Coherence Mechanisms for Provable Self-Improvement

Mehryar Mohri
Google Research & CIMS
mohri@google.com

Jon Schneider
Google Research
jschnei@google.com

Yifan Wu
Microsoft Research
yifan.wu2357@gmail.com

Abstract

Self-improvement is a critical capability for large language models and other intelligent systems, enabling them to refine their behavior and internal consistency without external supervision. Despite its importance, prior approaches largely rely on empirical heuristics and lack formal guarantees. In this paper, we propose a principled framework for *coherence-based* self-improvement: a model's outputs should remain consistent under task-preserving transformations of the input (e.g., paraphrases), and the set of coherent models is a convex set that can be accessed using only the model's own, lower-level abilities. We formalize self-improvement as a Bregman projection of a baseline model onto this set and prove that it yields *monotonic improvement*, measured by the expected Bregman divergence to the unknown ground truth, for every coherent ground truth; moreover, the projection is the *maximin-optimal* mechanism. We extend these guarantees to relaxed coherence constraints, which recover the penalized training objective used in practice, to non-realizable ground truths, and to finite samples and inexact optimization. We then study improvement across *families* of divergences: simultaneous improvement is impossible in general, even for two quadratic divergences, but is achieved by an orbit-averaging mechanism for jointly convex families, and we characterize precisely when it is possible. Finally, we prove characterization and *rigidity* results: any mechanism enjoying the Pythagorean improvement guarantee must be a Bregman projection, it inherits every symmetry of the problem, and under affine constraints its outputs across all generators are parametrized by a finite-dimensional dual subspace. Experiments with three instruction-tuned language models on TriviaQA show that coherence training improves both accuracy and consistency across paraphrases.

1 Introduction

The ability to self-improve is a central and defining challenge in the design of large language models (LLMs) and other advanced intelligent systems: it is essential for developing models that are internally consistent, robust, and capable of adapting to complex or evolving tasks without constant, granular human supervision. Despite its importance, the field has largely been guided by empirical heuristics. Without a formal understanding of *why* a given self-improvement mechanism works, there is no guarantee that it will not fail or degrade performance. This paper aims to bridge this gap by proposing a principled, mathematical framework for self-improvement grounded in the concept of *coherence*.

Our core thesis is that a model's internal consistency is a powerful and measurable proxy for its quality. We define coherence as the requirement that a model produce consistent outputs for inputs that are equivalent under a task-preserving transformation: a model should output the same answer distribution for a question x and its semantic paraphrase $\Phi(x)$. An *incoherent* model, one that contradicts itself on equivalent inputs, is demonstrably flawed, and we posit that actively reducing

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

incoherence provides a provable, non-heuristic path to improvement. Importantly, the paraphrases can be produced by the model itself: the model can *self-improve* by exploiting a more reliable, lower-level ability (paraphrasing, or more generally checking a task-preserving equivalence) to improve a higher-level one (answering).

Slightly more formally, consider a setting where we begin with an existing, imperfect model π_0. We have no labeled data with which to train it further, but we know that the ground truth π^* must belong to a convex set of "coherent" models $\mathcal{C}_{\mathrm{coh}}$ to which we have efficient access.[1] Can we use this information to construct an improved model $\widehat{\pi}$ that is *guaranteed* to be better than π_0, regardless of what the ground truth is?

Our contributions. We answer this affirmatively, with mechanisms formulated as *projections*: given a baseline π_0, we project it onto the coherent set to find the coherent model $\widehat{\pi}$ closest to π_0. We use the geometric language of *Bregman divergences*, a broad class of measures including the KL divergence and the squared Euclidean distance, and measure the loss of a candidate $\widehat{\pi}$ by its expected Bregman divergence to the ground truth, written $\mathsf{B}_F(\pi^* \parallel \widehat{\pi})$ here. Our results are as follows.

- **Improvement and optimality.** The Bregman projection $\widehat{\pi}$ of π_0 onto any convex set $\mathcal{C}$ containing π^* satisfies $\mathsf{B}_F(\pi^* \parallel \widehat{\pi}) \le \mathsf{B}_F(\pi^* \parallel \pi_0) - \mathsf{B}_F(\widehat{\pi} \parallel \pi_0)$, hence a quantifiable improvement (Section 4.1); when $\mathcal{C} = \mathcal{C}_{\mathrm{coh}}$ we bound the improvement from below by the measurable incoherence of π_0. Moreover, the projection onto $\mathcal{C}_{\mathrm{coh}}$ is the *maximin* mechanism: no coherent model guarantees a larger worst-case improvement (Section 4.2).
- **Relaxed coherence and the practical objective.** We analyze relaxed constraints, where coherence is only required on average up to a tolerance Λ, and prove an explicit improvement bound in terms of the incoherence of π_0 and Λ (Section 4.3). By Lagrangian duality, the relaxed projection is precisely the "KL-to-baseline plus coherence penalty" objective used in our experiments; the theory thus *derives* the training objective.
- **Robustness and two-step projection.** The guarantees extend to non-realizable ground truths (with explicit dependence on the distance of π^* to $\mathcal{C}_{\mathrm{coh}}$), to projections computed from a finite sample, and to *approximate projections* obtained by inexact optimization within a restricted (e.g., parametric) model class, with explicit error terms (Section 4.4, Section D). We also show that the direct projection coincides with a natural *two-step projection* (onto the cone of coherent functions, a closed-form Bregman centroid, then onto the model class) for all separable and quadratic generators (Section E).
- **Multiple divergences.** We ask whether one can improve simultaneously with respect to several Bregman divergences. We prove that this is impossible in general, even for two quadratic divergences (Section 5.1), but that for invariance mappings and *jointly convex* divergences, an orbit-averaging mechanism improves for all of them simultaneously; we characterize exactly when such universal improvement is possible (Section 5.2).
- **Characterization and rigidity.** Any mechanism enjoying the Pythagorean-type improvement guarantee over a convex set is the Bregman projection onto it; it inherits every symmetry of the problem for every generator; and under affine constraints the outputs across *all* generators are parametrized by a finite-dimensional dual subspace, yielding an exact description of the ties shared by all generators and a local rigidity property for polyhedral constraints (Section 6, Section G).
- **Experiments.** Coherence training of three instruction-tuned LLMs (`Phi-3`, `Qwen2.5`, `Llama-3.2`) on TriviaQA, with self-generated paraphrases, improves both accuracy and consistency across paraphrases (Section 7).

Scope. We study *coherence-based* (self-consistency) improvement: label-free improvement using an efficiently checkable convex constraint on output distributions. We do not claim a general theory of self-rewarding, RLAIF, or recursive data-generation schemes. Our guarantees are stated at the level of output distributions; parametric fine-tuning is addressed through the relaxed objective of Section 4.3 and the approximate-projection bounds of Section 4.4. Scope and limitations are discussed further in Sections 8 and I.

[1] This allows for a far more general notion of coherence than paraphrases alone: $\mathcal{C}_{\mathrm{coh}}$ can encode any efficiently checkable collection of convex constraints, e.g., monotonicity or order constraints (if y should be preferred to y' then $\pi(y \mid x) \ge \pi(y' \mid x)$), invariances or symmetries under task-preserving transformations ("A and B" versus "B and A"), or logical and calibration consistency conditions expressible as linear (in)equalities ($\pi(A \mid x) + \pi(\neg A \mid x) = 1$). While paraphrase coherence is our motivating example, most of our results apply to this general setting.

Related work. The closest line of work, on weak-to-strong generalization, also studies projection-based mechanisms with provable improvement guarantees [Burns et al., 2024, Charikar et al., 2024, Mulgund and Pabbaraju, 2025, Yao et al., 2025a,b]. Our perspective is different: we view coherence as an algorithmically accessible, without-loss way to construct a target-containing convex set directly from weaker model abilities, such as paraphrasing, and we go beyond single-divergence improvement with maximin optimality, relaxed and approximate projections, simultaneous improvement and rigidity results. Coherence is reminiscent of data augmentation and consistency regularization [Xie et al., 2020, Maini et al., 2024], and label-free paraphrase-consistency penalties have been used empirically in NLP [Elazar et al., 2021, Zhou et al., 2022]; the key differences are that we formalize label-free consistency enforcement as a projection onto a target-containing convex set, and that we prove *monotonic improvement with respect to the unknown ground truth* together with optimality and impossibility results, rather than generalization benefits of a supervised objective. Our perspective also connects to self-distillation [Hinton et al., 2015, Allen-Zhu and Li, 2023]: the coherence projection finds the single model that is the Bregman centroid of the "ensemble" of incoherent views $\pi_0(x)$ and $\pi_0(\Phi(x))$. A detailed discussion of related work is given in Section A.

2 Preliminaries

Notation and probabilistic setting. Let Σ be a finite alphabet (e.g., a set of tokens). Prompts are sequences $x \in \mathcal{X} \subseteq \Sigma^*$, and outputs lie in $\mathcal{Y} \subseteq \Sigma^*$. A language model is a conditional probability distribution $\pi\colon \mathcal{X} \to \Delta(\mathcal{Y})$, where $\Delta(\mathcal{Y})$ denotes the probability simplex over $\mathcal{Y}$; we write $\pi(x)$ for the corresponding distribution over outputs. We assume throughout that $\mathcal{X}$ and $\mathcal{Y}$ are finite, and write $d = |\mathcal{Y}|$ and $n = |\mathcal{X}|$. We write $\Pi_{\text{all}} = (\Delta(\mathcal{Y}))^{\mathcal{X}}$ for the set of all conditional distribution models, a compact convex subset of a Euclidean space. Throughout, we work with a closed and convex family of models $\Pi \subseteq \Pi_{\text{all}}$, which is therefore compact. We also define the extended space of unnormalized models $\Pi^{\dagger}_{\text{all}} = (\mathbb{R}^{\mathcal{Y}}_{+})^{\mathcal{X}} \cong \mathbb{R}^{n\times d}_{+}$, a closed convex cone. Expectations are with respect to a fixed prompt distribution $\mathcal{D}_{\mathcal{X}}$ over $\mathcal{X}$, with $\operatorname{supp}(\mathcal{D}_{\mathcal{X}}) = \mathcal{X}$ without loss of generality, and $\mathbb{P}[x] = \mathcal{D}_{\mathcal{X}}(x)$.

Bregman divergences. We adopt the standard definition following Rockafellar [1997] (see also [Bregman, 1967, Censor and Lent, 1981]).

Definition 2.1 (Bregman divergences)**.** *Let $F\colon \mathcal{K} \subseteq \mathbb{R}^d \to \mathbb{R}$ be a convex function defined on a convex set $\mathcal{K}$. The* Bregman divergence $\mathsf{B}_F(\mathsf{p} \,\|\, \mathsf{q})$ *is defined for any $\mathsf{p} \in \mathcal{K}$ and q in the interior of $\operatorname{dom}(F)$ where F admits a sub-gradient $g_{\mathsf{q}} \in \partial F(\mathsf{q})$, as the gap between $F(\mathsf{p})$ and the first-order approximation of F at q: $\mathsf{B}_F(\mathsf{p} \,\|\, \mathsf{q}) = F(\mathsf{p}) - F(\mathsf{q}) - \langle g_{\mathsf{q}}, \mathsf{p} - \mathsf{q}\rangle$.*

We restrict attention to the common case where F is differentiable on the interior of its domain, so that $\mathsf{B}_F(\mathsf{p} \,\|\, \mathsf{q}) = F(\mathsf{p}) - F(\mathsf{q}) - \langle \nabla F(\mathsf{q}), \mathsf{p} - \mathsf{q}\rangle$. Our primary setting is that of distributions in $\Delta(\mathcal{Y})$; to ensure that the divergence is well defined, we only assume that $\mathcal{K} \supseteq \Delta(\mathcal{Y})$ and that $\operatorname{int}(\mathcal{K})$ contains $\operatorname{relint}(\Delta(\mathcal{Y}))$ (e.g., $\mathcal{K} = \mathbb{R}^d_+$ or $\mathcal{K} = \mathbb{R}^d$). Throughout the paper, the generator $F\colon \mathcal{K} \to \mathbb{R}$ is of *Legendre type* (or simply *Legendre*): it is differentiable and strictly convex on $\operatorname{int}(\mathcal{K})$, and ∇F is a bijection from $\operatorname{int}(\mathcal{K})$ onto $\operatorname{int}(\operatorname{dom} F^*)$. The negative entropy (on $\mathcal{K} = \mathbb{R}^d_+$) and the squared norm (on $\mathcal{K} = \mathbb{R}^d$) are Legendre and generate the (unnormalized) relative entropy D_{KL} and the squared Euclidean distance. We assume that $\pi_0(x) \in \operatorname{int}(\mathcal{K})$ for all x and that every feasible set onto which we project meets $(\operatorname{int}\mathcal{K})^{\mathcal{X}}$; Bregman projections then lie in $(\operatorname{int}\mathcal{K})^{\mathcal{X}}$, so all divergences below are well defined (Section B.2). The properties of Bregman divergences and Legendre functions that we use are reviewed in Sections B.1 and B.2. A key fact, used repeatedly, is that $\pi \mapsto \mathsf{F}(\pi) = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[F(\pi(x))]$ is itself a Legendre function on the space of models, whose Bregman divergence is $\mathsf{B}_{\mathsf{F}}(\pi \,\|\, \pi') = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi(x) \,\|\, \pi'(x))]$ (Lemma B.6).

3 Coherence as a Principle of Self-Improvement

Invariance mappings. Let $\Phi\colon \mathcal{X} \to \mathcal{X}$ be a mapping on inputs (e.g., prompts). We call Φ an *invariance mapping* if, under an ideal model π^*, outputs are preserved: $\pi^*(x) = \pi^*(\Phi(x))$ for all $x \in \mathcal{X}$, where equality is interpreted under a task-preserving equivalence relation (e.g., syntactic, semantic, answer-based, or distributional). In terms of self-improvement, Φ can be obtained by asking the language model itself to paraphrase a prompt, or to judge the equivalence of two prompts. For example, the passive form of a sentence, a semantically equivalent paraphrase, or a rotated image should lead to the same output. Conceptually, these mappings are close to Zellig Harris's meaning-preserving

transformations [Harris, 1968]. When $\mathcal{X}$ is finite, Φ partitions $\mathcal{X}$ into orbits, i.e., equivalence classes $\mathcal{X}_i$, $i \in I$, of inputs treated as equivalent. For simplicity, in several results we further assume that Φ is an *involution*, $\Phi(\Phi(x)) = x$ for all x, so that orbits have size at most two. More generally, Φ is only a convenient device for specifying a partition of $\mathcal{X}$ into classes of task-equivalent prompts (e.g., obtained by asking the model to judge equivalence): $\mathcal{C}_{\mathrm{coh}}$ depends only on this partition, so all results that use only the convexity of $\mathcal{C}_{\mathrm{coh}}$ apply verbatim to an arbitrary partition, and those stated for an involution extend to classes of any finite size by replacing pairwise quantities with their class-level counterparts (Remark C.2).

Coherent models. We fix an invariance mapping Φ. A model π is *coherent* if $\pi(x) = \pi(\Phi(x))$ for all $x \in \mathcal{X}$. Let $\mathcal{C}^{\dagger\dagger}_{\mathrm{coh}} = \{\pi : \mathcal{X} \to \mathbb{R}^d : \pi(x) = \pi(\Phi(x)) \text{ for all } x \in \mathcal{X}\}$ be the set of all coherent functions, i.e., functions constant on the orbits of Φ. Each constraint being linear, $\mathcal{C}^{\dagger\dagger}_{\mathrm{coh}}$ is a linear subspace of $(\mathbb{R}^d)^{\mathcal{X}}$. The set of coherent unnormalized models is $\mathcal{C}^{\dagger}_{\mathrm{coh}} = \mathcal{C}^{\dagger\dagger}_{\mathrm{coh}} \cap \Pi^{\dagger}_{\mathrm{all}}$, a closed convex cone, and the family of coherent conditional probability models is

$$\mathcal{C}_{\mathrm{coh}} = \mathcal{C}^{\dagger}_{\mathrm{coh}} \cap \Pi_{\mathrm{all}},$$

a compact convex set. More generally, the arguments below apply to any closed convex set $\mathcal{C}$ of models (e.g., defined by order, symmetry or logical-consistency constraints), and we then write $\overline{\Pi} = \mathcal{C} \cap \Pi$, assumed non-empty.

Problem formulation. Let $\pi_0 \in \Pi_{\mathrm{all}}$ be a base model. For a fixed Bregman divergence B_F, we quantify the improvement of a model π over π_0 with respect to a ground truth π^* by

$$\mathrm{Improv}_{\pi^*}(\pi) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi(x))]. \tag{1}$$

Our goal is to use the invariance mapping Φ to construct, from π_0 and without labels, a model $\widehat{\pi} \in \Pi$ such that $\mathrm{Improv}_{\pi^*}(\widehat{\pi}) \geq 0$ (ideally, strictly) for *all* $\pi^* \in \mathcal{C}_{\mathrm{coh}} \cap \Pi$. When F is the negative entropy, so that B_F is the relative entropy, $\mathrm{Improv}_{\pi^*}(\pi)$ is exactly the reduction in expected log-loss, since $\mathbb{E}_x[\mathsf{D}_{\mathrm{KL}}(\pi^*(x) \,\|\, \pi(x))] = \mathbb{E}_{(x,y)}[-\log \pi(y \mid x)] - \mathbb{E}_{(x,y)}[-\log \pi^*(y \mid x)]$ for $y \sim \pi^*(x)$.

