Distributional Alignment Games

for Answer-Level Fine-Tuning

LTSS Copenhagen Workshop 2026

Mehryar Mohri

Google Research & Courant Institute

Joint work with Jon Schneider, Yifan Wu, and Yutao Zhong

June 27, 2026

Roadmap

- Part I: Answer-Level Fine-Tuning as a Game
  - Solving the Game: GRPO-Based Algorithms
  - Coherence: Self-Improvement via Consensus
  - Experimental Results
- Part II: Resolving the Estimation Bias
  - Summary & Open Directions

Part I: Answer-Level Fine-Tuning as a Game

Motivation: Why Answer-Level?

In reasoning-intensive tasks (math, code, QA), what matters is the final answer z, not the specific reasoning trace y.

Process Supervision

- Guide each step of CoT
- Requires step-level labels
- Expensive annotation

Outcome Supervision (ALFT)

- Only evaluate final answer
- Many valid reasoning paths
- Natural for math, code, QA

The fundamental difficulty

Even when we know the correct answer z, optimizing at the answer level is **intractable**: the policy $\pi(y \mid x)$ lives in *trace space*, but the objective depends on the *answer marginal* $\nu_\pi(z \mid x)$, which requires summing over an exponentially large pre-image.

And beyond correctness: What if z is unknown? What if we want distributional properties (diversity, coherence, safety)?

The ALFT Problem: Formal Setup

- $\mathcal{X}$: input prompts $\quad \mathcal{Y}$: reasoning traces $\quad \mathcal{Z}$: final answers
- Deterministic extraction: $\mathsf{E}\colon \mathcal{Y} \to \mathcal{Z}$ (e.g., parse last number)
- Policy $\pi(y \mid x)$ induces answer marginal:

$$\nu_\pi(z \mid x) \;=\; \sum_{y \,\in\, \mathsf{E}^{-1}(z)} \pi(y \mid x)$$

ALFT Optimization Problem

$$\min_{\pi \in \Pi} \; \mathcal{J}(\pi) \;=\; \mathbb{E}_x\big[\mathcal{R}\big(\nu_\pi(\cdot \mid x)\big) + \beta\, \mathsf{D}_{\mathrm{KL}}\big(\pi(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x)\big)\big]$$

- $\mathcal{R}$: convex functional encoding the answer-level goal
- $\beta > 0$: regularization strength
- π_0: reference policy (pre-trained LLM)

Why Is ALFT Hard?

The Intractability Bottleneck

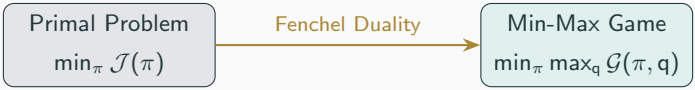
Computing $\nabla_\pi \mathcal{R}(\nu_\pi)$ requires marginalizing over the vast, unknown pre-image $\mathrm{E}^{-1}(z)$.

1. **Many-to-one mapping:** Thousands of traces y produce the same answer z
2. **Non-differentiable extraction:** E may involve code execution, parsing, etc.
3. **High variance:** REINFORCE must assign credit to a specific trace without knowing all alternatives

Standard gradient estimators are prohibitively high-variance and require impractically large sample sizes.

Key Idea: Lift to a Distributional Game

Shift in perspective: Instead of attacking the marginalization sum directly, introduce an auxiliary Target Distribution q as a variational proxy.



Fenchel duality: Since $\mathcal{R}$ is convex and l.s.c.,

$$\mathcal{R}(\nu) = \sup_{u \in \mathbb{R}^{|\mathcal{Z}|}} \left\{ \langle \nu, u \rangle - \mathcal{R}^*(u) \right\} \qquad \text{(Fenchel–Moreau)}$$

Reparametrize $u(z) = -\beta \log \mathrm{q}(z)$ with $\mathrm{q} \in \Delta(\mathcal{Z})$.

