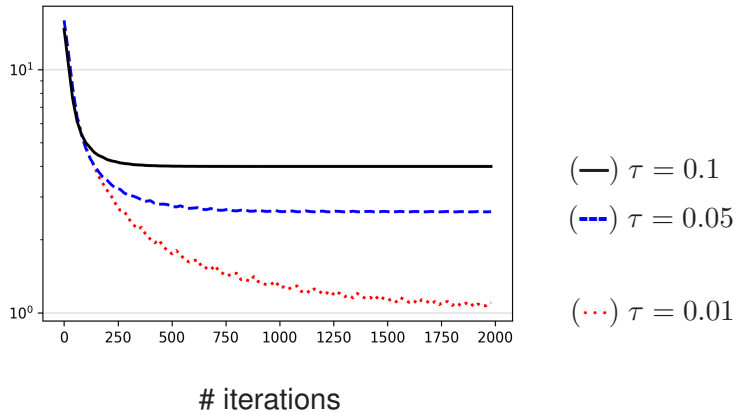
Dimension-free sublinear rate

stepsize $\eta = \epsilon^4$

regularization $\tau = \epsilon^2$

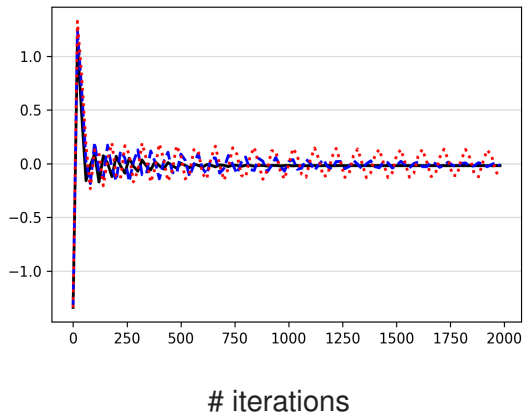
Sublinear convergence

$$\sum_s \|\pi_t(\cdot | s) - \pi^*(\cdot | s)\|^2$$



Smaller regularization $\tau \longrightarrow$ **better** accuracy

$$V_g^{\pi_t}(\rho)$$

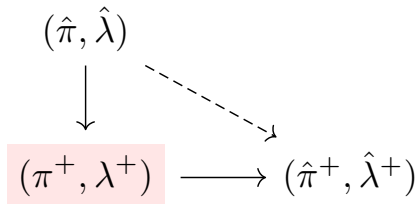


smaller regularization $\tau \longrightarrow$ larger oscillation

■ ACCURACY VS OSCILLATION TRADE-OFF

Optimistic method

Optimistic policy gradient primal-dual method



policy ascent direction: $Q_{r+\hat{\lambda}g}^{\hat{\pi}}(s, a)$

dual descent direction: $V_g^{\hat{\pi}}(\rho)$

policy ascent direction: $Q_{r+\lambda^+g}^{\pi^+}(s, a)$

dual descent direction: $V_g^{\pi^+}(\rho)$

■ SINGLE-TIME-SCALE PRIMAL-DUAL UPDATE

prediction step

$$\begin{aligned}\pi^+(\cdot | s) &= \operatorname{argmax}_{\pi' \in \Delta} \langle \pi'(\cdot | s), Q_{r+\lambda g}^\pi(s, \cdot) \rangle - \frac{1}{2\eta} \|\pi' - \hat{\pi}\|_s^2 \\ \lambda^+ &= \operatorname{argmin}_{\lambda' \in \Lambda} \lambda' V_g^\pi(\rho) + \frac{1}{2\eta} (\lambda' - \hat{\lambda})^2\end{aligned}$$

real step

$$\begin{aligned}\hat{\pi}^+(\cdot | s) &= \operatorname{argmax}_{\pi' \in \Delta} \langle \pi'(\cdot | s), Q_{r+\lambda^+ g}^{\pi^+}(s, \cdot) \rangle - \frac{1}{2\eta} \|\pi' - \hat{\pi}\|_s^2 \\ \hat{\lambda}^+ &= \operatorname{argmin}_{\lambda' \in \Lambda} \lambda' V_g^{\pi^+}(\rho) + \frac{1}{2\eta} (\lambda' - \hat{\lambda})^2\end{aligned}$$