4 Direct Coherence Projection

We study Bregman projections as a mechanism for enforcing coherence. We establish improvement guarantees for the projection onto $\mathcal{C}_{\mathrm{coh}}$ and, more generally, onto arbitrary closed convex sets (Section 4.1), show that this projection is maximin-optimal (Section 4.2), analyze *relaxed* coherence constraints, which yield the penalized objective used in practice (Section 4.3), and summarize guarantees for non-realizable targets, finite samples and approximate projections (Section 4.4).

4.1 Improvement guarantees

The following result shows that the Bregman projection onto a closed convex set $\mathcal{C}$, in particular the coherence set $\mathcal{C}_{\mathrm{coh}}$, improves the loss of the baseline whenever $\pi^* \in \mathcal{C} \cap \Pi$.

Theorem 4.1 (Bregman projection onto a convex set improves the baseline)**.** *Let $\mathcal{C}$ be a closed convex set such that $\mathcal{C} \cap \Pi$ is non-empty and contains the ideal model $\pi^* : \mathcal{X} \to \Delta(\mathcal{Y})$. Then, any solution $\widehat{\pi}$ of the convex optimization problem*

$$\min_{\pi \in \mathcal{C} \cap \Pi} \; \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi(x) \,\|\, \pi_0(x))] \tag{2}$$

satisfies

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] &\leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))] \\ &\leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))].\end{aligned}$$

If, in addition, Φ is an involution, $\mathcal{C} \subseteq \mathcal{C}_{\mathrm{coh}}$ (e.g., $\mathcal{C} = \mathcal{C}_{\mathrm{coh}}$), F is μ-strongly convex with respect to a norm $\|\cdot\|$, and $\lambda(x) = \frac{\mathbb{P}[x]}{\mathbb{P}[x]+\mathbb{P}[\Phi(x)]}$, then

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] \leq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))] \\ - \frac{\mu}{2} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}\Big[\lambda(x)(1-\lambda(x))\|\pi_0(x) - \pi_0(\Phi(x))\|^2\Big].\end{aligned}$$

The first statement follows from the Pythagorean theorem for Bregman divergences applied to the Legendre function F; the second quantifies the guaranteed gain in terms of the *measurable* incoherence of π_0 (proof in Section C.1). For the negative entropy, which is 1-strongly convex with respect to $\|\cdot\|_1$ (Pinsker's inequality [Rioul, 2023]), the theorem guarantees that the projection admits a log-loss no larger than that of π_0, with a gain of at least $\frac{1}{2}\,\mathbb{E}[\lambda(1-\lambda)\|\pi_0(x) - \pi_0(\Phi(x))\|_1^2]$. When $\mathcal{C} \cap \Pi$ is affine, the first inequality is an equality (Theorem E.1); since Π is bounded, this occurs only for a single point, but affine *directions* play a key role in Section 6.

4.2 Maximin optimality

Theorem 4.1 shows that

$$\mathrm{Improv}_{\pi^*}(\widehat{\pi}) \geq \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))] \quad \text{for all } \pi^* \in \mathcal{C}_{\mathrm{coh}} \cap \Pi. \tag{3}$$

Is there a coherent model with a better worst-case guarantee? The next result shows that the projection $\widehat{\pi}$ solves the max–min problem $\max_{\pi\in\mathcal{C}_{\mathrm{coh}}\cap\Pi} \min_{\pi^*\in\mathcal{C}_{\mathrm{coh}}\cap\Pi} \mathrm{Improv}_{\pi^*}(\pi)$ and that the right-hand side of (3) is the best possible worst-case improvement.

Theorem 4.2 (Maximin optimality)**.** *Let Π be a closed convex family of models, F a Legendre function, and $\widehat{\pi}$ the Bregman projection of π_0 onto $\mathcal{C}_{\mathrm{coh}} \cap \Pi$ (a solution of (2) with $\mathcal{C} = \mathcal{C}_{\mathrm{coh}}$). Then,*

$$\max_{\pi\in\mathcal{C}_{\mathrm{coh}}\cap\Pi} \min_{\pi^*\in\mathcal{C}_{\mathrm{coh}}\cap\Pi} \mathrm{Improv}_{\pi^*}(\pi) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))],$$

the maximum is attained at $\pi = \widehat{\pi}$, and the worst-case reference for $\widehat{\pi}$ is $\pi^* = \widehat{\pi}$.

The proof (Section C.2) shows that the improvement of any coherent π is smallest at the reference $\pi^* = \widehat{\pi}$.

4.3 Improvement guarantees for relaxed constraints

In practice, exact coherence $\pi(x) = \pi(\Phi(x))$ is rarely enforced; instead one penalizes incoherence. We now analyze this *relaxed* setting and show that the guarantees persist. Let D be a jointly convex divergence between distributions, such as the relative entropy D_{KL}, the squared Euclidean distance, the Jensen–Shannon divergence D_{JS}, the squared Hellinger distance $\mathsf{D}^2_{\mathrm{Hell}}$, or the symmetrized D_{KL}. For $\Lambda \geq 0$, define the convex set

$$\mathcal{C}^{\mathsf{D}}_\Lambda = \left\{\pi \in \Pi_{\mathrm{all}} \colon \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{D}(\pi(x), \pi(\Phi(x)))] \leq \Lambda\right\} \supseteq \mathcal{C}_{\mathrm{coh}},$$

and consider the projection of π_0 onto it,

$$\min_{\pi\in\Pi_{\mathrm{all}}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi(x) \,\|\, \pi_0(x))] \quad \text{subject to} \quad \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{D}(\pi(x), \pi(\Phi(x)))] \leq \Lambda. \tag{4}$$

This is a convex problem. By Lagrangian duality (Slater's condition holds for $\Lambda > 0$, the uniform model being coherent and interior; see Remark C.1 for both directions), it is equivalent to the penalized problem

$$\min_{\pi\in\Pi_{\mathrm{all}}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi(x) \,\|\, \pi_0(x))] + \lambda \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{D}(\pi(x), \pi(\Phi(x)))], \tag{5}$$

for some $\lambda \geq 0$. With $\mathsf{B}_F = \mathsf{D}_{\mathrm{KL}}$ and $\mathsf{D} = \mathsf{D}_{\mathrm{JS}}$, (5) is exactly the objective we use for fine-tuning in Section 7 (KL to the frozen baseline plus a pairwise coherence penalty). Two observations follow from Theorem 4.1, since $\mathcal{C}^{\mathsf{D}}_\Lambda$ is closed and convex: (i) the solution $\widehat{\pi}$ of (4) improves over π_0 for *every* $\pi^* \in \mathcal{C}^{\mathsf{D}}_\Lambda$, hence also for Λ-approximately coherent ones, a first form of robustness to imperfect transformations; (ii) the guaranteed gain is $\mathbb{E}[\mathsf{B}_F(\widehat{\pi} \,\|\, \pi_0)]$, which the next theorem bounds from below explicitly.

Theorem 4.3 (Improvement guarantees for relaxed constraints)**.** *Assume that Φ is an involution, that F is μ_F-strongly convex with respect to a norm $\|\cdot\|$ on $\mathbb{R}^d$, and that $\mathsf{D}(\mathsf{p},\mathsf{q}) \geq \frac{\mu_\mathsf{D}}{2}\|\mathsf{p}-\mathsf{q}\|^2$ for all $\mathsf{p},\mathsf{q} \in \Delta(\mathcal{Y})$. Let $\|\pi\|_{L_2} = [\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}\|\pi(x)\|^2]^{1/2}$ and define the incoherence gap of π_0 as $\Delta_{\mathrm{coh}} = \inf_{\pi\in\mathcal{C}_{\mathrm{coh}}}\|\pi_0 - \pi\|^2_{L_2}$. Let $\widehat{\pi}$ be the solution of (4). Then, for every $\pi^* \in \mathcal{C}^{\mathsf{D}}_\Lambda$,*

$$\mathrm{Improv}_{\pi^*}(\widehat{\pi}) \geq \frac{\mu_F}{2}\left[\sqrt{\Delta_{\mathrm{coh}}} - \sqrt{\frac{\Lambda}{2\mu_\mathsf{D}}}\right]_+^2.$$

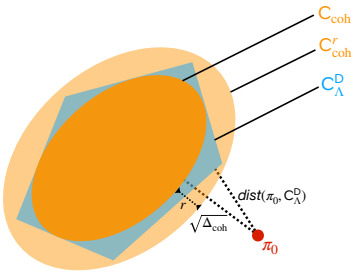


Figure 1: Illustration of the proof of Theorem 4.3: the relaxed set $\mathcal{C}^{\mathsf{D}}_{\Lambda}$ is contained in an L_2-tube of radius $r = \sqrt{\Lambda/(2\mu_{\mathsf{D}})}$ around the coherent set $\mathcal{C}_{\text{coh}}$, so the distance from π_0 to $\mathcal{C}^{\mathsf{D}}_{\Lambda}$ is at least $\sqrt{\Delta_{\text{coh}}} - r$; strong convexity of F then converts this distance into a Bregman improvement.

Furthermore, with $\gamma_0 = \|\pi_0 - \pi_0 \circ \Phi\|^2_{L_2}$ *the (computable) incoherence of* π_0 *and* $C_\Phi = \sup_{\pi \neq 0} \|\pi \circ \Phi\|_{L_2} / \|\pi\|_{L_2} = \max_x [\mathcal{D}_{\mathcal{X}}(\Phi(x))/\mathcal{D}_{\mathcal{X}}(x)]^{1/2}$ *the operator norm of the composition with* Φ*, we have* $\frac{\gamma_0}{(1+C_\Phi)^2} \leq \Delta_{\text{coh}} \leq \frac{\gamma_0}{4}$*; in particular* $\Delta_{\text{coh}} = \gamma_0/4$ *when* $\mathcal{D}_{\mathcal{X}}$ *is* Φ*-invariant.*

The proof (Section C.3, illustrated in Figure 1) builds a geometric bridge: the strong convexity of F converts the Bregman gain into an L_2 distance, the lower bound on D encloses $\mathcal{C}^{\mathsf{D}}_{\Lambda}$ in a tube around $\mathcal{C}_{\text{coh}}$, and a reverse triangle inequality yields the bound. The theorem covers both Problems (4) and (5), and shows that the improvement remains *strictly* positive as long as the tolerance Λ is small relative to the incoherence of π_0, equivalently, for sufficiently large penalty λ. For instance, with $\|\cdot\| = \|\cdot\|_1$, Pinsker-type inequalities give $\mathsf{D}_{\text{JS}}(\mathsf{p}, \mathsf{q}) \geq \frac{1}{8}\|\mathsf{p} - \mathsf{q}\|_1^2$ and $\mathsf{D}^2_{\text{Hell}}(\mathsf{p}, \mathsf{q}) \geq \frac{1}{8}\|\mathsf{p} - \mathsf{q}\|_1^2$, so $\mu_{\mathsf{D}_{\text{JS}}} = \mu_{\mathsf{D}^2_{\text{Hell}}} = 1/4$ and, for a Φ-invariant $\mathcal{D}_{\mathcal{X}}$,

$$\text{Improv}_{\pi^*}(\widehat{\pi}) \geq \frac{\mu_F}{2}\Big[\tfrac{1}{2}\sqrt{\gamma_0} - \sqrt{2\Lambda}\Big]_+^2, \qquad \gamma_0 = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}\big[\|\pi_0(x) - \pi_0(\Phi(x))\|_1^2\big],$$

a bound computable from the baseline; for the symmetrized D_{KL}, $\mathsf{D}_{\text{KL}}(\mathsf{p} \,\|\, \mathsf{q}) + \mathsf{D}_{\text{KL}}(\mathsf{q} \,\|\, \mathsf{p})$, $\mu_{\mathsf{D}} = 2$ and $\sqrt{2\Lambda}$ becomes $\sqrt{\Lambda}/2$.

4.4 Robustness: non-realizable targets, finite samples and approximate projections

The geometric structure of Bregman projections extends these guarantees to the imperfections encountered in practice; we state the results informally (precise statements and proofs in Section D).

Non-realizable ground truth (Theorem D.1). If the transformations are not perfectly task-preserving, π^* may lie outside $\mathcal{C}_{\text{coh}}$. Let $\epsilon = \mathbb{E}[\mathsf{B}_F(\pi^* \,\|\, \overline{\pi})]$ where $\overline{\pi}$ is the point of $\mathcal{C}_{\text{coh}} \cap \Pi$ closest to π^* in the sense of $\mathbb{E}[\mathsf{B}_F(\pi^* \,\|\, \cdot)]$, and $D = \mathbb{E}[\mathsf{B}_F(\widehat{\pi} \,\|\, \pi_0)]$ the displacement of the projection. If F is μ-strongly convex and L-smooth, then

$$\mathbb{E}[\mathsf{B}_F(\pi^* \,\|\, \widehat{\pi})] - \mathbb{E}[\mathsf{B}_F(\pi^* \,\|\, \pi_0)] \leq -D + \frac{2L}{\mu}\sqrt{\epsilon D},$$

so the improvement degrades gracefully with ϵ and remains strictly positive whenever $\epsilon < (\mu/2L)^2 D$: enforcing coherence helps as long as the ground truth is much closer to $\mathcal{C}_{\text{coh}}$ than the baseline.

Finite samples (Theorem D.2). When the projection is computed on an i.i.d. sample S of m prompts, the empirical projection $\widehat{\pi}_S$ satisfies $\mathbb{E}[\mathsf{B}_F(\pi^* \,\|\, \widehat{\pi}_S)] \leq \mathbb{E}[\mathsf{B}_F(\pi^* \,\|\, \widehat{\pi})] + 4\epsilon_m + \mathbb{E}[\mathsf{B}_F(\pi^* \,\|\, \pi_0) - \mathsf{B}_F(\widehat{\pi} \,\|\, \pi_0)]$, where ϵ_m is a uniform-deviation term controlled by standard Rademacher complexity bounds. The last *misalignment* term is unavoidable in general, because $\widehat{\pi}_S$ minimizes proximity to π_0 and not to π^* (Section D.3).

Approximate projections, model mismatch and optimization error (Theorem D.5). Parametric fine-tuning with non-convex optimization will not return the exact projection. Let J_λ be the penalized objective (5), with minimizer $\widehat{\pi}_\lambda$, and let $\widetilde{\pi} \in \Pi_{\text{all}}$ be *any* model, e.g., the output of a fine-tuning run in a parametric class, with suboptimality $\eta = J_\lambda(\widetilde{\pi}) - J_\lambda(\widehat{\pi}_\lambda)$. If F is μ_F-strongly convex and $\pi \mapsto \mathbb{E}[\mathsf{B}_F(\pi^* \,\|\, \pi)]$ is L_{π^*}-Lipschitz in L_2 (Euclidean norms), then

$$\text{Improv}_{\pi^*}(\widetilde{\pi}) \geq \text{Improv}_{\pi^*}(\widehat{\pi}_\lambda) - L_{\pi^*}\sqrt{2\eta/\mu_F}.$$

The bound is agnostic to the source of suboptimality: η jointly captures model-class mismatch (the projection may not lie in the parametric class) and optimization error. For the D_{KL}–D_{JS} objective of Section 7, with F the negative entropy and model outputs bounded below by α on the support of $\pi^*(x)$, $\mu_F = 1$ and $L_{\pi^*} \le 1/\alpha$, a dimension-free constant (Remark D.7), so that $\mathrm{Improv}_{\pi^*}(\widehat{\pi}) \ge \mathrm{Improv}_{\pi^*}(\widehat{\pi}_\lambda) - \sqrt{2\eta}/\alpha$: the gap to the relaxed projection is controlled by the training suboptimality alone, which can be monitored during fine-tuning. The same argument gives $\mathrm{Improv}(\widehat{\pi}_S) \ge \mathrm{Improv}(\widehat{\pi}) - \frac{2L_{\pi^*}}{\sqrt{\mu}}\sqrt{\epsilon_m}$ for the empirical projection (Corollary D.6).

Two-step projection (Section E). A natural alternative first projects π_0 onto the cone of coherent unnormalized functions $\mathcal{C}_{\mathrm{coh}}^{\dagger}$ (a closed-form Bregman centroid), and then onto $\mathcal{C}_{\mathrm{coh}} \cap \Pi$. For all quadratic and separable generators (squared Euclidean, Mahalanobis, D_{KL}, Itakura–Saito, α- and β-divergences), the two-step projection *coincides* with the direct one; this yields, for the relative entropy, a log-loss improvement of at least $\mathbb{E}[2\min\{\lambda(x), 1-\lambda(x)\}\mathsf{D}^2_{\mathrm{Hell}}(\pi_0(x) \,\|\, \pi_0(\Phi(x)))]$ (Corollary E.6).