The Distributional Alignment Game

Game Objective

$$\mathcal{G}(\pi, \mathsf{q}) \;=\; \mathop{\mathbb{E}}_{x}\Big[\beta\, \mathrm{D}_{\mathrm{KL}}\big(\pi(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x)\big) - \beta \mathop{\mathbb{E}}_{y\sim\pi(\cdot|x)}\big[\log \mathsf{q}(\mathsf{E}(y))\big] - \mathcal{R}^*(-\beta \log \mathsf{q})\Big]$$

Two players:

Policy π (minimizer)

- Generates reasoning traces
- Tries to align with Target
- Regularized by π_0

Target q (maximizer)

- Auxiliary answer distribution
- Encodes the alignment goal
- Adapts to challenge Policy

Why the Game Works: Key Decoupling

The crucial algebraic step:

$$\begin{aligned}\big\langle \nu_\pi(\cdot \mid x),\ -\beta \log \mathsf{q} \big\rangle &= -\beta \sum_{z} \nu_\pi(z \mid x) \log \mathsf{q}(z) \\ &= -\beta \sum_{z} \sum_{y \in \mathsf{E}^{-1}(z)} \pi(y \mid x)\ \log \mathsf{q}(z) \\ &= -\beta \sum_{y \in \mathcal{Y}} \pi(y \mid x)\ \log \mathsf{q}(\mathsf{E}(y)) \\ &= -\beta \mathop{\mathbb{E}}_{y \sim \pi(\cdot \mid x)} \big[\log \mathsf{q}(\mathsf{E}(y))\big]\end{aligned}$$

What happened?

The intractable sum over $\mathsf{E}^{-1}(z)$ *disappeared*.

We moved from marginals $\nu_\pi(z)$ to trace-level expectations $\mathbb{E}_{y \sim \pi}[\cdot]$, which we can estimate by simply sampling traces.

Consistency: Nash Equilibrium = Optimal ALFT

Theorem (Consistency of the Game)

Let $\mathcal{R}$ be convex and l.s.c. Then:

1. **Equivalence:** $\min_{\pi \in \Pi} \mathcal{J}(\pi) = \min_{\pi \in \Pi} \max_{q \in \Delta(\mathcal{Z})} \mathcal{G}(\pi, q)$
2. **Optimal Policy:** For fixed q, the unique optimum is

$$\pi^*(y \mid x) \propto \pi_0(y \mid x)\, q(\mathsf{E}(y) \mid x)$$

Interpretation:

- π^* is the KL-projection of π_0 onto $\{\pi : \mathbb{E}_{y \sim \pi}[\log q^*(\mathsf{E}(y))] \geq c^*\}$
- Traces producing high-q^* answers are upweighted; unlikely traces under π_0 are penalized
- Strict convexity–concavity $\Rightarrow$ **unique Nash Equilibrium**; no-regret dynamics converge at $O(1/\sqrt{T})$

Approximation Guarantee

Proposition

Let $(\widehat{\pi}, \widehat{\mathrm{q}})$ be an ϵ-approximate equilibrium of $\mathcal{G}$. Then

$$\mathcal{J}(\widehat{\pi}) \;\leq\; \mathcal{J}(\pi^*) + \epsilon.$$

Why this matters:

- In practice, we never find exact equilibria
- The game formulation is *robust*: approximate play $\Rightarrow$ approximately optimal ALFT policy

Instantiations: One Game, Many Alignment Goals

Different choices of $\mathcal{R} \Rightarrow$ different games and targets:

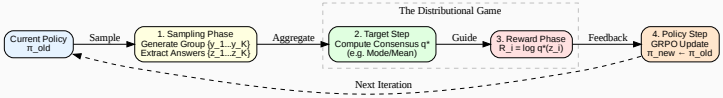
Goal	**Functional** $\mathcal{R}(\nu)$	**Optimal Target** q^*
Standard RL	$-\mathbb{E}_{z\sim\nu}[r(z)]$	$\mathsf{q}^* \propto \exp(r(z)/\beta)$
SFT / RLVR	$\mathrm{D}_{\mathrm{KL}}(\delta_{\mathsf{gt}}\|\nu)$	$\mathsf{q}^* = \delta_{\text{ground truth}}$
Diversity	$-H(\nu)$ (neg. entropy)	$\mathsf{q}^* \propto 1/\nu_\pi(z)$ (inverse freq.)
Safety	$\mathbb{I}_{\mathcal{C}}(\nu)$ (constraint set)	$\mathsf{q}^* = \mathrm{Proj}_{\mathcal{C}}(\nu_\pi)$
Coherence	expected divergence	$\mathsf{q}^* =$ consensus (geo. mean)