Non-asymptotic last-iterate performance

distance of $(\hat{\pi}_t, \hat{\lambda}_t)$ to $\Pi^* \times \Lambda^*$: $\mathbb{E} [\text{Dist}_s(\hat{\pi}_t, \Pi^*)] + \text{Dist}(\hat{\lambda}_t, \Lambda^*)$

Theorem (informal)

★ **Distance of $(\hat{\pi}_t, \hat{\lambda}_t)$ to $\Pi^* \times \Lambda^*$ is bounded by**

$$\left(\frac{1}{1 + C} \right)^t$$

Exponential decay to zero

C, η – problem-dependent constants

optimality gap: $V_r^*(\rho) - V_r^{\pi_t}(\rho)$

feasibility gap: $0 - V_g^{\pi_t}(\rho)$

Implication (informal)

★ **Achieving ϵ -optimality gap and ϵ -feasibility gap** requires

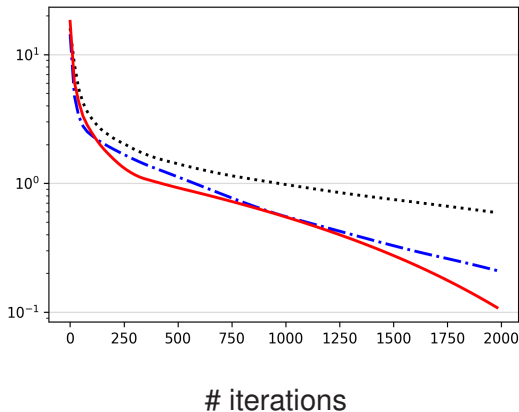
$$\log^2 \frac{1}{\epsilon} \text{ iterations}$$

Linear rate

problem-dependent

Linear convergence

$$\sum \|\hat{\pi}_t(\cdot | s) - \pi^*(\cdot | s)\|^2$$



($\cdots$) $\eta = 0.05$

($-\cdot-$) $\eta = 0.1$

($—$) $\eta = 0.2$

Linear rate holds for a range of stepsizes

From reinforcement learning to generative modeling

■ REQUIREMENT-DRIVEN REINFORCEMENT LEARNING

QUESTION

Can direct policy search methods find
an optimal constrained policy?



**Best- or last-iterate
convergence**

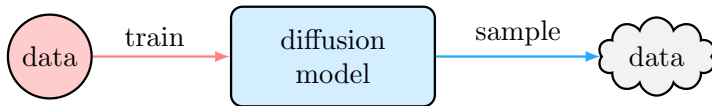
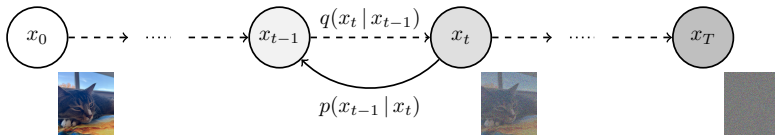
- ★ Natural policy gradient primal-dual method
- ★ Regularized policy gradient primal-dual method
- ★ Optimistic policy gradient primal-dual method

Part II

Requirement-Driven Generative Modeling

(controlled generation of diffusion model and large language model)

Requirement-driven diffusion model



data – image, trajectory, video, protein

Image generation (DALL-E)



Goal	generating images
Requirements	diversity
	aestheticity
	⋮

Constrained diffusion model

■ CONSTRAINED MODEL OPTIMIZATION

$$\begin{aligned} & \underset{\theta}{\text{minimize}} && D_{\text{KL}}(q(x_{0:T}) \parallel p_{\theta}(x_{0:T})) \\ & \text{subject to} && D_{\text{KL}}(q^i(x_{0:T}) \parallel p_{\theta}(x_{0:T})) \leq b_i \quad \text{for } i = 1, \dots, m \end{aligned}$$