5 Improvement under Multiple Divergences

The previous section fixed a single divergence B_F. We now ask whether a single model can improve over π_0 simultaneously for a *family* of Bregman divergences generated by a convex set $\mathcal{F}$ of Legendre generators. We first show that a natural robust (minimax) solution does not guarantee improvement (Section 5.1), then identify exactly when universal improvement is possible (Section 5.2).

5.1 Failure of Pythagorean improvement for minimax projections

A natural robust solution is the minimax projection $\min_{\pi\in\Pi\cap\mathcal{C}_{\mathrm{coh}}} \max_{F\in\mathcal{F}} \mathbb{E}_x[\mathsf{B}_F(\pi(x) \,\|\, \pi_0(x))]$, a convex problem since each divergence is convex in its first argument. As the next result shows, it does not in general guarantee improvement, even for the convex hull of two quadratic generators.

Theorem 5.1 (Failure of minimax projections)**.** *There exist finite sets $\mathcal{X}, \mathcal{Y}$ with $|\mathcal{Y}| = 4$, a prompt distribution $\mathcal{D}_{\mathcal{X}}$, a closed convex model class $\Pi \subseteq \Pi_{\mathrm{all}}$, an involution Φ on $\mathcal{X}$ with coherence set $\mathcal{C}_{\mathrm{coh}}$, a source model $\pi_0 \notin \Pi$, and a convex family $\mathcal{F} = \mathrm{conv}\{F_1, F_2\}$ of two quadratic Legendre generators such that the unique minimax solution*

$$\pi_{mm} = \operatorname*{argmin}_{\pi\in\Pi\cap\mathcal{C}_{\mathrm{coh}}} \max_{F\in\mathcal{F}} \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\pi(x) \,\|\, \pi_0(x))]$$

violates the improvement condition: there exist $F \in \mathcal{F}$ and a feasible reference $\pi^* \in \Pi \cap \mathcal{C}_{\mathrm{coh}}$ with $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_{mm}(x))] > \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))]$.

The explicit construction is in Section F.1: the two generators weigh different coordinates heavily, the minimax solution sits halfway between the two single-generator projections, and a reference placed at one of them is farther from π_{mm} than from π_0 in the other geometry. The example also shows that for some pairs of Bregman divergences *no* coherent model in Π improves for both.

5.2 Universal improvement guarantees and impossibility results

Can a single model $\widehat{\pi}$ improve over π_0 for *every* divergence in $\mathcal{F}$? The following theorem shows that this is attainable under two conditions: convexity of the divergences in their second argument (implied by joint convexity), and feasibility of a specific *orbit-averaged* model.

Theorem 5.2 (Improvement via orbitwise averaging for jointly convex divergences)**.** *Let Π be a closed convex family of models, Φ an involution, $\pi_0 \in \Pi$ a baseline and $\pi^* \in \Pi \cap \mathcal{C}_{\mathrm{coh}}$ a coherent reference. Let $\mathcal{F}$ be a family of Legendre generators such that, for every $F \in \mathcal{F}$ and every $\mathsf{p} \in \Delta(\mathcal{Y})$, the map $\mathsf{q} \mapsto \mathsf{B}_F(\mathsf{p} \,\|\, \mathsf{q})$ is convex on $\Delta(\mathcal{Y})$; this holds in particular when B_F is jointly convex, as for the squared Euclidean, Mahalanobis and relative entropy divergences. Define the orbitwise averaged (coherent) model*

$$\widehat{\pi}_{\mathrm{orbit}}(x) = \frac{\mathbb{P}[x]\pi_0(x) + \mathbb{P}[\Phi(x)]\pi_0(\Phi(x))}{\mathbb{P}[x] + \mathbb{P}[\Phi(x)]} = \lambda(x)\pi_0(x) + (1-\lambda(x))\pi_0(\Phi(x)).$$

Then, for all $F \in \mathcal{F}$, $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_{\mathrm{orbit}}(x))] \le \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))]$. Moreover, $\widehat{\pi}_{\mathrm{orbit}} \in \Pi$ whenever Π is closed under the swap $\pi \mapsto \pi \circ \Phi$ and $\mathcal{D}_{\mathcal{X}}$ is Φ-invariant (so that $\lambda \equiv 1/2$).

The proof decomposes $\mathcal{X}$ into orbits $\{x, \Phi(x)\}$ and applies convexity in the second argument (Section F.2); Corollary F.1 specializes it to the Mahalanobis family. Note that $\widehat{\pi}_{\mathrm{orbit}}$ is the (weighted) Euclidean projection of π_0 onto $\mathcal{C}_{\mathrm{coh}}$, and that, unlike a Bregman projection, it does not depend on the divergence. The next two results show that both conditions are essentially necessary. We say that a coherent model $\widehat{\pi}$ has the *universal improvement property* over $\mathcal{F}$ if

$$\forall F \in \mathcal{F},\ \forall \pi^* \in \mathcal{C}_{\mathrm{coh}}, \qquad \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] \le \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))]. \tag{6}$$

Theorem 5.3 (Universal improvement over quadratic divergences forces orbit averaging)**.** *Let Φ be an involution and let $\mathcal{F}$ contain the quadratic generators $F_{\mathsf{M}}(\mathsf{p}) = \frac{1}{2}\mathsf{p}^\top \mathsf{M}\mathsf{p}$ for all positive definite M. Let $\widehat{\pi} \in \mathcal{C}_{\mathrm{coh}}$ satisfy (6), and let $u(x) = \widehat{\pi}(x) - \widehat{\pi}_{\mathrm{orbit}}(x)$ and $\delta_0(x) = \pi_0(x) - \pi_0(\Phi(x))$. Then:*

(a) $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[u(x)u(x)^\top] \preceq \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\lambda(x)(1-\lambda(x))\delta_0(x)\delta_0(x)^\top]$*: the deviation from $\widehat{\pi}_{\mathrm{orbit}}$ is dominated by the incoherence of π_0.*

(b) *If $|\mathcal{Y}| \ge 3$ and $\mathcal{X}$ consists of a single orbit $\{x, \Phi(x)\}$, then $\widehat{\pi} = \widehat{\pi}_{\mathrm{orbit}}$. In particular, if $\widehat{\pi}_{\mathrm{orbit}} \notin \Pi$, no model in $\Pi \cap \mathcal{C}_{\mathrm{coh}}$ has the universal improvement property.*

The proof (Section F.4) cancels the quadratic terms in π^* and uses rank-one generators $\mathsf{M} = vv^\top$ with v nearly orthogonal to δ_0 to force $u = 0$. Part (b) cannot be extended verbatim to several orbits: incoherence on one orbit creates slack that can absorb a small deviation on another, which (a) quantifies (Remark F.2). The second condition, convexity in the second argument, is also necessary.

Theorem 5.4 (Impossibility without convexity in the second argument)**.** *Let $|\mathcal{Y}| \ge 3$ and let $\mathcal{F}$ be a convex family of Legendre generators containing all quadratic generators F_{M}, $\mathsf{M} \succ 0$, and at least one generator F whose Bregman divergence is not convex in its second argument on $\Delta(\mathcal{Y})$, i.e., for which $\mathsf{q} \mapsto \mathsf{B}_F(\mathsf{p}^* \,\|\, \mathsf{q})$ is not convex on $\operatorname{relint}\Delta(\mathcal{Y})$ for some $\mathsf{p}^* \in \Delta(\mathcal{Y})$. Then, there exists an instance, with $\mathcal{X}$ a single orbit $\{x, \Phi(x)\}$, a distribution $\mathcal{D}_{\mathcal{X}}$ and a baseline π_0, on which* no *coherent model $\widehat{\pi} \in \mathcal{C}_{\mathrm{coh}}$ has the universal improvement property (6) over $\mathcal{F}$.*

The proof (Section F.5) places the witnesses of the failure of convexity on an orbit: by Theorem 5.3, any universal improver must be the orbit average, which by construction is strictly worse than π_0 for F. Together, these results give an exact dichotomy for families containing the quadratic generators: universal improvement is possible (on every instance) if and only if every divergence in the family is convex in its second argument, and it is then achieved by $\widehat{\pi}_{\mathrm{orbit}}$ whenever it is feasible.

6 Rigidity of Improvement Mechanisms

We have shown that Bregman projections onto coherent sets are provable improvement mechanisms. We now turn to the converse question: what can be said about *any* mechanism with such a guarantee? In geometry and topology, rigidity theorems show that certain structures resist perturbation [Mostow, 1973, Gromov and Pansu, 1991, Spatzier, 2004]; we find an analogous phenomenon here. Fix a closed convex set $\Pi \subseteq \Pi_{\mathrm{all}}$ (for instance $\mathcal{C}_{\mathrm{coh}} \cap \Pi'$ for a model class Π'), a baseline π_0, and $\mathcal{D}_{\mathcal{X}}$. A *mechanism* assigns to every Legendre generator F in a family $\mathcal{F}$ an output $\widehat{\pi}_F \in \Pi$. We say that it satisfies the *strong improvement inequality* (SII) if, for all $\pi^* \in \Pi$,

$$\mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_F(x))] \le \mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))] - \mathbb{E}[\mathsf{B}_F(\widehat{\pi}_F(x) \,\|\, \pi_0(x))], \tag{7}$$

which is precisely the guarantee of Theorem 4.1. All results below are proven in Section G.

Theorem 6.1 (Characterization of improvement mechanisms)**.** *Let F be Legendre and assume $\pi_0(x)$ and $\widehat{\pi}_F(x)$ lie in $\operatorname{int}(\operatorname{dom} F)$ for all x. The following are equivalent: (i) $\widehat{\pi}_F$ satisfies the SII (7) for all $\pi^* \in \Pi$; (ii) $\widehat{\pi}_F$ satisfies the variational inequality $\mathbb{E}[\langle \nabla F(\widehat{\pi}_F(x)) - \nabla F(\pi_0(x)), \pi(x) - \widehat{\pi}_F(x)\rangle] \ge 0$ for all $\pi \in \Pi$; (iii) $\widehat{\pi}_F$ is the (unique) B_F-projection of π_0 onto Π.*

Thus, provable improvement in the Pythagorean sense *is* Bregman projection onto a convex set: there is no other mechanism. The proof is short: by the three-point identity, the SII is equivalent to the variational inequality (ii), the first-order optimality condition of the projection problem, which is sufficient by convexity and singles out a unique point by strict convexity. The output varies with F, but its structure is severely constrained: first, every symmetry of the problem is inherited by the output, for every generator.

Theorem 6.2 (Symmetry rigidity)**.** *Let $\sigma: \mathcal{X} \to \mathcal{X}$ be a bijection such that Π is σ-invariant ($\pi \in \Pi \Rightarrow \pi \circ \sigma \in \Pi$), $\pi_0 \circ \sigma = \pi_0$, and $\mathcal{D}_{\mathcal{X}} \circ \sigma = \mathcal{D}_{\mathcal{X}}$. Then, for every Legendre generator F for which $\widehat{\pi}_F$ satisfies the SII, $\widehat{\pi}_F(\sigma(x)) = \widehat{\pi}_F(x)$ for all $x \in \mathcal{X}$.*

In other words, any tie forced by a symmetry of $(\Pi, \pi_0, \mathcal{D}_{\mathcal{X}})$ holds for *all* generators, with no continuity or strictness assumption. Section G develops this further in two directions.

(1) Affine structure. Suppose that $\Pi = A \cap \Pi_{\mathrm{all}}$ is *affinely constrained*: A is an affine subspace with direction space $\Pi_0 \subseteq Z^{\mathcal{X}}$, where $Z = \{v \in \mathbb{R}^d : \sum_k v_k = 0\}$ (e.g., calibration, moment, fairness or coherence constraints). Let P_Z be the orthogonal projection onto Z, $\Pi_0^{\perp}$ the $\mathcal{D}_{\mathcal{X}}$-weighted orthogonal complement of Π_0 within $Z^{\mathcal{X}}$, and $r = \dim \Pi_0^{\perp} = n(d-1) - \dim \Pi_0$. We say that (RI) holds at F if $\widehat{\pi}_F \in \operatorname{relint} \Pi$, which is automatic for steep generators such as the negative entropy whenever Π meets $\operatorname{relint} \Pi_{\mathrm{all}}$ (Remark G.1).

Theorem 6.3 (Finite-dimensional rigidity for affine constraints)**.** *Let $\Pi = A \cap \Pi_{\mathrm{all}}$ be affinely constrained and let $\mathcal{F}_{\mathrm{RI}}$ be the set of Legendre generators F for which $\widehat{\pi}_F$ satisfies the SII and (RI). Then: (a) for every $F \in \mathcal{F}_{\mathrm{RI}}$, there exist $w_F \in \Pi_0^{\perp}$ and a scalar function $c_F : \mathcal{X} \to \mathbb{R}$ such that*

$$\nabla F(\widehat{\pi}_F(x)) = \nabla F(\pi_0(x)) + w_F(x) + c_F(x)\mathbf{1}, \qquad x \in \mathcal{X}; \tag{8}$$

(b) for a fixed F, the dual residual w_F determines $\widehat{\pi}_F$: if $\pi, \pi' \in \Pi_{\mathrm{all}}$ with values in $\operatorname{int}(\operatorname{dom} F)$ satisfy $P_Z(\nabla F(\pi(x)) - \nabla F(\pi_0(x))) = P_Z(\nabla F(\pi'(x)) - \nabla F(\pi_0(x))) = w(x)$ for all x, then $\pi = \pi'$; (c) consequently, the family $\{\widehat{\pi}_F : F \in \mathcal{F}_{\mathrm{RI}}\}$ is parametrized by the r-dimensional space $\Pi_0^{\perp}$, however large the family of generators; in particular, if $r = 0$ (i.e., $\Pi = \Pi_{\mathrm{all}}$), then $\widehat{\pi}_F = \pi_0$ for all F.

Thus, the generator only selects a point of an r-dimensional space, r being the number of independent constraints, and the nonlinear map $(\nabla F)^{-1}$ then fixes the scale of the correction. Moreover, a purely structural equivalence on $\mathcal{X}$, computable from $(\Pi, \mathcal{D}_{\mathcal{X}})$ alone, determines which pairs of inputs are tied by every generator.

Theorem 6.4 (Universal tie criterion for affine constraints)**.** *Let $\Pi = A \cap \Pi_{\mathrm{all}}$ be affinely constrained and define the* structural equivalence *on $\mathcal{X}$ by $x \approx x'$ iff $w(x) = w(x')$ for every $w \in \Pi_0^{\perp}$. Then, for every baseline $\pi_0 \in \Pi_{\mathrm{all}}$, every Legendre generator F such that $\widehat{\pi}_F$ satisfies the SII and (RI), and every pair $x, x' \in \mathcal{X}$: if $x \approx x'$ and $\pi_0(x) = \pi_0(x')$, then $\widehat{\pi}_F(x) = \widehat{\pi}_F(x')$. In particular, $\widehat{\pi}_F$ belongs to the coherence set $\mathcal{C}_{\approx} = \{\pi \in \Pi : \pi(x) = \pi(x') \text{ for all } x \approx x' \text{ with } \pi_0(x) = \pi_0(x')\}$ for every such F and every π_0.*

When Π is given by r independent linear constraints $\sum_z \langle a_k(z), \pi(z) \rangle = c_k$ with $a_k(z) \in Z$, the criterion is verifiable: $x \approx x'$ iff $a_k(x)/\mathcal{D}_{\mathcal{X}}(x) = a_k(x')/\mathcal{D}_{\mathcal{X}}(x')$ for all k, i.e., the $\mathcal{D}_{\mathcal{X}}$-normalized constraint coefficients coincide; this detects *weighted symmetries* invisible to Theorem 6.2. Under a mild richness condition this description is exact (Theorem G.4), and the Pythagorean inequality becomes an identity quantifying the gap between the outputs of two generators (Theorem G.5). *(2) Local rigidity.* When Π is a polytope (e.g., coherence, order or calibration constraints), the face of Π exposed by the mechanism at F_0 absorbs the outputs of all nearby generators, so that every tie holding on that face is inherited *exactly*, not approximately, by $\widehat{\pi}_F$ for F near F_0 (Theorem G.8); along connected families of generators, all outputs lie in the relative interior of a single face under a standard non-degeneracy condition (Corollary G.11); for general convex Π, a conic sharp-minimum condition yields a quantitative version in which the output approaches the coherence set to second order (Proposition G.10); "near" refers to uniform convergence of $(F, \nabla F)$ on $\Delta(\mathcal{Y})$, so these local results concern generators finite and C^1 near the simplex, whereas the characterization, symmetry and affine results also cover steep generators such as the negative entropy. These results show that the behavior of any provable improvement mechanism is dictated by the geometry of the problem $(\Pi, \mathcal{D}_{\mathcal{X}})$ rather than by the choice of divergence.