Key observation: The SFT/RLVR row is a fixed Dirac target — exactly what DeepSeek-R1, Gemini, etc. do. Our framework gives it theory, but more importantly, it **generalizes** to the rows below, where no ground truth is needed.

RLVR is the special case where someone hands you the answer. The interesting question is: what happens when nobody does?

Solving the Game: GRPO-Based Algorithms

The Game-Theoretic Alignment Loop



Alternating Best Response:

1. **Target Step:** Estimate optimal q^* from sampled group (consensus, inverse-freq, etc.)
2. **Policy Step:** Compute rewards $R_i = \log q^*(\mathrm{E}(y_i))$, derive advantages, update via GRPO

Game-Derived Rewards for GRPO

GRPO (Group Relative Policy Optimization): For each input x, sample a group $G = \{y_1, \ldots, y_K\}$ from π_{old}.

Our insight: The Nash Equilibrium condition $\pi^*(y) \propto \pi_0(y)\, \mathsf{q}^*(\mathsf{E}(y))$ defines the natural reward:

$$R(y) \;=\; \beta \, \log \mathsf{q}^*(\mathsf{E}(y))$$

Advantage (group-relative baseline for variance reduction):

$$A_i \;=\; \frac{R_i - \mathrm{Mean}(\{R_j\}_{j=1}^K)}{\mathrm{StdDev}(\{R_j\}_{j=1}^K) + \epsilon}$$

Policy update:

$$\mathcal{L}_{\text{GRPO}}(\pi) \;=\; \mathop{\mathbb{E}}_{x}\left[\mathop{\mathbb{E}}_{G\sim\pi_{\text{old}}}\left[\frac{1}{K}\sum_{i=1}^{K}\frac{\pi(y_i \mid x)}{\pi_{\text{old}}(y_i \mid x)}\, A_i - \beta\, \mathrm{D}_{\mathrm{KL}}(\pi \,\|\, \pi_{\text{ref}})\right]\right]$$

Coherence: Self-Improvement via Consensus

Coherence: The Self-Improvement Principle

Core idea [Mohri, Schneider, Wu 2025]: A reliable model should produce consistent answer distributions for semantically equivalent inputs.

Let $\Phi\colon \mathcal{X} \to \mathcal{X}$ be a task-preserving transformation (e.g., paraphrase).

$$\mathcal{C}_{\mathrm{coh}} \;=\; \big\{\pi \;\mid\; \nu_\pi(\cdot \mid x) = \nu_\pi(\cdot \mid \Phi(x)) \quad \forall x\big\}$$

Self-Improvement as Bregman Projection

$$\widehat{\pi} \;=\; \operatorname*{argmin}_{\pi \in \mathcal{C}_{\mathrm{coh}}} \mathop{\mathbb{E}}_{x}\big[D_F\big(\pi(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x)\big)\big]$$

Guarantee (monotonic improvement): $\widehat{\pi}$ is strictly closer to π^* than π_0, provided π^* is coherent.

Consensus Targets: From Geometric Mean to Majority Vote

Ideal target (Geometric Mean):

$$q^*_{\mathrm{GM}}(z) \;=\; \frac{1}{Z}\left(\prod_{x' \in \mathcal{O}_x} \nu(z \mid x')\right)^{1/|\mathcal{O}_x|}$$

(exact solution for KL; intractable partition function)

Practical relaxation (Arithmetic Mean):

$$q_{\mathrm{AM}}(z) \;=\; \frac{1}{|\mathcal{O}_x|} \sum_{x' \in \mathcal{O}_x} \nu(z \mid x')$$