θ – diffusion model parameter

★ Use case of q^i : fairness to minority classes

High-dimensional, nonconvex $\implies$ no strong duality

Optimal dualization

■ CONSTRAINED DISTRIBUTION OPTIMIZATION

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

$$\star \text{ Lagrangian } L(p, \lambda) = D_{\text{KL}}(q \parallel p) + \sum_{i=1}^m \lambda_i (D_{\text{KL}}(q^i \parallel p) - b_i)$$

Infinite-dimensional, convex $\implies$ **strong duality**

$$p^\star(x_{0:T}) = \underset{p}{\text{minimize}} L(p, \lambda^\star)$$

$$\text{optimal dual } \lambda^\star \in \underset{\lambda \geq 0}{\operatorname{argmax}} D(\lambda) := \min_p L(p, \lambda)$$

Gap between p and p_θ

My contribution: Dual training

K[†]D[†]R, NeurIPS '24

■ CONSTRAINED DIFFUSION MODEL TRAINING METHOD

dual subgradient ascent w/ approximate primal minimization

Theorem (informal)

★ Gap between trained and optimal models is bounded by $\frac{1}{\sqrt{T}} + \nu + \epsilon$.

T – total diffusion steps

ν, ϵ – parametrization and optimization errors

Optimal constrained model

primal problem: $\underset{p}{\text{minimize}} \quad L(p, \lambda)$

mixture distribution: $q_{\text{mix}}^{(\lambda)}(x_{0:T}) := \frac{1}{1 + \mathbf{1}^\top \lambda} \sum_{i=1}^m \lambda_i q^i(x_{0:T})$

■ EQUIVALENT PRIMAL PROBLEM

$$\underset{p}{\text{minimize}} \quad D_{\text{KL}} \left(q_{\text{mix}}^{(\lambda)}(x_{0:T}) \parallel p(x_{0:T}) \right)$$

★ Optimal constrained model $p^*(x_{0:T}) = q_{\text{mix}}^{(\lambda^*)}(x_{0:T})$

Score-based dual training

score matching: $\mathbb{E}_{q(x_0), t, x_t} [\|\hat{s}_\theta(x_t, t) - \nabla \log q(x_t)\|^2]$

dual function: $\bar{g}_s(\lambda) = \operatorname{argmin}_{\theta} L_s(\hat{s}_\theta, \lambda)$

■ DUAL TRAINING

$$\begin{aligned}\hat{s}_\theta^+ &\in \operatorname{argmin}_{\theta} L_s(\hat{s}_\theta, \lambda) \\ \lambda_i^+ &= \left[\lambda_i + \eta \left(\mathbb{E}_{q^i(x_0), t, x_t} [\|\hat{s}_\theta^+(x_t, t) - \nabla \log q^i(x_t)\|^2] - \tilde{b}_i \right) \right]_+\end{aligned}$$

$\tilde{b}_i$ – rescaled margin

Dual approximation $\bar{\lambda}^* \in \operatorname{argmax}_{\lambda \geq 0} \bar{g}_s(\lambda)$

Non-asymptotic performance

trained model: $\bar{p}^*(\bar{\lambda}^*) \in \operatorname{argmin}_{\theta} L_s(\hat{s}_{\theta}, \bar{\lambda}^*)$

optimality gap: $\operatorname{TV}(q_{\text{mix}}^*, \bar{p}^*(\bar{\lambda}^*))$

Theorem (informal)

★ **Optimality gap of $\bar{p}^*(\bar{\lambda}^*)$ is bounded by**

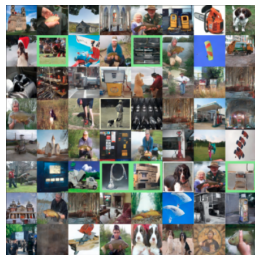
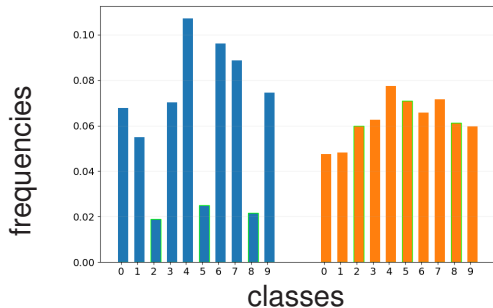
$$\operatorname{TV}(q_{\text{mix}}^*, q_{\text{mix}}^{(\bar{\lambda}^*)}) + \operatorname{TV}(q_{\text{mix}}^{(\bar{\lambda}^*)}, \bar{p}^*(\bar{\lambda}^*)) \approx \frac{1}{\sqrt{T}} + \sqrt{\nu} + \epsilon_{\text{score}}$$