7 Empirical Evaluations

We evaluate coherence training on the TriviaQA benchmark (17,944 validation examples) with three instruction-tuned language models: `Phi-3-mini-4k-instruct`, `Qwen2.5-3B-Instruct`, and `Llama-3.2-3B-Instruct`. Starting from each baseline, we minimize the penalized objective (5) with $\mathsf{B}_F = \mathsf{D}_{\mathrm{KL}}$ (to the frozen baseline) and $\mathsf{D} = \mathsf{D}_{\mathrm{JS}}$ (between the answer distributions on a question and its paraphrase), at the completion level, with QLoRA (4-bit NF4, rank $r = 8$) for 200 batches of 32 paraphrase classes and coherence weight $\lambda = 1$, without any hyperparameter search.

		Greedy		Sampling	
Model	**Training**	**EM**	**F1 (95% CI)**	**EM**	**F1 (95% CI)**
Phi-3	baseline	32.03	42.76 ± 0.63	29.49	40.12 ± 0.62
Phi-3	trained	**35.69**	**45.35 ± 0.65**	**31.24**	**41.07 ± 0.64**
Qwen2.5	baseline	32.85	39.65 ± 0.67	31.94	38.65 ± 0.67
Qwen2.5	trained	**33.89**	**40.40 ± 0.68**	**32.75**	**39.23 ± 0.67**
Llama-3.2	baseline	39.12	48.56 ± 0.66	37.02	46.07 ± 0.66
Llama-3.2	trained	**43.53**	**51.10 ± 0.68**	**40.18**	**47.86 ± 0.68**

Table 1: TriviaQA validation performance (17,944 examples) for baseline and coherence-trained models, under greedy decoding and sampling (temperature 0.7, top-p = 0.9). F1 is reported as mean ± half-width of the 95% confidence interval.

Paraphrases are generated by `Qwen2.5-3B-Instruct` with greedy decoding: the coherent set is constructed from the lower-level paraphrasing ability of a model of the same scale, the model's own for `Qwen2.5` and a peer's for `Phi-3` and `Llama-3.2`, which keeps Φ fixed across architectures. Full details are given in Section H.

Coherence training improves downstream QA performance across all three models under both greedy and sampling-based decoding (Table 1; absolute and relative gains in Table 2, Section H): exact match improves by 3.66 and 4.41 points for Phi-3 and Llama-3.2 under greedy decoding (relative gains of $\sim 11\%$), and gains under sampling are smaller but consistently positive. Semantic pairwise coherence (the rate at which answers to paraphrased questions express the same factual answer) also improves for all three models (Table 3, Section H): from 0.533 to 0.608 for Phi-3 (relative gain 14.07%), from 0.472 to 0.491 for Qwen2.5 and from 0.655 to 0.690 for Llama-3.2 (Wilson 95% half-widths 0.010). Gains are smallest for `Qwen2.5`, the model trained on its own paraphrases and with the lowest baseline coherence (see Section H). Regarding ablations, dropping the coherence term leaves π_0 unchanged, since π_0 minimizes the KL term alone, so the coherence penalty is the sole source of improvement; supervised baselines use labels that our mechanism does not; label-free alternatives (unanchored consistency minimization, majority-vote pseudo-labeling, inference-time orbit averaging) are discussed in Section H in the light of the theory.

8 Conclusion

We presented a mathematical framework for coherence-based self-improvement, in which a model is improved, without labels, by projecting it onto the convex set of models that are coherent under task-preserving transformations. The framework delivers guarantees of several kinds: monotonic improvement with respect to every coherent ground truth, with a quantified gain and maximin optimality; robustness to relaxed constraints, non-realizable targets, finite samples and inexact optimization; a sharp picture of when improvement can hold simultaneously for several divergences; and characterization and rigidity results showing that provable improvement mechanisms are Bregman projections whose structure is dictated by the problem geometry; experiments confirm that coherence training improves consistency and accuracy.

The framework has limitations, discussed further in Section I. The guarantees require a coherent set that contains, or approximately contains, the ground truth: if the transformations alter the task, enforcing coherence can introduce bias, a risk that Theorem D.1 quantifies but does not eliminate. The analysis is at the level of output distributions; the connection to parametric fine-tuning is made through the penalized objective and the approximate-projection bound of Section 4.4, which says nothing about the optimization dynamics themselves. Finally, our experiments are a proof of concept with small models on a single benchmark; a larger-scale study, including reasoning tasks, is carried out in follow-up work [Mohri et al., 2026]. Natural directions for future research include richer coherence sets (order, logical and calibration constraints), mechanisms that learn the invariance mapping jointly with the model, and extending the rigidity theory to general convex constraint sets.

References

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-policy distillation of language models: Learning from self-generated mistakes. In *The twelfth international conference on learning representations*, 2024.

Charalambos D Aliprantis and Kim C Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer Science & Business Media, 3rd edition, 2006.

Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. In *The Eleventh International Conference on Learning Representations*, 2023.

Yasemin Altun and Alex Smola. Unifying divergence minimization and statistical inference via convex duality. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT 2006)*, pages 139–153, 2006.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*, 2022.

A. Banerjee, I. Dhillon, J. Ghosh, and S. Merugu. A generalized maximum entropy approach to Bregman co-clustering and matrix approximation. *Journal of Machine Learning Research*, 8: 1919–1986, 2007.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, and Joydeep Ghosh. Clustering with Bregman divergences. *Journal of machine learning research*, 6(Oct):1705–1749, 2005.

Peter L. Bartlett and Shahar Mendelson. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Heinz H. Bauschke and Jonathan M. Borwein. Legendre functions and the method of random Bregman projections. *Journal of Convex Analysis*, 4(1):27–67, 1997.

Claude Berge. *Topological Spaces: Including a Treatment of Multi-Valued Functions, Vector Spaces and Convexity*. Oliver & Boyd, Edinburgh and London, 1963.

Quentin Bertrand, Joey Bose, Alexandre Duplessis, Marco Jiralerspong, and Gauthier Gidel. On the stability of iterative retraining of generative models on their own data. In *The Twelfth International Conference on Learning Representations*, 2024.

Lev M. Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Computational Mathematics and Mathematical Physics*, 7:200–217, 1967.

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, et al. Weak-to-strong generalization: eliciting strong capabilities with weak supervision. In *Proceedings of the 41st International Conference on Machine Learning*, pages 4971–5012, 2024.

Yair Censor and Arnold Lent. An iterative row-action method for interval convex programming. *Journal of Optimization Theory and Applications*, 34(3):321–353, 1981.

Moses Charikar, Chirag Pabbaraju, and Kirankumar Shiragur. Quantifying the gain in weak-to-strong generalization. *Advances in neural information processing systems*, 37:126474–126499, 2024.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR, 2020.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

Michael Collins, Robert E. Schapire, and Yoram Singer. Logistic regression, Adaboost and Bregman distances. *Machine Learning*, 48(1-3):253–285, 2002.

Corinna Cortes, Vitaly Kuznetsov, Mehryar Mohri, and Umar Syed. Structural Maxent models. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 391–399, Lille, France, 07–09 Jul 2015. PMLR.

Ekin D Cubuk, Barret Zoph, Dandelion Mane, Vijay Vasudevan, and Quoc V Le. Autoaugment: Learning augmentation strategies from data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 113–123, 2019.

Ekin D Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V Le. Randaugment: Practical automated data augmentation with a reduced search space. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 702–703, 2020.

Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Maximum entropy density estimation with generalized regularization and an application to species distribution modeling. *Journal of Machine Learning Research*, 8:1217–1260, 2007.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031, 2021.

Matthew Finlayson, John Hewitt, Alexander Koller, Swabha Swayamdipta, and Ashish Sabharwal. Closing the curious case of neural text degeneration. In *The Twelfth International Conference on Learning Representations*, 2024.

Béla A Frigyik, Santosh Srivastava, and Maya R Gupta. Functional Bregman divergence and Bayesian estimation of distributions. *IEEE Transactions on Information Theory*, 54(11):5130–5139, 2008.

Rafael Frongillo and Mark D. Reid. Convex foundations for generalized maxent models. In *Bayesian Inference and Maximum Entropy Methods in Science and Engineering (MaxEnt 2013), AIP Conference Proceedings*, volume 1636, pages 11–16. AIP Publishing, 2014.

Yichao Fu, Xuewei Wang, Yuandong Tian, and Jiawei Zhao. Deep think with confidence. *arXiv preprint arXiv:2508.15260*, 2025.

Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Tomasz Korbak, Henry Sleight, Rajashree Agrawal, John Hughes, Dhruv Bhandarkar Pai, Andrey Gromov, et al. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. In *First Conference on Language Modeling*, 2024.

Mikhael Gromov and Pierre Pansu. Rigidity of lattices: An introduction. In Paolo de Bartolomeis and Franco Tricerri, editors, *Geometric Topology: Recent Developments*, volume 1504 of *Lecture Notes in Mathematics*, pages 39–137. Springer, 1991.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge distillation of large language models. In *The Twelfth International Conference on Learning Representations*, 2024.

P. Harremoës. Maximum entropy and sufficiency. *arXiv preprint arXiv:1607.02259*, 2016.

Zellig S. Harris. *Mathematical structures of language*, volume 21 of *Interscience tracts in pure and applied mathematics*. Interscience Publ., 1968.

Ryuichiro Hataya, Han Bao, and Hiromi Arai. Will large-scale generative models corrupt future datasets? In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20555–20565, 2023.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Audrey Huang, Adam Block, Dylan J. Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T. Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. *arXiv preprint arXiv:2412.01951*, 2024. URL https://arxiv.org/abs/2412.01951.

Jiaxin Huang, Shixiang Gu, Le Hou, Yuexin Wu, Xuezhi Wang, Hongkun Yu, and Jiawei Han. Large language models can self-improve. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1051–1068, 2023.

Haozhe Ji, Pei Ke, Hongning Wang, and Minlie Huang. Language model decoding as direct metrics optimization. In *The Twelfth International Conference on Learning Representations*, 2024.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, 2017.

Sham M. Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13:1865–1890, 2012.

Konwoo Kim, Suhas Kotha, Percy Liang, and Tatsunori Hashimoto. Pre-training under infinite compute. *arXiv preprint arXiv:2509.14786*, 2025.

Jyrki Kivinen and Manfred K. Warmuth. Boosting as entropy projection. In *Proceedings of the 12th Annual Conference on Computational Learning Theory (COLT '99)*, pages 134–144, Santa Cruz, CA, USA, 1999.

Vladimir Koltchinskii and Dmitry Panchenko. Rademacher processes and bounding the risk of function learning. In *High Dimensional Probability II*, pages 443–459. Birkhäuser, 2000.

Vladimir Koltchinskii and Dmitry Panchenko. Empirical margin distributions and bounding the generalization error of combined classifiers. *Annals of Statistics*, 30(1):1–50, 2002.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

John Lafferty. Additive models, boosting, and inference for generalized divergences. In *Proceedings of the Twelfth Annual Conference on Computational Learning Theory (COLT '99)*, 1999.

Meizhu Liu and Baba C. Vemuri. Robust and efficient regularized boosting using total Bregman divergence (tBRLPBoost). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

Ximing Lu, Sean Welleck, Peter West, Liwei Jiang, Jungo Kasai, Daniel Khashabi, Ronan Le Bras, Lianhui Qin, Youngjae Yu, Rowan Zellers, et al. NeuroLogic A*esque decoding: Constrained text generation with lookahead heuristics. In *2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022*, pages 780–799. Association for Computational Linguistics (ACL), 2022.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594, 2023.

Pratyush Maini, Skyler Seto, Richard Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephrasing the web: A recipe for compute and data-efficient language modeling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14044–14072, 2024.

Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juarez, and Rik Sarkar. Combining generative artificial intelligence (AI) and the internet: Heading towards evolution or degradation? *arXiv preprint arXiv:2303.01255*, 2023.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. Adaptive computation and machine learning. MIT Press, second edition, 2018.

Mehryar Mohri, Jon Schneider, and Yifan Wu. Distributional alignment games for answer-level fine-tuning. In *Proceedings of the 43rd International Conference on Machine Learning*, ICML, 2026. URL https://arxiv.org/abs/2604.27166. To appear. Also available as arXiv:2604.27166.

G. Daniel Mostow. *Strong Rigidity of Locally Symmetric Spaces*, volume 78 of *Annals of Mathematics Studies*. Princeton University Press, 1973.

Abhijeet Mulgund and Chirag Pabbaraju. Relating misfit to gain in weak-to-strong generalization beyond the squared loss. In *International Conference on Machine Learning*, pages 45129–45150. PMLR, 2025.

Noboru Murata, Takashi Takenouchi, Takafumi Kanamori, and Shinto Eguchi. Information geometry of U-boost and Bregman divergence. *Neural Computation*, 16(7):1437–1481, 2004.

Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10):1100, 2020.

Richard Nock, Aditya Krishna Menon, and Cheng Soon Ong. A scaled Bregman theorem with applications. In *Advances in Neural Information Processing Systems 29*, 2016.

Mina Rezaei, Farzin Soleymani, Bernd Bischl, and Shekoofeh Azizi. Deep Bregman divergence for contrastive learning of visual representations. *arXiv preprint arXiv:2109.07455*, 2021.

Olivier Rioul. A historical perspective on Schützenberger-Pinsker inequalities. In Frank Nielsen and Frédéric Barbaresco, editors, *Geometric Science of Information - 6th International Conference, GSI 2023, St. Malo, France, August 30 - September 1, 2023, Proceedings, Part I*, volume 14071 of *Lecture Notes in Computer Science*, pages 291–306. Springer, 2023.

R. Tyrrell Rockafellar. *Convex analysis*. Princeton University Press, 1997.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019. URL https://arxiv.org/abs/1910.01108.

Rico Sennrich, Barry Haddow, and Alexandra Birch. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, 2016.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759, 2024.

Yuda Song, Hanlin Zhang, Carson Eisenach, Sham Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. *arXiv preprint arXiv:2412.02674*, 2024. URL https://arxiv.org/abs/2412.02674.

R. J. Spatzier. An invitation to rigidity theory. In M. Brin, B. Hasselblatt, and Y. Pesin, editors, *Modern Dynamical Systems and Applications*, pages 211–231. Cambridge University Press, 2004.

Koji Tsuda, Gunnar Rätsch, and Manfred K. Warmuth. Matrix exponentiated gradient updates for on-line learning and Bregman projection. *Journal of Machine Learning Research*, 6:995–1018, 2005. URL http://jmlr.org/papers/v6/tsuda05a.html.

Tim van Erven and Peter Harremoës. Rényi divergence and Kullback-Leibler divergence. *IEEE Trans. Inf. Theory*, 60(7):3797–3820, 2014.

Xinyi Wang, Hieu Pham, Zihang Dai, and Graham Neubig. Switchout: an efficient data augmentation algorithm for neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 856–861, 2018.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023a.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, 2023b.

Manfred K. Warmuth, Karen A. Glocer, and S. V. N. Vishwanathan. Entropy regularised LPBoost. In Yoav Freund, László Györfi, Gábor Turán, and T. Zeugmann, editors, *Algorithmic Learning Theory — Proceedings of the 19th International Conference (ALT 2008), Lecture Notes in Computer Science*, volume 5254 of *Lecture Notes in Computer Science*, pages 256–271. Springer, 2008. doi: 10.1007/978-3-540-87987-9_23. URL https://link.springer.com/chapter/10.1007/978-3-540-87987-9_23.

Jason Wei and Kai Zou. EDA: Easy data augmentation techniques for boosting performance on text classification tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6382–6388, 2019.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268, 2020.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, 2024.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. Synthetic continued pretraining. In *The Thirteenth International Conference on Learning Representations*, 2025.

Wei Yao, Gengze Xu, Huayi Tang, Wenkai Yang, Donglin Di, Ziqiao Wang, and Yong Liu. On weak-to-strong generalization and f-divergence. *arXiv preprint arXiv:2506.03109*, 2025a.

Wei Yao, Wenkai Yang, Ziqiao Wang, Yankai Lin, and Yong Liu. Revisiting weak-to-strong generalization in theory and practice: Reverse KL vs. forward KL. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 2860–2888, 2025b.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, pages 57905–57923, 2024.

Shun Zhang, Zhenfang Chen, Yikang Shen, Mingyu Ding, Joshua B Tenenbaum, and Chuang Gan. Planning with large language models for code generation. In *The Eleventh International Conference on Learning Representations*, 2023.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Prompt consistency for zero-shot task generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2613–2626, 2022.