Stability of Arithmetic Consensus (Theorem)

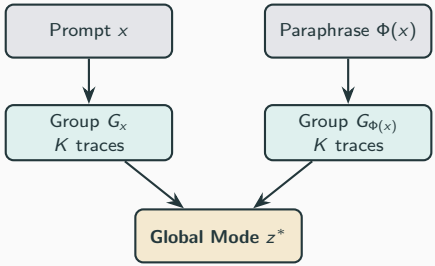
Improvement degradation relative to ideal geometric mean is bounded by the **Generalized Squared Hellinger distance**:

$$\mathrm{Improv}(\widehat{\pi}_{\mathrm{Hybrid}}) \;\geq\; \mathrm{Improv}(\widehat{\pi}_{\mathrm{Ideal}}) - 2LB\, \mathop{\mathbb{E}}_{x}\!\left[H^2\big(\{\nu_k(\cdot \mid x)\}\big)\right]$$

Coherence-GRPO: The Algorithm

For each prompt x:

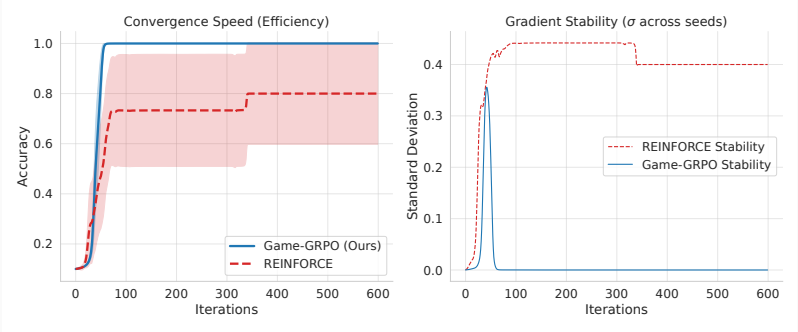
1. **Orbit Sampling:**
   - Generate $\mathcal{O}_x = \{x, \Phi(x)\}$
   - Sample K traces per input
2. **Target Step:**
   - Extract all answers in orbit
   - Compute Global Mode z^*
3. **Reward:**
   - $R_i = \mathbb{I}(\mathsf{E}(y_i) = z^*)$
   - Match to *global* consensus
4. **Update:** GRPO on all traces



Pairwise-GRPO: For open-ended domains, replace exact-match mode with pairwise semantic distance $d(y, y')$. Reward = average agreement with the orbit group.

Experimental Results

Synthetic Validation: Variance Reduction



- $|\mathcal{Y}| = 10{,}000$, $|\mathcal{Z}| = 10$ (1000:1 redundancy)
- Stochastic difficulty: Easy (bias 8.0) vs. Hard (bias 0.0)
- **Game-GRPO** converges to 100% accuracy; REINFORCE baseline plateaus at ≈80%

GSM8K Results: Self-Improvement Without Ground Truth

Algorithm	Model	Baseline ACC	Our ACC	Abs. ↑	Rel. ↑
Pairwise-GRPO	Qwen-3B	75.06	79.61	**+4.55**	+6.06%
	Llama	66.19	72.71	**+6.52**	+9.85%
	Phi-3	73.69	82.87	**+9.18**	+12.46%
Coherence-GRPO	Qwen-3B	75.06	80.36	**+5.30**	+7.06%
	Llama	66.19	69.37	**+3.18**	+4.80%
	Phi-3	73.69	81.50	**+7.81**	+10.60%

- **No ground-truth labels used.** Improvement is purely from coherence-based self-play
- Consistent gains across all 3 models (3B-class)
- Greedy decoding at test time

TriviaQA Results: Pairwise vs. Coherence

Algorithm	Model	Base EM	Our EM	Rel. ↑
Pairwise-GRPO	Qwen	32.95	35.32	+7.19%
	Llama	39.12	47.85	+22.32%
	Phi-3	32.03	45.50	**+42.06%**
Coherence-GRPO	Qwen	32.95	32.90	−0.15%
	Llama	39.12	40.25	+2.89%
	Phi-3	32.03	32.00	−0.09%