Sublinear decay up to two errors

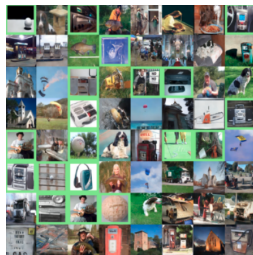
T – total diffusion steps

$\nu, \epsilon_{\text{score}}$ – parametrization and score matching errors

Constraints promote diversity



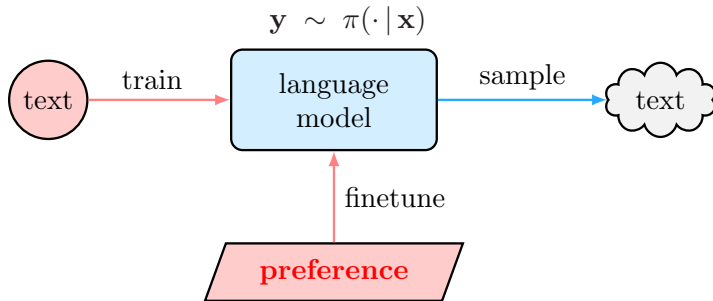
unconstrained



constrained

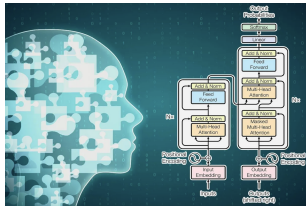
- ★ Unconstrained model (■)
- ★ Our constrained model (■)

Requirement-driven large language model



preference – reward / human feedback

Text generation (ChatGPT)



Goal	generating responses
Requirements	harmlessness
	truthfulness
	⋮

Constrained alignment

■ CONSTRAINED DISTRIBUTION OPTIMIZATION

$$\underset{\pi}{\text{maximize}} \quad \mathbb{E}_{\mathbf{x}} \left[\mathbb{E}_{\mathbf{y} \sim \pi(\cdot | \mathbf{x})} [r(\mathbf{x}, \mathbf{y})] - \beta D_{\text{KL}}(\pi(\cdot | \mathbf{x}) \parallel \pi_{\text{ref}}(\cdot | \mathbf{x})) \right]$$

$$\text{subject to} \quad \mathbb{E}_{\mathbf{x}} \left[\mathbb{E}_{\mathbf{y} \sim \pi(\cdot | \mathbf{x})} [g_i(\mathbf{x}, \mathbf{y})] - \mathbb{E}_{\mathbf{y} \sim \pi_{\text{ref}}(\cdot | \mathbf{x})} [g_i(\mathbf{x}, \mathbf{y})] \right] \geq b_i$$

for $i = 1, \dots, m$

★ Use case of g_i : harmlessness

High-dimensional, convex $\implies$ strong duality

Optimal dualization

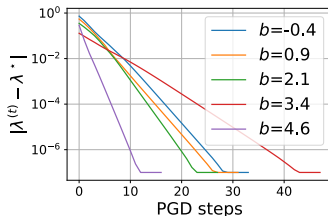
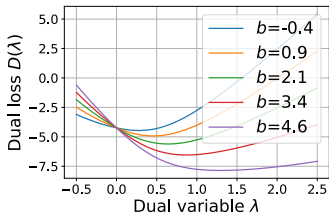
■ EXPLICIT DUAL FUNCTION

$$D(\lambda) = \beta \mathbb{E}_{\mathbf{x}} \left[\ln \mathbb{E}_{\mathbf{y} \sim \pi_{\text{ref}}(\cdot | \mathbf{x})} \left[\exp \left(\frac{r(\mathbf{x}, \mathbf{y}) + \langle \lambda, h(\mathbf{x}, \mathbf{y}) \rangle}{\beta} \right) \right] \right]$$

$$h_i(\mathbf{x}, \mathbf{y}) = g_i(\mathbf{x}, \mathbf{y}) - \mathbb{E}_{\mathbf{y} \sim \pi_{\text{ref}}(\cdot | \mathbf{x})} [g_i(\mathbf{x}, \mathbf{y})] - b_i - \text{safety margin}$$