Contents of the Appendix

A Related Work **19**

B Additional Preliminaries on Bregman Divergences **20**

B.1 Properties of Bregman divergences 20

B.2 Legendre functions 21

B.3 Centroid properties 21

B.4 Fenchel-Bregman inequality 22

B.5 Bregman divergence expectation 23

C Proofs for Section 4 **23**

C.1 Proof of Theorem 4.1 23

C.2 Proof of Theorem 4.2 25

C.3 Proof of Theorem 4.3 25

D Further Improvement Guarantees: Non-Realizable Targets, Finite Samples and Approximate Projections **27**

D.1 Improvement guarantees in the non-realizable setting 27

D.2 Improvement guarantees for the empirical Bregman projection 28

D.3 Discussion of Theorem D.2 30

D.4 Guarantees under strong convexity: empirical and approximate projections 30

E Two-Step Coherence Projection **32**

E.1 Pythagorean theorem equality 33

E.2 Comparison of direct and two-step projections 34

E.3 Improvement guarantees via two-step projection 35

E.4 Improvement guarantees for the relative entropy two-step projection 36

F Proofs for Section 5 **37**

F.1 Proof of Theorem 5.1 37

F.2 Proof of Theorem 5.2 38

F.3 Universal improvement for the Mahalanobis family 39

F.4 Proof of Theorem 5.3 39

F.5 Proof of Theorem 5.4 40

G Characterization and Rigidity of Improvement Mechanisms **41**

G.1 Characterization: provable improvement is Bregman projection 41

G.2 Symmetry rigidity 42

G.3 Affine structure 42

G.3.1 Finite-dimensional rigidity 43

G.3.2 Universal tie criterion 44

G.3.3 Exact universal partition 44

G.3.4 Bregman Pythagorean identity and cross-generator gap 45
G.4 Local rigidity for polyhedral constraints 46
G.5 Example: an emergent tie in the hyperplane case 48
G.6 Discussion: rigidity versus characterization 49

H Empirical Evaluations: Details **49**

I Limitations **54**

A Related Work

Our work sits at the intersection of several active research areas: heuristic self-improvement, invariance in machine learning, and the geometric theory of projections. We situate our contribution relative to these fields.

Weak-to-Strong Generalization and Projection-Based Improvement. Recent work on weak-to-strong generalization and self-improvement also studies projection-based mechanisms with Bregman-style guarantees [Burns et al., 2024, Charikar et al., 2024, Mulgund and Pabbaraju, 2025, Yao et al., 2025a,b]. This line of work is especially close to ours because it likewise identifies geometric post-processing as a route to provable improvement. Our results differ in several ways. First, we give realizable, non-realizable, relaxed and finite-sample guarantees for general Bregman projections (Sections 4 and D), whereas existing analyses are typically narrower in the divergences they consider [Charikar et al., 2024, Yao et al., 2025b]. Second, Theorem 4.2 proves a maximin characterization: the direct coherence projection is worst-case optimal among coherent candidates in terms of guaranteed improvement. Third, Sections 5 and 6 study simultaneous improvement across families of divergences and the characterization and rigidity consequences of demanding such strong guarantees, going beyond the single-divergence improvement perspective. Finally, we emphasize that the convex set is constructed from the model's own abilities (paraphrasing), which is what makes the mechanism a *self*-improvement mechanism.

Heuristic Self-Improvement and Self-Correction. A dominant paradigm for improving LLMs without human labels is to fine-tune them on their own outputs, a process often called self-instruction [Wang et al., 2023b]. This typically involves a "judge" model (which can be the model itself) to filter, score, or refine generations [Madaan et al., 2023, Huang et al., 2023, Zheng et al., 2023]. This approach has been formalized as Reinforcement Learning from AI Feedback (RLAIF), where the model learns from its own self-generated rewards [Yuan et al., 2024], or through principled self-critique, as in Constitutional AI [Bai et al., 2022]. While empirically powerful, these methods rely on the model's own, often biased [Xu et al., 2024], assessment of "quality" or "harmlessness." This can lead to instability, "model collapse" [Hataya et al., 2023, Martínez et al., 2023, Bertrand et al., 2024, Shumailov et al., 2024], or reinforcement of the model's own blind spots. Our framework differs fundamentally. Instead of relying on a subjective "quality" score, we use the objective, formal criterion of *coherence* as the improvement signal. We do not ask the model to *judge* its outputs, but rather enforce that its outputs be *mathematically consistent* under transformation, providing a more robust, non-heuristic foundation for improvement. We stress that our scope is correspondingly narrower: we do not provide a theory of self-rewarding, RLAIF or recursive data generation, but of coherence-based (self-consistency) improvement.

Inference-Time Mechanisms and Self-Consistency. A separate line of research modifies the decoding process at inference time. This includes generating candidate sets via sampling [Ji et al., 2024, Finlayson et al., 2024] or search [Zhang et al., 2023], and then re-weighting or selecting outputs based on model-internal signals [Fu et al., 2025, Chuang et al., 2024] or external constraints [Lu et al., 2022]. Perhaps the most relevant work in this area is *self-consistency* [Wang et al., 2023a], which improves reasoning by sampling multiple outputs from varied prompts (e.g., chain-of-thought variants [Wei et al., 2022]) and taking a majority vote. Self-consistency *leverages* a model's incoherence, that is, its tendency to give different answers to equivalent prompts, to find a more robust aggregate answer. Our work is the conceptual inverse: we seek to *eliminate* this incoherence from the model's parameters directly, not merely average it out at inference time. Our coherence projection can be seen as a training-time mechanism to "bake in" self-consistency, yielding a model that is inherently more reliable. The theoretical work of Huang et al. [2024] on training a model to learn its best-of-N (BoN) generations is also related, though our focus is on enforcing coherence across model outputs, rather than amplifying a model-defined "best" output.

Invariance, Data Augmentation, Consistency Regularization and Self-Distillation. The use of label-preserving transformations, or *data augmentation*, is a cornerstone of modern machine learning, especially in vision [Krizhevsky et al., 2012, Cubuk et al., 2019, 2020, Chen et al., 2020, DeVries and Taylor, 2017]. In NLP, this includes paraphrasing, back-translation, and token-level perturbations [Wei and Zou, 2019, Sennrich et al., 2016, Wang et al., 2018, Xie et al., 2020, Yang et al., 2025]. These methods all heuristically enforce invariance by showing the model examples of inputs (x) and transformations ($\Phi(x)$) that share a label, and existing theoretical analyses of data augmentation explain its benefits in terms of the *generalization* of a *supervised* objective (e.g., variance reduction

or implicit regularization induced by the augmented empirical risk). Our setting and claims are different in three respects. (i) No labels are used: the only information exploited is the equivalence between x and $\Phi(x)$, so the mechanism is a projection of the *model's own* output distributions onto a convex set, not a supervised risk minimization over an augmented sample. (ii) The guarantee is a *monotonic improvement with respect to the unknown ground truth* for every coherent ground truth (Theorem 4.1), together with an optimality statement (Theorem 4.2) and impossibility statements (Section 5); naive consistency regularization offers no such guarantee and may degrade a model. (iii) The characterization results of Section 6 show that any mechanism with such a guarantee must be a Bregman projection, which delineates exactly what label-free consistency enforcement can and cannot achieve. The work of Xie et al. [2020], which adds an unsupervised consistency loss, and Maini et al. [2024], which studies pre-training with paraphrases, are particularly close practical realizations of our mechanism, for which we now provide a general theoretical guarantee. Closer still, in NLP, are label-free *paraphrase-consistency* fine-tuning methods: Elazar et al. [2021] measure the inconsistency of pretrained language models across paraphrased factual queries and reduce it by fine-tuning on unlabeled paraphrases with a pairwise divergence penalty between the predictions on paraphrases, combined with a term that retains the original model's behavior, and Zhou et al. [2022] regularize the agreement of a model's predictions across paraphrased prompts on unlabeled data to improve zero-shot task generalization. These methods share the functional form of the penalized objective (5) that we evaluate in Section 7, and are empirical evidence for the mechanism we analyze. Our contribution relative to this line of work is to identify the objective as a Bregman projection onto a convex set containing the ground truth, which *derives* it from first principles (Section 4.3), and to establish what such label-free consistency enforcement can and cannot achieve: monotonic improvement with respect to the unknown ground truth, maximin optimality, robustness to relaxed, non-realizable, finite-sample and approximate projections, the multi-divergence dichotomy, and the characterization and rigidity results. Our perspective also connects to *self-distillation* [Hinton et al., 2015, Sanh et al., 2019, Gerstgrasser et al., 2024, Kim et al., 2025, Gu et al., 2024, Agarwal et al., 2024]. The theory of Allen-Zhu and Li [2023] suggests that gains from self-distillation can arise from an "implicit ensemble" effect. Our coherence projection offers a precise formulation of this: it finds a single model $\widehat{\pi}$ that is the Bregman centroid of the "ensemble" of incoherent views ($\pi_0(x)$ and $\pi_0(\Phi(x))$), thereby distilling their shared knowledge into one consistent model (Section E). Finally, Song et al. [2024] observe an incoherence between a model's ability to make decisions under uncertainty and its ability to predict that uncertainty; such "self-coherence" constraints also define convex sets to which our guarantees apply.

Bregman Projections in Machine Learning. The mathematical tools we use are grounded in a rich history of Bregman divergences in machine learning. Bregman divergences are central to maximum entropy models [Collins et al., 2002, Dudík et al., 2007, Altun and Smola, 2006, Banerjee et al., 2007, Frongillo and Reid, 2014, Harremoës, 2016, Nock et al., 2016, Cortes et al., 2015, Mohri et al., 2018], boosting algorithms [Collins et al., 2002, Kivinen and Warmuth, 1999, Warmuth et al., 2008, Tsuda et al., 2005, Lafferty, 1999, Murata et al., 2004, Liu and Vemuri, 2011, Rezaei et al., 2021], clustering [Banerjee et al., 2005], and Bayesian estimation [Frigyik et al., 2008]. Our contribution is to apply this geometric tool to the problem of model self-improvement, using projections onto coherence-defined sets as the core mechanism for provably enhancing model consistency and performance.

B Additional Preliminaries on Bregman Divergences

B.1 Properties of Bregman divergences

Bregman divergences admit several well-known properties (see Rockafellar [1997] or Appendix E of [Mohri et al., 2018], and [Nielsen, 2020] for an introductory text):

- Non-negativity: $\mathsf{B}_F \geq 0$, with $\mathsf{B}_F(\mathsf{p} \parallel \mathsf{q}) = 0$ iff $\mathsf{p} = \mathsf{q}$ when F is strictly convex.
- Convexity in the first argument: for a fixed q, the function $\mathsf{p} \mapsto \mathsf{B}_F(\mathsf{p} \parallel \mathsf{q})$ is convex in p.
- Asymmetry: in general, $\mathsf{B}_F(\mathsf{p} \parallel \mathsf{q}) \neq \mathsf{B}_F(\mathsf{q} \parallel \mathsf{p})$.
- Three-point identity: for any $\mathsf{p} \in \mathcal{K}$ and any $\mathsf{q}, \mathsf{r} \in \operatorname{int}(\mathcal{K}) = \Omega$,

$$\mathsf{B}_F(\mathsf{p} \parallel \mathsf{r}) + \mathsf{B}_F(\mathsf{r} \parallel \mathsf{q}) = \mathsf{B}_F(\mathsf{p} \parallel \mathsf{q}) + \langle \nabla F(\mathsf{q}) - \nabla F(\mathsf{r}), \mathsf{p} - \mathsf{r} \rangle.$$

  In particular, $\mathsf{B}_F(\mathsf{p} \parallel \mathsf{q}) + \mathsf{B}_F(\mathsf{q} \parallel \mathsf{p}) = \langle \nabla F(\mathsf{p}) - \nabla F(\mathsf{q}), \mathsf{p} - \mathsf{q} \rangle$.

- Projection: let $\mathcal{C} \subseteq \mathcal{K}$ be a non-empty, closed, and convex set; then the projection $\mathrm{Proj}_{\mathcal{C}}^{F}(\mathsf{q}) = \operatorname{argmin}_{\mathsf{p}\in\mathcal{C}} \mathsf{B}_F(\mathsf{p} \parallel \mathsf{q})$ is well defined for any q allowed in the definition of B_F. In our setting, where $\mathcal{C}$ is a closed subset of Π_{all}, $\mathcal{C}$ is compact and lower semicontinuity in p ensures the existence of a minimizer.
- Generalized Pythagorean theorem: let $\mathcal{C}$ be a non-empty, closed, and convex set, and let $\overline{\mathsf{q}} = \mathrm{Proj}_{\mathcal{C}}^{F}(\mathsf{q})$. Then, for any point $\mathsf{p} \in \mathcal{C}$,
$$\mathsf{B}_F(\mathsf{p} \parallel \mathsf{q}) \geq \mathsf{B}_F(\mathsf{p} \parallel \overline{\mathsf{q}}) + \mathsf{B}_F(\overline{\mathsf{q}} \parallel \mathsf{q}).$$
Equality holds for all $\mathsf{p} \in \mathcal{C}$ if $\mathcal{C}$ is an affine space (Theorem E.1). The inequality is equivalent to the first-order optimality condition $\langle \nabla F(\overline{\mathsf{q}}) - \nabla F(\mathsf{q}), \mathsf{p} - \overline{\mathsf{q}} \rangle \geq 0$ for all $\mathsf{p} \in \mathcal{C}$, by the three-point identity.
- Positive linearity in the generator: if F and G are two convex functions and $\alpha, \beta > 0$, then $\mathsf{B}_{\alpha F + \beta G} = \alpha \mathsf{B}_F + \beta \mathsf{B}_G$.
- Strong convexity and smoothness: if F is μ-strongly convex with respect to a norm $\|\cdot\|$, then $\mathsf{B}_F(\mathsf{p} \parallel \mathsf{q}) \geq \frac{\mu}{2}\|\mathsf{p} - \mathsf{q}\|^2$. We say that F is *L-smooth* with respect to $\|\cdot\|$ on a convex set U if ∇F is L-Lipschitz from $(U, \|\cdot\|)$ to $(\mathbb{R}^d, \|\cdot\|_*)$, where $\|\cdot\|_*$ is the dual norm: $\|\nabla F(\mathsf{p}) - \nabla F(\mathsf{q})\|_* \leq L\|\mathsf{p} - \mathsf{q}\|$; integrating along the segment $[\mathsf{q}, \mathsf{p}]$ then gives $\mathsf{B}_F(\mathsf{p} \parallel \mathsf{q}) \leq \frac{L}{2}\|\mathsf{p} - \mathsf{q}\|^2$. For instance, the negative entropy is 1-strongly convex with respect to $\|\cdot\|_1$ on $\Delta(\mathcal{Y})$ (Pinsker's inequality) and $(1/\alpha)$-smooth with respect to $\|\cdot\|_1$ on $\mathcal{S}_\alpha = \{\mathsf{p} : \mathsf{p}_i \geq \alpha\}$, since $\|\nabla F(\mathsf{p}) - \nabla F(\mathsf{q})\|_\infty = \max_i |\log \mathsf{p}_i - \log \mathsf{q}_i| \leq \frac{1}{\alpha}\|\mathsf{p} - \mathsf{q}\|_\infty \leq \frac{1}{\alpha}\|\mathsf{p} - \mathsf{q}\|_1$ by the mean value theorem.

B.2 Legendre functions

The following definition is based on [Rockafellar, 1997].

Definition B.1 (Functions of Legendre type)**.** *A proper, convex, and closed function $F : \mathcal{K} \to \mathbb{R}$ is of* Legendre type *if its behavior on its open interior $\Omega = \mathit{int}(\mathcal{K})$ is as follows:*

1. *F is differentiable on Ω.*
2. *The gradient mapping $\nabla F : \Omega \to \mathrm{int}(\mathrm{dom}(F^*))$ is a bijection.*

The bijection property is equivalent to F being strictly convex on Ω and *steep* (that is, for any sequence $\mathsf{p}_k \in \Omega$ converging to a boundary point $\mathsf{p}_\partial \in \partial\mathcal{K}$, the norm of the gradient explodes: $\|\nabla F(\mathsf{p}_k)\| \to +\infty$), which ensures that ∇F is surjective. When $\mathcal{K} = \mathbb{R}^d$ (e.g., quadratic generators), the steepness condition is vacuous.

When F is of Legendre type, its Fenchel conjugate F^* is also of Legendre type, and a number of standard properties hold:

- Uniqueness and interiority of projection: for any non-empty, closed, and convex set $\mathcal{C} \subseteq \mathcal{K}$ with $\mathcal{C} \cap \Omega \neq \varnothing$, the Bregman projection of a point $\mathsf{q} \in \Omega$ onto $\mathcal{C}$, $\mathrm{Proj}_{\mathcal{C}}^{F}(\mathsf{q})$, exists, is unique, and belongs to Ω [Bauschke and Borwein, 1997]. The interiority follows from the fact that a Legendre function has an empty subdifferential at every boundary point of its domain [Rockafellar, 1997, Theorem 26.1], while a minimizer of $F - \langle \nabla F(\mathsf{q}), \cdot \rangle$ over $\mathcal{C}$ must admit a subgradient. Throughout the paper, baselines satisfy $\pi_0(x) \in \Omega$ and every feasible set onto which we project meets $\Omega^{\mathcal{X}}$; by this property (applied to the Legendre function F of Lemma B.6), all projections $\widehat{\pi}$ considered satisfy $\widehat{\pi}(x) \in \Omega$ for all x, so that $\nabla F(\widehat{\pi}(x))$ and the divergences $\mathsf{B}_F(\cdot \parallel \widehat{\pi}(x))$ are well defined.
- Fenchel-Young equality: for any $\mathsf{p} \in \Omega$, $F(\mathsf{p}) + F^*(\nabla F(\mathsf{p})) = \langle \mathsf{p}, \nabla F(\mathsf{p}) \rangle$.
- Divergence duality: for any $\mathsf{p}, \mathsf{q} \in \Omega$, $\mathsf{B}_F(\mathsf{p} \parallel \mathsf{q}) = \mathsf{B}_{F^*}(\nabla F(\mathsf{q}) \parallel \nabla F(\mathsf{p}))$.