Key insight:

- COHERENCE-GRPO relies on exact-match extractor → fails on open-ended QA
- PAIRWISE-GRPO uses semantic distance → handles "NYC" vs. "New York City"

⇒ Use COHERENCE for discrete answers, PAIRWISE for open-ended

Safety Extension: Safety-GRPO

Safety as a distributional constraint: $\mathcal{R}(\nu) = \mathbb{I}_{\mathcal{C}}(\nu)$ (indicator of safe set)

Metric	Base	SAFETY-GRPO	Δ
Benign Non-Refusal	1.000	0.990	-0.010
Harmful Toxicity	0.073	0.068	-0.005
Harmful Toxic @0.5	0.055	0.040	-0.015

Primal-dual update successfully enforces safety budgets:
reduces harmful toxicity while preserving benign helpfulness.

Mechanism: Lagrange multipliers $\rightarrow$ dynamic penalty weights;
target $= \mathrm{q}^* = \mathrm{Proj}_{\mathcal{C}}(\nu_\pi)$ (information projection onto safe set).

Part II: Resolving the Estimation Bias

The Hidden Bias in Small-Batch ALFT

Recall: the game-derived reward is $R(z) = \beta \log \mathrm{q}^*(z)$.

In practice, q^* is estimated empirically: $\widehat{\mathrm{q}}(z) = X/K$ where $X \sim \mathrm{Binomial}(K, \mathrm{q}^*(z))$.

Jensen's Inequality Bias

$$\mathbb{E}\big[\log \widehat{\mathrm{q}}(z)\big] \;\approx\; \log \mathrm{q}^*(z) \;-\; \frac{1 - \mathrm{q}^*(z)}{2K\,\mathrm{q}^*(z)}$$

Consequences:

- $\mathcal{O}(1/K)$ systematic bias — does *not* vanish for fixed K
- Bias magnitude $\propto 1/\mathrm{q}^*(z)$: penalizes rare answers most
- Acts as an artificial **anti-exploration** cost $\Rightarrow$ premature mode collapse
- Rare answers: $X = 0 \Rightarrow \log 0 = -\infty$

Solution 1: Change the Geometry (Polynomial Rewards)

Generalize the game from KL to arbitrary Bregman divergences D_F:

$$\min_{\pi \in \Pi} \; \mathop{\mathbb{E}}_{x}\big[\mathcal{R}(\nu_\pi(\cdot \mid x)) + \beta\, \mathsf{D}_F\big(\pi(\cdot \mid x) \,\|\, \pi_0(\cdot \mid x)\big)\big]$$

Key Insight

If the distance-generating function $\Phi(\mathsf{q})$ is a polynomial of degree $d+1$, then the reward $R(z) = \beta\, \nabla\Phi(\mathsf{q})_z$ is a **polynomial of degree** d in $\mathsf{q}(z)$.

Theorem (Exact Unbiased Estimation via U-statistics)

For polynomial reward of degree $d \le K$, the falling factorial

$$\widehat{\mathsf{q}}^{\,m}(z) \;=\; \frac{X(X-1)\cdots(X-m+1)}{K(K-1)\cdots(K-m+1)}$$

is the unique minimum-variance unbiased estimator for $\mathsf{q}(z)^m$.

Example (Euclidean game): $R(z) = \beta(1-\mathsf{q}(z)) \;\Rightarrow\; \widehat{R}(z) = \beta(1 - X/K)$ (exactly unbiased!)

Solution 2: Optimal Approximation for KL Games

For the canonical KL divergence (log reward), *exact* unbiased estimation is impossible.

But: In policy gradients, the bias is weighted by p, and $\lim_{p \to 0} p \log p = 0$.

Minimax polynomial estimator: Pre-compute optimal coefficients $\mathbf{c}^* = (c_0^*, \ldots, c_K^*)$ by solving:

$$\min_{\mathbf{c} \in \mathbb{R}^{K+1}} \max_{p \in [0,1]} \left| p\, P_{\mathbf{c}}(p) - \beta\, p \log p \right|$$

via a linear program (solved *offline*, once per K).