Independent of π , smooth, and strong convexity

$\Rightarrow$ **easy** to find λ^*



My Contribution: One-shot safety alignment

H[†]L[†]DBLH^{D†}, NeurIPS '24 (**spotlight**)

■ NON-ITERATIVE CONSTRAINED ALIGNMENT METHOD

stage 1: find an optimal dual $\lambda^* \in \underset{\lambda \geq 0}{\operatorname{argmax}} D(\lambda)$

stage 2: solve the primal problem $\pi^* = \underset{\pi}{\operatorname{maximize}} L(\pi, \lambda^*)$

■ PRACTICAL IMPLEMENTATIONS

model-based setting

model-based dual

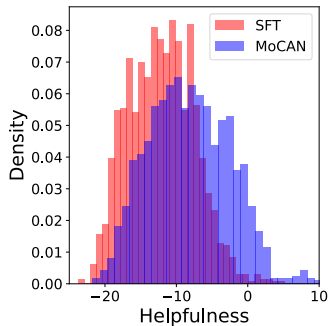
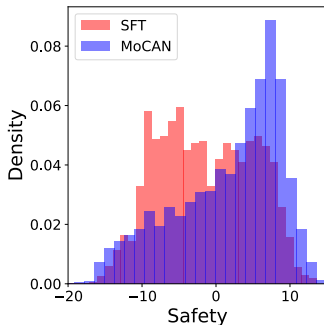
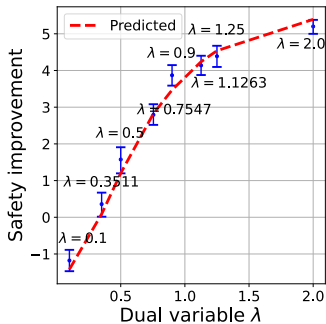
pseudo-preference optimization

preference-based setting

preference-based dual

pseudo-preference optimization

Non-iterative method reduces computational cost



Cost of **non-iterative** method $\ll$ cost of previous **iterative** PPO, DPO

From reinforcement learning to generative modeling

■ REQUIREMENT-DRIVEN GENERATIVE MODELING

QUESTION

How to efficiently impose constraints
on **large generative models**?



**Convexity in
distribution space**

- ★ Dual-based training method
- ★ One-shot safety alignment method

From reinforcement learning to generative modeling

■ REQUIREMENT-DRIVEN REINFORCEMENT LEARNING

QUESTION

Can direct policy search methods find
an optimal constrained policy?



**Best- or last-iterate
convergence**

■ REQUIREMENT-DRIVEN GENERATIVE MODELING

QUESTION

How to efficiently impose constraints
on **large generative models?**



**Convexity in
distribution space**

■ FUTURE DIRECTIONS

Future directions

■ RISK-AWARENESS, MULTI-AGENCY, ADAPTIVITY OF CONSTRAINED REINFORCEMENT LEARNING

Autonomy



Energy



Robotics



Healthcare



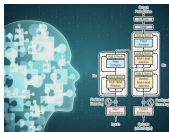
Collaboration areas: Autonomous and Cyber-Physical Systems (EECS, MABE), Energy Networks (EECS, CURENT), Robotics (EECS), Healthcare (EECS, ISE)

- ★ Risk-constrained reinforcement learning
- ★ Multi-agent constrained reinforcement learning
- ★ Online/offline constrained reinforcement learning

Future directions

■ MATH FOUNDATION OF CONSTRAINED GENERATIVE MODELING

Language



Image



Robotics



Medical



Collaboration areas: Generative Modeling (EECS), Robotics (EECS)

- ★ Safety / physics constraints
- ★ Composition of generative models
- ★ Optimal control for generative modeling

■ INTERPLAY BETWEEN GENERATIVE MODELING AND REINFORCEMENT LEARNING

Thank you for your attention.