The negative entropy function and the norm squared functions are examples of Legendre functions generating the (unnormalized) D_{KL} and the Euclidean distance squared Bregman divergences.

B.3 Centroid properties

Bregman divergences admit closed-form expressions for centroids and projections, which are essential for our self-improvement mechanisms, in particular for the two-step projection method of Section E.

Lemma B.2. *Let $\lambda \in \Delta_p$ and $\mathsf{p}_1, \dots, \mathsf{p}_p \in \mathcal{K}$, and $\mathsf{q} \in \operatorname{int}(\operatorname{dom}(F))$. Then, the following holds:*

$$\sum_{k=1}^{p} \lambda_k \mathsf{B}_F(\mathsf{p}_k \parallel \mathsf{q}) - \mathsf{B}_F\Big(\sum_{k=1}^{p} \lambda_k \mathsf{p}_k \parallel \mathsf{q}\Big) = \sum_{k=1}^{p} \lambda_k F(\mathsf{p}_k) - F\Big(\sum_{k=1}^{p} \lambda_k \mathsf{p}_k\Big).$$

Proof. By definition of the Bregman divergence, $\mathsf{B}_F(\mathsf{p}_k \parallel \mathsf{q}) = F(\mathsf{p}_k) - F(\mathsf{q}) - \langle \nabla F(\mathsf{q}), \mathsf{p}_k - \mathsf{q}\rangle$. Summing over k with weights λ_k gives $\sum_k \lambda_k \mathsf{B}_F(\mathsf{p}_k \parallel \mathsf{q}) = \sum_k \lambda_k F(\mathsf{p}_k) - F(\mathsf{q}) - \langle \nabla F(\mathsf{q}), \sum_k \lambda_k \mathsf{p}_k - \mathsf{q}\rangle$. Similarly, $\mathsf{B}_F(\sum_k \lambda_k \mathsf{p}_k \parallel \mathsf{q}) = F(\sum_k \lambda_k \mathsf{p}_k) - F(\mathsf{q}) - \langle \nabla F(\mathsf{q}), \sum_k \lambda_k \mathsf{p}_k - \mathsf{q}\rangle$. Subtracting the second from the first yields $\sum_k \lambda_k F(\mathsf{p}_k) - F(\sum_k \lambda_k \mathsf{p}_k)$, which is independent of q, as claimed. □

This identity shows that the Bregman divergence from a weighted combination of points to a reference q decomposes in a way independent of q. This property underlies the computation of centroids and justifies that Bregman projections can be expressed in terms of gradient averages.

Lemma B.3. *Let F be Legendre and $\mathsf{q}_1, \dots, \mathsf{q}_p \in \Omega$ with weights $\lambda \in \Delta_p$. Then, the minimizer of $\mathsf{p} \mapsto \sum_{k=1}^{p} \lambda_k \mathsf{B}_F(\mathsf{p} \parallel \mathsf{q}_k)$ over Ω satisfies*

$$\nabla F(\mathsf{p}^*) = \sum_{k=1}^{p} \lambda_k \nabla F(\mathsf{q}_k), \quad \text{or equivalently } \mathsf{p}^* = (\nabla F)^{-1}\Big(\sum_{k=1}^{p} \lambda_k \nabla F(\mathsf{q}_k)\Big).$$

If $\mathcal{C} \subseteq \mathcal{K}$ is closed and convex, the unique minimizer over $\mathcal{C}$ is the Bregman projection of p^* onto $\mathcal{C}$: $\mathsf{p}^*_{\mathcal{C}} = \operatorname{argmin}_{\mathsf{p} \in \mathcal{C}} \mathsf{B}_F(\mathsf{p} \parallel \mathsf{p}^*)$.

Proof. By the duality identity for Legendre functions, $\sum_k \lambda_k \mathsf{B}_F(\mathsf{p} \parallel \mathsf{q}_k) = \sum_k \lambda_k \mathsf{B}_{F^*}(\nabla F(\mathsf{q}_k) \parallel \nabla F(\mathsf{p}))$. Using Lemma B.2 (applied to F^*), this is equal, up to an additive constant independent of p, to

$$\mathsf{B}_{F^*}\Big(\sum_{k=1}^{p} \lambda_k \nabla F(\mathsf{q}_k) \parallel \nabla F(\mathsf{p})\Big) = \mathsf{B}_F\Big(\mathsf{p} \parallel (\nabla F)^{-1}\Big(\sum_{k=1}^{p} \lambda_k \nabla F(\mathsf{q}_k)\Big)\Big).$$

Thus, the unconstrained minimizer is $\mathsf{p}^* = (\nabla F)^{-1}(\sum_k \lambda_k \nabla F(\mathsf{q}_k))$. The identity $\sum_k \lambda_k \mathsf{B}_F(\mathsf{p} \parallel \mathsf{q}_k) = \mathsf{B}_F(\mathsf{p} \parallel \mathsf{p}^*) + \text{const}$ in fact holds for all $\mathsf{p} \in \mathcal{K}$ (not only $\mathsf{p} \in \Omega$): expanding the definitions, both sides equal $F(\mathsf{p})$ minus the same linear function of p, since $\nabla F(\mathsf{p}^*) = \sum_k \lambda_k \nabla F(\mathsf{q}_k)$, up to a constant. Hence, for a closed convex set $\mathcal{C} \subseteq \mathcal{K}$, the minimizer over $\mathcal{C}$ is the Bregman projection of p^* onto $\mathcal{C}$, as required. □

Corollary B.4. *Let p^* and $\mathsf{p}^*_{\mathcal{C}}$ be as above. Then,*

$$\min_{\mathsf{p} \in \mathcal{C}} \sum_{k=1}^{p} \lambda_k \mathsf{B}_F(\mathsf{p} \parallel \mathsf{q}_k) = \sum_{k=1}^{p} \lambda_k F^*(\nabla F(\mathsf{q}_k)) - F^*\Big(\sum_{k=1}^{p} \lambda_k \nabla F(\mathsf{q}_k)\Big) + \mathsf{B}_F(\mathsf{p}^*_{\mathcal{C}} \parallel \mathsf{p}^*).$$

Proof. Immediate from the proof of Lemma B.3: the additive constant is $\sum_k \lambda_k F^*(\nabla F(\mathsf{q}_k)) - F^*(\sum_k \lambda_k \nabla F(\mathsf{q}_k))$ by Lemma B.2, and the minimum of the remaining term over $\mathcal{C}$ is $\mathsf{B}_F(\mathsf{p}^*_{\mathcal{C}} \parallel \mathsf{p}^*)$. □

B.4 Fenchel-Bregman inequality

The following general inequality bounds the inner product of differences by Bregman divergences; it generalizes the familiar Cauchy–Schwarz-type control of cross terms using the geometry of Bregman divergences and is a standard tool for bounding inner-product terms in the analysis of projections.

Lemma B.5 (Fenchel–Bregman inequality)**.** *For any Legendre function F and $u, v \in \Omega = \operatorname{int}(\operatorname{dom}(F))$ with $\alpha \in \operatorname{int}(\operatorname{dom}(F^*)) = \nabla F(\Omega)$ and $\beta = \nabla F(v)$, the following inequality holds:*

$$\langle u - v, \alpha - \beta\rangle \le \mathsf{B}_F(u \parallel v) + \mathsf{B}_{F^*}(\alpha \parallel \beta).$$

Proof. By definition of the Bregman divergences:

$$\begin{aligned}
&\mathsf{B}_F(u \parallel v) + \mathsf{B}_{F^*}(\alpha \parallel \beta) - \langle u - v, \alpha - \beta\rangle \\
&= F(u) - F(v) - \langle \nabla F(v), u - v\rangle + F^*(\alpha) - F^*(\beta) - \langle \nabla F^*(\beta), \alpha - \beta\rangle - \langle u - v, \alpha - \beta\rangle \\
&= \big(F(u) + F^*(\alpha) - \langle u, \alpha\rangle\big) - \big(F(v) + F^*(\beta) - \langle v, \beta\rangle\big) \\
&\quad + \langle u - v, \beta - \nabla F(v)\rangle - \langle \nabla F^*(\beta) - v, \alpha - \beta\rangle.
\end{aligned}$$

Now set $\beta = \nabla F(v)$. Since F is Legendre, $\nabla F^* = (\nabla F)^{-1}$, so $\nabla F^*(\beta) = v$. The last two inner-product terms vanish. By Fenchel duality, $F(u) + F^*(\alpha) - \langle u, \alpha\rangle \geq 0$ and $F(v) + F^*(\beta) - \langle v, \beta\rangle = 0$, completing the proof. □

B.5 Bregman divergence expectation

Our analysis is based on the expectation over $\mathcal{X}$ of a Bregman divergence. We equip the (finite-dimensional) space of all functions $(\mathbb{R}^d)^{\mathcal{X}} = \{\pi : \mathcal{X} \to \mathbb{R}^d\}$ with the $\mathcal{D}_{\mathcal{X}}$-weighted inner product $\langle \pi, \pi'\rangle = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\langle \pi(x), \pi'(x)\rangle]$ and the associated norm $\|\pi\|_{L_2} = \langle \pi, \pi\rangle^{1/2}$. Throughout the paper, gradients, Fenchel conjugates, orthogonal complements and subdifferentials of functionals on $(\mathbb{R}^d)^{\mathcal{X}}$ are taken with respect to this inner product; all statements (optimality conditions, Pythagorean identities, projections) are of course independent of this choice, but the pointwise expressions below are specific to it (with the unweighted Euclidean pairing, the gradient of F at π would be $x \mapsto \mathcal{D}_{\mathcal{X}}(x)\nabla F(\pi(x))$). We write $\mathcal{K}^{\mathcal{X}} = \{\pi : \pi(x) \in \mathcal{K}\ \forall x\}$ and $\Omega^{\mathcal{X}} = \{\pi : \pi(x) \in \Omega\ \forall x\}$, where $\Omega = \operatorname{int}(\mathcal{K})$; since $\mathcal{X}$ is finite, $\Omega^{\mathcal{X}}$ is the interior of $\mathcal{K}^{\mathcal{X}}$ and the boundary points of $\mathcal{K}^{\mathcal{X}}$ are the $\pi \in \mathcal{K}^{\mathcal{X}}$ with $\pi(x) \in \partial\mathcal{K}$ for some x.

Lemma B.6. *Let $F : \mathcal{K} \to \mathbb{R}$ be convex, extended by $+\infty$ outside $\mathcal{K}$, and differentiable on $\Omega = \operatorname{int}(\mathcal{K})$, and define $\mathsf{F}(\pi) = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[F(\pi(x))] = \sum_{x\in\mathcal{X}} \mathcal{D}_{\mathcal{X}}(x)F(\pi(x))$ for $\pi \in (\mathbb{R}^d)^{\mathcal{X}}$, so that $\operatorname{dom}\mathsf{F} = \mathcal{K}^{\mathcal{X}}$. Then F is convex, differentiable on $\Omega^{\mathcal{X}}$ with gradient $(\nabla\mathsf{F}(\pi))(x) = \nabla F(\pi(x))$, and its Bregman divergence satisfies $\mathsf{B}_{\mathsf{F}}(\pi \,\|\, \pi') = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi(x) \,\|\, \pi'(x))]$ for $\pi \in \mathcal{K}^{\mathcal{X}}$ and $\pi' \in \Omega^{\mathcal{X}}$. If further F is of Legendre type, then F is a Legendre function on $(\mathbb{R}^d)^{\mathcal{X}}$ with $\operatorname{int}(\operatorname{dom}\mathsf{F}) = \Omega^{\mathcal{X}}$, and its conjugate is $\mathsf{F}^*(\theta) = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[F^*(\theta(x))]$.*

Proof. Convexity of F follows from that of F, since $\mathcal{D}_{\mathcal{X}}(x) > 0$ for all x. For $\pi \in \Omega^{\mathcal{X}}$ and $v \in (\mathbb{R}^d)^{\mathcal{X}}$, the directional derivative of F at π in the direction v is $\sum_x \mathcal{D}_{\mathcal{X}}(x)\langle \nabla F(\pi(x)), v(x)\rangle = \langle \nabla F \circ \pi, v\rangle$, so F is differentiable on $\Omega^{\mathcal{X}}$ with gradient $(\nabla\mathsf{F}(\pi))(x) = \nabla F(\pi(x))$ with respect to the weighted inner product. This gives the Bregman divergence formula $\mathsf{B}_{\mathsf{F}}(\pi \,\|\, \pi') = \mathsf{F}(\pi) - \mathsf{F}(\pi') - \langle \nabla\mathsf{F}(\pi'), \pi - \pi'\rangle = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi(x) \,\|\, \pi'(x))]$.

Now, assume F is Legendre. F is proper, since it is finite on $\mathcal{K}^{\mathcal{X}} \neq \varnothing$ and nowhere $-\infty$. It is closed (lower semicontinuous on $(\mathbb{R}^d)^{\mathcal{X}}$), as a finite sum of the lower semicontinuous functions $\pi \mapsto \mathcal{D}_{\mathcal{X}}(x)F(\pi(x))$, each of which is the composition of the closed function F with a continuous linear map. Its domain is $\mathcal{K}^{\mathcal{X}}$, whose interior is $\Omega^{\mathcal{X}}$, on which F is differentiable, and F is strictly convex on $\Omega^{\mathcal{X}}$: if $\pi \neq \pi'$ in $\Omega^{\mathcal{X}}$, they differ at some x, where strict convexity of F on Ω gives a strict inequality, and all other terms are convex. Finally, F is steep: if (π_n) in $\Omega^{\mathcal{X}}$ converges to a boundary point π of $\mathcal{K}^{\mathcal{X}}$, then $\pi_n(x) \to \pi(x) \in \partial\mathcal{K}$ for some x, so $\|\nabla F(\pi_n(x))\| \to \infty$ by steepness of F, and $\|\nabla\mathsf{F}(\pi_n)\|_{L_2}^2 \geq \mathcal{D}_{\mathcal{X}}(x)\|\nabla F(\pi_n(x))\|_2^2 \to \infty$. Hence F is essentially smooth and essentially strictly convex, i.e., of Legendre type [Rockafellar, 1997, Section 26]; in particular, $\nabla\mathsf{F}$ is a bijection from $\Omega^{\mathcal{X}}$ onto $\operatorname{int}(\operatorname{dom}\mathsf{F}^*)$. Explicitly, since ∇F is a bijection from Ω onto $\operatorname{int}(\operatorname{dom}F^*)$, the pointwise mapping $\pi \mapsto (x \mapsto \nabla F(\pi(x)))$ is a bijection from $\Omega^{\mathcal{X}}$ onto the set of functions $\theta : \mathcal{X} \to \operatorname{int}(\operatorname{dom}F^*)$, which is the interior of the domain of F^*: indeed, the supremum defining the conjugate with respect to the weighted pairing decouples across x, $\mathsf{F}^*(\theta) = \sup_\pi \sum_x \mathcal{D}_{\mathcal{X}}(x)[\langle \pi(x), \theta(x)\rangle - F(\pi(x))] = \sum_x \mathcal{D}_{\mathcal{X}}(x)F^*(\theta(x))$, whence $\operatorname{dom}\mathsf{F}^* = (\operatorname{dom}F^*)^{\mathcal{X}}$. □

In particular, by the interiority property recalled in Section B.2, if $\pi_0 \in \Omega^{\mathcal{X}}$ and $\mathcal{C}$ is a closed convex subset of $(\mathbb{R}^d)^{\mathcal{X}}$ with $\mathcal{C} \cap \Omega^{\mathcal{X}} \neq \varnothing$, the B_{F}-projection of π_0 onto $\mathcal{C} \cap \mathcal{K}^{\mathcal{X}}$ exists, is unique and lies in $\Omega^{\mathcal{X}}$, even when $\mathcal{C}$ meets the boundary of $\mathcal{K}^{\mathcal{X}}$ (as Π_{all} and $\mathcal{C}_{\text{coh}}$ do when $\mathcal{K} = \mathbb{R}^d_+$).