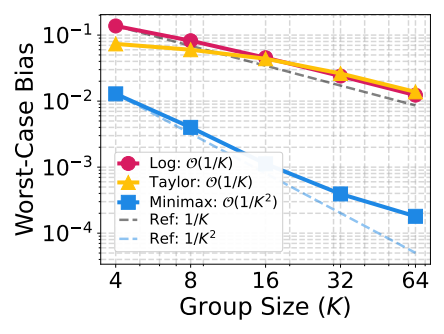
Theorem (Fundamental Limit — Ditzian–Totik)

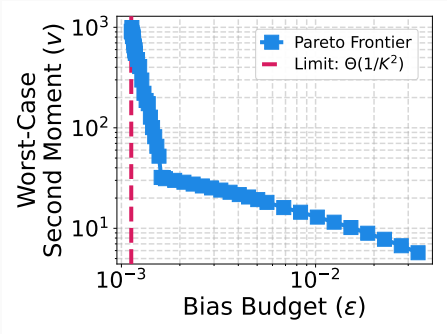
No estimator from K samples can achieve gradient bias better than $\Theta(1/K^2)$.

Our minimax LP estimator **achieves this optimal rate**.

Practical cost: Replace $\log(X/K)$ with table lookup $c_X^* \Rightarrow$ zero online overhead, 3.5× faster.

Bias Rates and the Pareto Frontier





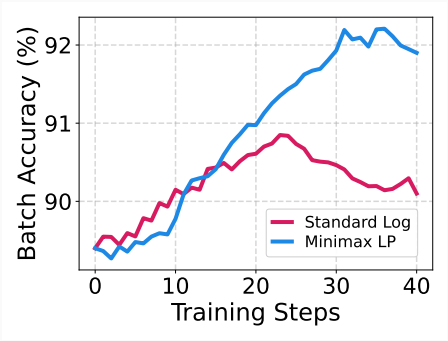
Left: Taylor correction fails at boundary ($\mathcal{O}(1/K)$ worst case).

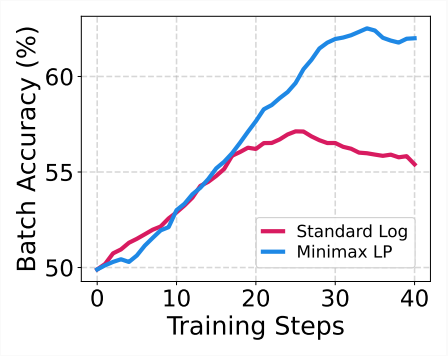
Minimax LP achieves uniform $\mathcal{O}(1/K^2)$.

Right: Bias–Variance Pareto frontier via AQP formulation.

Trade off exactness for lower variance $\Rightarrow$ accelerated convergence.

Impact on LLM Fine-Tuning





Model	Estimator	GSM8K Acc. (%)	MathQA Acc. (%)
Qwen-7B	Std. Log	90.7 ± 1.1	57.0 ± 1.4
	Minimax LP	$\mathbf{92.1 \pm 0.4}$	$\mathbf{62.3 \pm 0.5}$
Llama-8B	Std. Log	85.6 ± 1.2	49.1 ± 0.9
	Minimax LP	$\mathbf{87.2 \pm 0.3}$	$\mathbf{52.8 \pm 0.6}$

Summary & Open Directions

Summary

1. **Distributional Alignment Games** Lift intractable ALFT to a tractable min-max game via Fenchel duality. Nash Equilibrium = optimal ALFT policy.

2. **Unified framework** Diversity, coherence, safety, standard RL = different Target strategies in the same game.

3. **Scalable algorithms** COHERENCE-GRPO and PAIRWISE-GRPO: significant gains on GSM8K (+3–9 pp) and TriviaQA (+42% EM), entirely *without ground truth*.

4. **Resolving estimation bias** Generalized Bregman games admit exact unbiased estimators. For KL: minimax polynomial achieves the $\Theta(1/K^2)$ fundamental limit with zero overhead.