C Proofs for Section 4

C.1 Proof of Theorem 4.1

Theorem 4.1 (Bregman projection onto a convex set improves the baseline)**.** *Let $\mathcal{C}$ be a closed convex set such that $\mathcal{C} \cap \Pi$ is non-empty and contains the ideal model $\pi^* : \mathcal{X} \to \Delta(\mathcal{Y})$. Then, any solution $\widehat{\pi}$ of the convex optimization problem*

$$\min_{\pi\in\mathcal{C}\cap\Pi} \quad \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi(x) \,\|\, \pi_0(x))] \tag{2}$$

satisfies

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x)\,\|\,\widehat{\pi}(x))] \le \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x)\,\|\,\pi_0(x))] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\widehat{\pi}(x)\,\|\,\pi_0(x))]$$
$$\le \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x)\,\|\,\pi_0(x))].$$

If, in addition, Φ is an involution, $\mathcal{C}\subseteq\mathcal{C}_{\mathrm{coh}}$ (e.g., $\mathcal{C}=\mathcal{C}_{\mathrm{coh}}$), F is μ-strongly convex with respect to a norm $\|\cdot\|$, and $\lambda(x)=\frac{\mathbb{P}[x]}{\mathbb{P}[x]+\mathbb{P}[\Phi(x)]}$, then

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x)\,\|\,\widehat{\pi}(x))] \le \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x)\,\|\,\pi_0(x))]$$
$$-\frac{\mu}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[\lambda(x)(1-\lambda(x))\|\pi_0(x)-\pi_0(\Phi(x))\|^2\Big].$$

Proof. By Lemma B.6, the expectation of the Bregman divergence B_F, denoted B_F, is the Bregman divergence associated to $\mathsf{F}(\pi)=\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[F(\pi(x))]$. Thus, $\widehat{\pi}$ is the F-Bregman projection of π_0 on the closed convex set $\mathcal{C}\cap\Pi$. By the generalized Pythagorean theorem for Bregman divergences, for $\pi^*\in\mathcal{C}\cap\Pi$:

$$\mathsf{B}_\mathsf{F}(\pi^*\,\|\,\pi_0)\ge\mathsf{B}_\mathsf{F}(\pi^*\,\|\,\widehat{\pi})+\mathsf{B}_\mathsf{F}(\widehat{\pi}\,\|\,\pi_0).$$

Thus, we have

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x)\,\|\,\widehat{\pi}(x))] = \mathsf{B}_\mathsf{F}(\pi^*\,\|\,\widehat{\pi})$$
$$\le \mathsf{B}_\mathsf{F}(\pi^*\,\|\,\pi_0)-\mathsf{B}_\mathsf{F}(\widehat{\pi}\,\|\,\pi_0)$$
$$= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x)\,\|\,\pi_0(x))]-\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\widehat{\pi}(x)\,\|\,\pi_0(x))]$$
$$\le \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x)\,\|\,\pi_0(x))].$$

This proves the first inequalities. For the explicit bound, assume $\mathcal{C}\subseteq\mathcal{C}_{\mathrm{coh}}$, so that $\widehat{\pi}$ is coherent, and that Φ is an involution. Since Φ is a bijection of $\mathcal{X}$, $\sum_x g(\Phi(x))=\sum_x g(x)$ for any g, and we can write:

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\widehat{\pi}(x)\,\|\,\pi_0(x))]$$
$$=\frac{1}{2}\sum_{x\in\mathcal{X}}[\mathbb{P}[x]\mathsf{B}_F(\widehat{\pi}(x)\,\|\,\pi_0(x))+\mathbb{P}[\Phi(x)]\mathsf{B}_F(\widehat{\pi}(\Phi(x))\,\|\,\pi_0(\Phi(x)))]$$
$$=\frac{1}{2}\sum_{x\in\mathcal{X}}[\mathbb{P}[x]\mathsf{B}_F(\widehat{\pi}(x)\,\|\,\pi_0(x))+\mathbb{P}[\Phi(x)]\mathsf{B}_F(\widehat{\pi}(x)\,\|\,\pi_0(\Phi(x)))]$$
$$=\frac{1}{2}\sum_{x\in\mathcal{X}}(\mathbb{P}[x]+\mathbb{P}[\Phi(x)])\Big[\lambda(x)\mathsf{B}_F(\widehat{\pi}(x)\,\|\,\pi_0(x))+(1-\lambda(x))\mathsf{B}_F(\widehat{\pi}(x)\,\|\,\pi_0(\Phi(x)))\Big].$$

Let the term inside the square brackets be denoted $W(x,\widehat{\pi})$. By the μ-strong convexity of F with respect to $\|\cdot\|$, we have $\mathsf{B}_F(\mathsf{p}\,\|\,\mathsf{q})\ge\frac{\mu}{2}\|\mathsf{p}-\mathsf{q}\|^2$, thus

$$W(x,\widehat{\pi})\ge\frac{\mu}{2}\Big[\lambda(x)\|\widehat{\pi}(x)-\pi_0(x)\|^2+(1-\lambda(x))\|\widehat{\pi}(x)-\pi_0(\Phi(x))\|^2\Big].$$

Now, for any norm, any $\mathsf{r},\mathsf{p},\mathsf{q}$ and $\lambda\in[0,1]$, writing $a=\|\mathsf{r}-\mathsf{p}\|$ and $b=\|\mathsf{r}-\mathsf{q}\|$, the triangle inequality $\|\mathsf{p}-\mathsf{q}\|\le a+b$ and the identity $\lambda a^2+(1-\lambda)b^2-\lambda(1-\lambda)(a+b)^2=(\lambda a-(1-\lambda)b)^2\ge 0$ give

$$\lambda\|\mathsf{r}-\mathsf{p}\|^2+(1-\lambda)\|\mathsf{r}-\mathsf{q}\|^2\ge\lambda(1-\lambda)\|\mathsf{p}-\mathsf{q}\|^2.$$

Applying this with $\mathsf{r}=\widehat{\pi}(x)$, $\mathsf{p}=\pi_0(x)$ and $\mathsf{q}=\pi_0(\Phi(x))$ yields $W(x,\widehat{\pi})\ge\frac{\mu}{2}\lambda(x)(1-\lambda(x))\|\pi_0(x)-\pi_0(\Phi(x))\|^2$, and therefore

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\widehat{\pi}(x)\,\|\,\pi_0(x))]\ge\frac{1}{2}\sum_{x\in\mathcal{X}}(\mathbb{P}[x]+\mathbb{P}[\Phi(x)])\frac{\mu}{2}\lambda(x)(1-\lambda(x))\|\pi_0(x)-\pi_0(\Phi(x))\|^2$$
$$=\frac{\mu}{2}\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\Big[\lambda(x)(1-\lambda(x))\|\pi_0(x)-\pi_0(\Phi(x))\|^2\Big],$$

where the last equality uses again the invariance of the sum under $x\mapsto\Phi(x)$ and the fact that $\lambda(\Phi(x))(1-\lambda(\Phi(x)))=\lambda(x)(1-\lambda(x))$. Combining this with the main inequality yields the explicit bound stated in the theorem. □

C.2 Proof of Theorem 4.2

Theorem 4.2 (Maximin optimality). *Let Π be a closed convex family of models, F a Legendre function, and $\widehat{\pi}$ the Bregman projection of π_0 onto $\mathcal{C}_{\mathrm{coh}} \cap \Pi$ (a solution of (2) with $\mathcal{C} = \mathcal{C}_{\mathrm{coh}}$). Then,*

$$\max_{\pi \in \mathcal{C}_{\mathrm{coh}} \cap \Pi} \min_{\pi^* \in \mathcal{C}_{\mathrm{coh}} \cap \Pi} \mathrm{Improv}_{\pi^*}(\pi) = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))],$$

the maximum is attained at $\pi = \widehat{\pi}$, and the worst-case reference for $\widehat{\pi}$ is $\pi^* = \widehat{\pi}$.

Proof. We know $\widehat{\pi} \in \mathcal{C}_{\mathrm{coh}} \cap \Pi$. For any $\pi \in \mathcal{C}_{\mathrm{coh}} \cap \Pi$, taking the ground truth $\pi^* = \widehat{\pi}$ gives

$$\begin{aligned}\min_{\pi^* \in \mathcal{C}_{\mathrm{coh}} \cap \Pi} \mathrm{Improv}_{\pi^*}(\pi) \le \mathrm{Improv}_{\widehat{\pi}}(\pi) &= \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))] - \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi(x))] \\ &\le \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))],\end{aligned}$$

hence $\max_{\pi \in \mathcal{C}_{\mathrm{coh}} \cap \Pi} \min_{\pi^* \in \mathcal{C}_{\mathrm{coh}} \cap \Pi} \mathrm{Improv}_{\pi^*}(\pi) \le \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))]$. On the other hand, by Theorem 4.1 (see (3)),

$$\max_{\pi \in \mathcal{C}_{\mathrm{coh}} \cap \Pi} \min_{\pi^* \in \mathcal{C}_{\mathrm{coh}} \cap \Pi} \mathrm{Improv}_{\pi^*}(\pi) \ge \min_{\pi^* \in \mathcal{C}_{\mathrm{coh}} \cap \Pi} \mathrm{Improv}_{\pi^*}(\widehat{\pi}) \ge \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))].$$

The two bounds coincide, so the maximin value equals $\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))]$, the maximum is attained at $\pi = \widehat{\pi}$, and, since $\mathrm{Improv}_{\widehat{\pi}}(\widehat{\pi}) = \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))]$, the inner minimum for $\pi = \widehat{\pi}$ is attained at $\pi^* = \widehat{\pi}$. □

C.3 Proof of Theorem 4.3

Theorem 4.3 (Improvement guarantees for relaxed constraints). *Assume that Φ is an involution, that F is μ_F-strongly convex with respect to a norm $\|\cdot\|$ on $\mathbb{R}^d$, and that $\mathsf{D}(\mathsf{p}, \mathsf{q}) \ge \frac{\mu_{\mathsf{D}}}{2} \|\mathsf{p} - \mathsf{q}\|^2$ for all $\mathsf{p}, \mathsf{q} \in \Delta(\mathcal{Y})$. Let $\|\pi\|_{L_2} = [\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}} \|\pi(x)\|^2]^{1/2}$ and define the incoherence gap of π_0 as $\Delta_{\mathrm{coh}} = \inf_{\pi \in \mathcal{C}_{\mathrm{coh}}} \|\pi_0 - \pi\|_{L_2}^2$. Let $\widehat{\pi}$ be the solution of (4). Then, for every $\pi^* \in \mathcal{C}_\Lambda^{\mathsf{D}}$,*

$$\mathrm{Improv}_{\pi^*}(\widehat{\pi}) \ge \frac{\mu_F}{2} \left[\sqrt{\Delta_{\mathrm{coh}}} - \sqrt{\frac{\Lambda}{2\mu_{\mathsf{D}}}} \right]_+^2.$$

Furthermore, with $\gamma_0 = \|\pi_0 - \pi_0 \circ \Phi\|_{L_2}^2$ the (computable) incoherence of π_0 and $C_\Phi = \sup_{\pi \neq 0} \|\pi \circ \Phi\|_{L_2} / \|\pi\|_{L_2} = \max_x [\mathcal{D}_{\mathcal{X}}(\Phi(x)) / \mathcal{D}_{\mathcal{X}}(x)]^{1/2}$ the operator norm of the composition with Φ, we have $\frac{\gamma_0}{(1+C_\Phi)^2} \le \Delta_{\mathrm{coh}} \le \frac{\gamma_0}{4}$; in particular $\Delta_{\mathrm{coh}} = \gamma_0/4$ when $\mathcal{D}_{\mathcal{X}}$ is Φ-invariant.

Proof. By the joint convexity and lower semicontinuity of D (which hold for all the divergences listed in Section 4.3), $\pi \mapsto \mathbb{E}[\mathsf{D}(\pi(x), \pi(\Phi(x)))]$ is convex and lower semicontinuous, thus $\mathcal{C}_\Lambda^{\mathsf{D}}$ is closed and convex; it contains $\mathcal{C}_{\mathrm{coh}}$, hence meets $(\operatorname{int} \mathcal{K})^{\mathcal{X}}$, and the B_F-projection $\widehat{\pi}$ of π_0 onto $\mathcal{C}_\Lambda^{\mathsf{D}}$ is well defined.

By Theorem 4.1 applied with $\mathcal{C} = \mathcal{C}_\Lambda^{\mathsf{D}}$ and $\Pi = \Pi_{\mathrm{all}}$, for every $\pi^* \in \mathcal{C}_\Lambda^{\mathsf{D}}$, the improvement of $\widehat{\pi}$ over π_0 is lower bounded as follows:

$$\begin{aligned}\mathrm{Improv}_{\pi^*}(\widehat{\pi}) &= \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))] - \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] \\ &\ge \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))].\end{aligned}$$

By the μ_F-strong convexity of F, we have

$$\mathrm{Improv}_{\pi^*}(\widehat{\pi}) \ge \inf_{\pi \in \mathcal{C}_\Lambda^{\mathsf{D}}} \mathbb{E}[\mathsf{B}_F(\pi(x) \,\|\, \pi_0(x))] \ge \inf_{\pi \in \mathcal{C}_\Lambda^{\mathsf{D}}} \frac{\mu_F}{2} \|\pi - \pi_0\|_{L_2}^2 = \frac{\mu_F}{2} \left[\mathrm{dist}_{L_2}(\pi_0, \mathcal{C}_\Lambda^{\mathsf{D}}) \right]^2.$$

For any $\pi \in \mathcal{C}_\Lambda^{\mathsf{D}}$, by the lower bound on D, we can write:

$$\Lambda \ge \mathbb{E}[\mathsf{D}(\pi(x), \pi(\Phi(x)))] \ge \frac{\mu_{\mathsf{D}}}{2} \|\pi - \pi \circ \Phi\|_{L_2}^2.$$

Note that for any $\pi \in \Pi_{\text{all}}$, $\pi_{\text{sym}} = \frac{1}{2}[\pi + \pi \circ \Phi]$ is in $\mathcal{C}_{\text{coh}}$, since Φ is an involution. Thus, for any $\pi \in \mathcal{C}^{\mathsf{D}}_{\Lambda}$, we have

$$\operatorname{dist}_{L_2}(\pi, \mathcal{C}_{\text{coh}}) \le \left\|\pi - \pi_{\text{sym}}\right\|_{L_2} = \frac{1}{2}\|\pi - \pi \circ \Phi\|_{L_2} \le \frac{1}{2}\sqrt{\frac{2\Lambda}{\mu_{\mathsf{D}}}} = \sqrt{\frac{\Lambda}{2\mu_{\mathsf{D}}}} =: r.$$

Thus, $\mathcal{C}^{\mathsf{D}}_{\Lambda}$ lies in $\mathcal{C}^{r}_{\text{coh}}$, the L_2-tube of radius r around $\mathcal{C}_{\text{coh}}$, which implies $\operatorname{dist}_{L_2}(\pi_0, \mathcal{C}^{\mathsf{D}}_{\Lambda}) \ge \operatorname{dist}_{L_2}(\pi_0, \mathcal{C}^{r}_{\text{coh}})$ (see Figure 1). For any $c^r \in \mathcal{C}^{r}_{\text{coh}}$, there exists $c \in \mathcal{C}_{\text{coh}}$ with $\|c - c^r\|_{L_2} \le r$. Thus, by the triangle inequality, $\|\pi_0 - c^r\|_{L_2} \ge \|\pi_0 - c\|_{L_2} - \|c^r - c\|_{L_2} \ge \|\pi_0 - c\|_{L_2} - r$. Taking the infimum over $c \in \mathcal{C}_{\text{coh}}$ on the right-hand side gives $\|\pi_0 - c^r\|_{L_2} \ge \operatorname{dist}_{L_2}(\pi_0, \mathcal{C}_{\text{coh}}) - r$. Next, taking the infimum over $c^r \in \mathcal{C}^{r}_{\text{coh}}$ on the left-hand side yields $\operatorname{dist}_{L_2}(\pi_0, \mathcal{C}^{r}_{\text{coh}}) \ge \operatorname{dist}_{L_2}(\pi_0, \mathcal{C}_{\text{coh}}) - r$. Since the distance is non-negative and $\sqrt{\Delta_{\text{coh}}} = \operatorname{dist}_{L_2}(\pi_0, \mathcal{C}_{\text{coh}})$, we have:

$$\operatorname{dist}_{L_2}(\pi_0, \mathcal{C}^{\mathsf{D}}_{\Lambda}) \ge \operatorname{dist}_{L_2}(\pi_0, \mathcal{C}^{r}_{\text{coh}}) \ge \left[\sqrt{\Delta_{\text{coh}}} - r\right]_{+}.$$

This proves the main improvement bound.