Open Directions

- **Richer orbit structures** Extension to multi-lingual, multi-modal, and compositional equivalences is straightforward
- **Adaptive group sizes** Dynamically adjust K based on task difficulty and estimation precision requirements
- **Online target learning** Learning q^* with function approximation instead of computing it from batch statistics
- **Beyond single-turn** Extend to multi-turn conversations and compositional reasoning with intermediate verification
- **Tighter convergence theory** Finite-time rates for the alternating best-response dynamics in the practical LLM setting

Thank you!

Questions?

mohri@google.com

Backup: Duality with DPO

Proposition (Duality with Direct Preference Optimization)

The optimal target distribution q^* corresponds to the exponentiated ground-truth reward of the equivalent RL problem:

$$q^*(E(y)) = \frac{1}{Z} \exp\left(\frac{r^*(y)}{\beta}\right)$$

Complementary perspectives:

- **DPO:** Eliminates the reward model $\rightarrow$ solves for policy directly
- **Our game:** Eliminates the policy $\rightarrow$ solves for the optimal target distribution

Implication: For *dynamic* answer-level objectives (diversity, coherence), DPO would require iterative re-generation. Our GRPO-based algorithm is a more direct online solver.

References

- M. Mohri, J. Schneider, Y. Wu. *Distributional Alignment Games for Answer-Level Fine-Tuning*. ICML 2026.
- M. Mohri, J. Schneider, Y. Zhong. *Generalized Distributional Alignment Games for Unbiased Answer-Level Fine-Tuning*. 2026.
- M. Mohri, J. Schneider, Y. Wu. *Self-Improvement via Coherence*. 2025.
- D. Shao et al. *DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models*. 2024. (GRPO)
- R. Rafailov et al. *Direct Preference Optimization*. NeurIPS 2023.
- Y. Li et al. *Inverse-Frequency Reward for Diversity*. 2025.
- Z. Ditzian, V. Totik. *Moduli of Smoothness*. Springer, 1987.

Backup: Generalized Game (Bregman)

Generalized Game Objective

$$\mathcal{G}_F(\pi, \mathsf{q}) \;=\; \mathop{\mathbb{E}}_{x}\Big[\beta\, \mathsf{D}_F(\pi \| \pi_0) + s\beta \mathop{\mathbb{E}}_{y\sim\pi}\big[\nabla\Phi(\mathsf{q})_{\mathsf{E}(y)}\big] - \Psi(\mathsf{q})\Big]$$

Consistency (Theorem): $\min_\pi \mathcal{J}_F(\pi) = \min_\pi \max_{\mathsf{q}} \mathcal{G}_F(\pi, \mathsf{q})$

Optimal policy: Bregman projection

$$\pi^*(\cdot \mid x) \;=\; \mathrm{Proj}^F_{\Delta(\mathcal{Y})}\Big((\nabla F)^{-1}\big(\nabla F(\pi_0(\cdot \mid x)) - s\,\nabla\Phi(\mathsf{q}(\cdot \mid x))_{\mathsf{E}(\cdot)}\big)\Big)$$

Backup: AQP Formulation

Variance-Optimal Augmented Quadratic Program:

$$\begin{aligned} \min_{\mathbf{c}\in\mathbb{R}^{K+1},\, v\in\mathbb{R}} \quad & v \\ \text{s.t.} \quad & \sum_{k=0}^{K} c_k^2\, B_{k,K}(p_m) \le v, \quad \forall\, p_m \in \mathcal{P} \\ & \left| p_m P_{\mathbf{c}}(p_m) - \beta\, p_m \log p_m \right| \le \epsilon, \quad \forall\, p_m \in \mathcal{P} \end{aligned}$$

- Minimizes worst-case second moment subject to bias budget ϵ
- Varying $\epsilon \in [\epsilon^*_{\text{min}}, \infty)$ traces the complete Pareto frontier
- Strictly dominates sample-splitting approaches (Theorem)
- Guarantees $\mathcal{O}(\epsilon)$-approximate equilibrium with accelerated convergence