We now prove the bounds on Δ_{coh}. Since $\pi_{\text{sym}} = \frac{1}{2}(\pi_0 + \pi_0 \circ \Phi)$ is in $\mathcal{C}_{\text{coh}}$, we have the upper bound

$$\sqrt{\Delta_{\text{coh}}} = \operatorname{dist}_{L_2}(\pi_0, \mathcal{C}_{\text{coh}}) \le \|\pi_0 - \pi_{\text{sym}}\|_{L_2} = \frac{1}{2}\|\pi_0 - \pi_0 \circ \Phi\|_{L_2} = \frac{1}{2}\sqrt{\gamma_0}.$$

For the lower bound, let π be any model in $\mathcal{C}_{\text{coh}}$, so $\pi = \pi \circ \Phi$. By the triangle inequality,

$$\sqrt{\gamma_0} = \|\pi_0 - \pi_0 \circ \Phi\|_{L_2} = \|(\pi_0 - \pi) - (\pi_0 - \pi) \circ \Phi\|_{L_2} \le \|\pi_0 - \pi\|_{L_2} + \|(\pi_0 - \pi) \circ \Phi\|_{L_2} \le (1 + C_\Phi)\|\pi_0 - \pi\|_{L_2},$$

by definition of the operator norm C_Φ (note that, Φ being a bijection with $\Phi^{-1} = \Phi$, the change of variables $y = \Phi(x)$ gives $\|\pi \circ \Phi\|^2_{L_2} = \sum_x \mathcal{D}_{\mathcal{X}}(x)\|\pi(\Phi(x))\|^2 = \sum_y \mathcal{D}_{\mathcal{X}}(\Phi(y))\|\pi(y)\|^2$, whence the expression of C_Φ in the statement). Since this holds for any $\pi \in \mathcal{C}_{\text{coh}}$, it holds for the infimum: $\sqrt{\Delta_{\text{coh}}} \ge \sqrt{\gamma_0}/(1 + C_\Phi)$. Squaring the upper and lower bounds gives the result. Finally, when $\mathcal{D}_{\mathcal{X}}$ is Φ-invariant, $\|\pi \circ \Phi\|_{L_2} = \|\pi\|_{L_2}$ for all π, so $C_\Phi = 1$ and the two bounds coincide: $\Delta_{\text{coh}} = \gamma_0/4$. □

Remark C.1 (Duality between the constrained and penalized problems)**.** *We make explicit the equivalence between Problems (4) and (5) invoked in Section 4.3. Let* $\operatorname{pen}(\pi) = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{D}(\pi(x), \pi(\Phi(x)))]$ *and* $B(\pi) = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi(x) \parallel \pi_0(x))]$. *(i)* From Λ to λ. *For* $\Lambda > 0$, *the uniform model* $\pi_{\text{unif}}(y \mid x) = 1/d$ *is coherent, so* $\operatorname{pen}(\pi_{\text{unif}}) = 0 < \Lambda$, *and it lies in* $\operatorname{relint}\Pi_{\text{all}} \cap (\operatorname{int}\mathcal{K})^{\mathcal{X}}$, *where* B *is finite; thus Slater's condition holds for the convex problem (4), and there exists a multiplier* $\lambda \ge 0$ *such that the solution* $\widehat{\pi}$ *of (4) minimizes* $B + \lambda \operatorname{pen}$ *over* Π_{all}, *i.e., solves (5), with* $\lambda(\operatorname{pen}(\widehat{\pi}) - \Lambda) = 0$. *(ii)* From λ to Λ. *Conversely, fix* $\lambda \ge 0$. *The objective* $J_\lambda = B + \lambda \operatorname{pen}$ *is lower semicontinuous on the compact set* Π_{all}, *so it admits a minimizer* $\widehat{\pi}_\lambda$. *Writing* $J_\lambda = \mathsf{F} + g$ *with* $g = \lambda \operatorname{pen} - \langle \nabla \mathsf{F}(\pi_0), \cdot \rangle + I_{\Pi_{\text{all}}} + \text{const}$, *where* $\operatorname{dom} g$ *is sandwiched between* $\operatorname{relint}\Pi_{\text{all}}$ *(on which* D *is finite) and* Π_{all}, *so that* $\operatorname{relint}\operatorname{dom} g = \operatorname{relint}\Pi_{\text{all}}$ *meets* $\operatorname{relint}\operatorname{dom}\mathsf{F} \supseteq (\operatorname{int}\mathcal{K})^{\mathcal{X}}$, *the subdifferential sum rule [Rockafellar, 1997, Theorem 23.8] gives* $0 \in \partial\mathsf{F}(\widehat{\pi}_\lambda) + \partial g(\widehat{\pi}_\lambda)$; *in particular* $\partial\mathsf{F}(\widehat{\pi}_\lambda) \ne \varnothing$, *hence* $\widehat{\pi}_\lambda \in (\operatorname{int}\mathcal{K})^{\mathcal{X}}$ *by essential smoothness of* F *(Lemma B.6). Since* F *is strictly convex on* $(\operatorname{int}\mathcal{K})^{\mathcal{X}}$ *and* g *is convex, the minimizer is unique. Now let* $\Lambda(\lambda) = \operatorname{pen}(\widehat{\pi}_\lambda)$. *For every* $\pi \in \mathcal{C}^{\mathsf{D}}_{\Lambda(\lambda)}$, $B(\pi) \ge B(\widehat{\pi}_\lambda) + \lambda[\operatorname{pen}(\widehat{\pi}_\lambda) - \operatorname{pen}(\pi)] \ge B(\widehat{\pi}_\lambda)$, *so* $\widehat{\pi}_\lambda$ *is the (unique)* B_F*-projection of* π_0 *onto* $\mathcal{C}^{\mathsf{D}}_{\Lambda(\lambda)}$, *i.e., the solution of (4) with* $\Lambda = \Lambda(\lambda)$. *Consequently, Theorem 4.3 and the observations preceding it apply verbatim to the solution of the penalized problem (5), with* $\Lambda = \Lambda(\lambda)$, *a quantity computable from the trained model;* $\Lambda(\lambda)$ *is non-increasing in* λ *and tends to* 0 *as* $\lambda \to \infty$.

Remark C.2 (Arbitrary partitions of $\mathcal{X}$)**.** *The invariance mapping* Φ *serves only to specify a partition of* $\mathcal{X}$ *into equivalence classes of task-equivalent prompts, and* $\mathcal{C}_{\text{coh}}$ *is the set of models that are constant on these classes. Consequently, all results that depend on* $\mathcal{C}_{\text{coh}}$ *only through its being a closed convex set, namely the first statement of Theorem 4.1, Theorem 4.2, the results of Section D, Theorems E.1 and E.4 and all of Section G, hold verbatim for an arbitrary partition, however it is obtained. The results stated for an involution extend to classes* $O \subseteq \mathcal{X}$ *of arbitrary finite size by replacing pairwise quantities with their class-level counterparts. The orbit average* $\widehat{\pi}_{\text{orbit}}(x)$ *becomes the* $\mathcal{D}_{\mathcal{X}}$*-weighted average of* π_0 *over the class of* x, *and Theorem 5.2 follows from the same Jensen argument. The gain* $\lambda(x)(1 - \lambda(x))\|\pi_0(x) - \pi_0(\Phi(x))\|^2$ *in Theorem 4.1 becomes the weighted within-class dispersion* $\frac{1}{4W_O}\sum_{x,x' \in O} \mathbb{P}[x]\,\mathbb{P}[x']\|\pi_0(x) - \pi_0(x')\|^2$, $W_O = \sum_{x \in O} \mathbb{P}[x]$, *with the same proof, the two-point identity being replaced by* $\|\mathsf{p} - \mathsf{q}\|^2 \le 2(a^2 + b^2)$ *(at the price of a*

factor 2 in the constant for non-Euclidean norms). The pairwise penalty in Theorem 4.3 becomes the average pairwise divergence within the class, with π_{sym} replaced by the class average, and the first-step projection of Section E remains the right Bregman centroid of the class (Lemma B.3 applies to any finite number of points). Only the two-point computations of Corollary E.6 and of the rank-one argument in Theorem 5.3(b) are specific to classes of size two.

D Further Improvement Guarantees: Non-Realizable Targets, Finite Samples and Approximate Projections

D.1 Improvement guarantees in the non-realizable setting

For the following result, we no longer assume that π^* is in $\mathcal{C}_{\mathrm{coh}} \cap \Pi$.

Theorem D.1 (Coherent Bregman projection with non-realizable π^*)**.** *Let $\overline{\pi}$ be any model in $\mathcal{C}_{\mathrm{coh}} \cap \Pi$ with $\overline{\pi}(x) \in \mathrm{int}(\mathrm{dom}\, F)$ for all x, and let $\epsilon = \mathbb{E}_{x\sim\mathcal{D}_X}[\mathsf{B}_F(\pi^*(x) \,\|\, \overline{\pi}(x))]$; the bounds below are most favorable when $\overline{\pi}$ minimizes this quantity, in which case ϵ is the B_F-distance of π^* to $\mathcal{C}_{\mathrm{coh}} \cap \Pi$. Then, the B_F-projection $\widehat{\pi}$ of π_0 onto $\mathcal{C}_{\mathrm{coh}} \cap \Pi$ satisfies:*

$$\begin{aligned}&\mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] - \mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))]\\ &\quad\le -\mathbb{E}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))] + \mathbb{E}[\langle \pi^*(x) - \overline{\pi}(x), \nabla F(\pi_0(x)) - \nabla F(\widehat{\pi}(x))\rangle].\end{aligned}$$

If F is μ-strongly convex with respect to a norm $\|\cdot\|$, with dual norm $\|\cdot\|_*$, then

$$\begin{aligned}&\mathop{\mathbb{E}}_{x\sim\mathcal{D}_X}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_X}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))]\\ &\quad\le -\mathop{\mathbb{E}}_{x\sim\mathcal{D}_X}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))] + \sqrt{\frac{2\epsilon}{\mu}}\sqrt{\mathop{\mathbb{E}}_{x\sim\mathcal{D}_X}[\|\nabla F(\pi_0(x)) - \nabla F(\widehat{\pi}(x))\|_*^2]}.\end{aligned}$$

If, furthermore, F is L-smooth with respect to $\|\cdot\|$ on a convex set containing the values of π_0 and $\widehat{\pi}$ (that is, $\|\nabla F(\mathsf{p}) - \nabla F(\mathsf{q})\|_* \le L\|\mathsf{p} - \mathsf{q}\|$ there, see Section B.1), then

$$\mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] - \mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))] \le -D + \frac{2L}{\mu}\sqrt{\epsilon D}, \qquad D = \mathbb{E}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))],$$

and the right-hand side is strictly negative whenever $D > 0$ (i.e., $\pi_0 \notin \mathcal{C}_{\mathrm{coh}} \cap \Pi$) and $\epsilon < \left[\frac{\mu}{2L}\right]^2 D$. If F is L-smooth with respect to $\|\cdot\|$ on all of $\mathbb{R}^d$ (e.g., a quadratic generator), the bound improves to $-D + 2\sqrt{(L/\mu)\epsilon D}$, which is strictly negative whenever $D > 0$ and $\epsilon < \frac{\mu}{4L} D$.

Proof. By the three-point identity for Bregman divergences, we can write:

$$\begin{aligned}&\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))\\ &= \mathsf{B}_F(\pi^*(x) \,\|\, \overline{\pi}(x)) + \mathsf{B}_F(\overline{\pi}(x) \,\|\, \widehat{\pi}(x)) + \langle \pi^*(x) - \overline{\pi}(x), \nabla F(\overline{\pi}(x)) - \nabla F(\widehat{\pi}(x))\rangle\\ &\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))\\ &= \mathsf{B}_F(\pi^*(x) \,\|\, \overline{\pi}(x)) + \mathsf{B}_F(\overline{\pi}(x) \,\|\, \pi_0(x)) + \langle \pi^*(x) - \overline{\pi}(x), \nabla F(\overline{\pi}(x)) - \nabla F(\pi_0(x))\rangle.\end{aligned}$$

Subtracting the second identity from the first one and taking expectations yields:

$$\begin{aligned}&\mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] - \mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))]\\ &= \mathbb{E}[\mathsf{B}_F(\overline{\pi}(x) \,\|\, \widehat{\pi}(x)) - \mathsf{B}_F(\overline{\pi}(x) \,\|\, \pi_0(x))] + \mathbb{E}[\langle \pi^*(x) - \overline{\pi}(x), \nabla F(\pi_0(x)) - \nabla F(\widehat{\pi}(x))\rangle].\end{aligned}$$

Now, by the Pythagorean theorem, since $\widehat{\pi}$ is the projection of π_0 onto $\mathcal{C}_{\mathrm{coh}} \cap \Pi$ and since $\overline{\pi}$ is in $\mathcal{C}_{\mathrm{coh}} \cap \Pi$, we have $\mathbb{E}[\mathsf{B}_F(\overline{\pi}(x) \,\|\, \widehat{\pi}(x))] \le \mathbb{E}[\mathsf{B}_F(\overline{\pi}(x) \,\|\, \pi_0(x))] - \mathbb{E}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))]$. Using this to bound the first term on the right-hand side yields the first statement. By the strong convexity of F, $\epsilon = \mathbb{E}_{x\sim\mathcal{D}_X}[\mathsf{B}_F(\pi^*(x) \,\|\, \overline{\pi}(x))] \ge \frac{\mu}{2}\mathbb{E}_{x\sim\mathcal{D}_X}\left[\|\pi^*(x) - \overline{\pi}(x)\|^2\right]$. Thus, by Hölder's and the Cauchy–Schwarz inequalities,

$$\begin{aligned}&\mathbb{E}[\langle \pi^*(x) - \overline{\pi}(x), \nabla F(\pi_0(x)) - \nabla F(\widehat{\pi}(x))\rangle]\\ &\le \sqrt{\mathbb{E}[\|\pi^*(x) - \overline{\pi}(x)\|^2]}\sqrt{\mathbb{E}[\|\nabla F(\pi_0(x)) - \nabla F(\widehat{\pi}(x))\|_*^2]}\\ &\le \sqrt{\frac{2\epsilon}{\mu}}\sqrt{\mathbb{E}[\|\nabla F(\pi_0(x)) - \nabla F(\widehat{\pi}(x))\|_*^2]},\end{aligned}$$

which gives the second statement. If F is L-smooth, then

$$\mathbb{E}\big[\|\nabla F(\pi_0(x)) - \nabla F(\widehat{\pi}(x))\|_*^2\big] \le L^2\,\mathbb{E}\big[\|\pi_0(x) - \widehat{\pi}(x)\|^2\big] \le \frac{2L^2}{\mu}\,\mathbb{E}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))] = \frac{2L^2}{\mu} D.$$

Thus, the right-hand side of the second statement is at most $-D + \sqrt{2\epsilon/\mu}\sqrt{2L^2 D/\mu} = -D + \frac{2L}{\mu}\sqrt{\epsilon D}$, which, for $D > 0$, is negative iff $\frac{2L}{\mu}\sqrt{\epsilon D} < D$, that is $\epsilon < \left[\frac{\mu}{2L}\right]^2 D$.

Finally, if F is L-smooth on all of $\mathbb{R}^d$, then its conjugate F^* is $(1/L)$-strongly convex with respect to $\|\cdot\|_*$ (strong convexity/smoothness duality, see Kakade et al. 2012), and the duality identity $\mathsf{B}_F(\mathsf{p} \,\|\, \mathsf{q}) = \mathsf{B}_{F^*}(\nabla F(\mathsf{q}) \,\|\, \nabla F(\mathsf{p}))$ yields the cocoercivity inequality $\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x)) \ge \frac{1}{2L}\|\nabla F(\pi_0(x)) - \nabla F(\widehat{\pi}(x))\|_*^2$. Taking expectations, $\mathbb{E}[\|\nabla F(\pi_0(x)) - \nabla F(\widehat{\pi}(x))\|_*^2] \le 2LD$, and the right-hand side of the second statement is at most $-D + \sqrt{2\epsilon/\mu}\sqrt{2LD} = -D + 2\sqrt{(L/\mu)\epsilon D}$, which is negative iff $\epsilon < \frac{\mu}{4L} D$ (for $D > 0$). Since $L \ge \mu$, this threshold improves on the previous one by the factor L/μ. □

The theorem shows that when F is μ-strongly convex, for π^* approximately realizable (ϵ small), the coherent projection still guarantees an improvement over the baseline π_0. Moreover, if F is also L-smooth, a strict improvement is guaranteed whenever $\epsilon < \left[\frac{\mu}{2L}\right]^2 D$: enforcing coherence helps whenever the ground truth is closer to the coherent set than the displacement D it induces on the baseline. Conversely, if the transformations are far from task-preserving (ϵ large), no improvement is guaranteed, and the projection may indeed introduce bias; the quantity ϵ is the right measure of the quality of the transformations. In the special case of the unnormalized relative entropy where F is the negative entropy, F is 1-strongly convex with respect to $\|\cdot\|_1$ on $\Delta(\mathcal{Y})$ (Pinsker's inequality), thus $\mu = 1$ with $\|\cdot\|_* = \|\cdot\|_\infty$. In general, however, F is not globally smooth, since $(\nabla F)_i(\mathsf{p}) = 1 + \log \mathsf{p}_i$ and the Hessian is diagonal with entries $1/\mathsf{p}_i$. Nevertheless, if the models considered take values in $\mathcal{S}_\alpha = \{\mathsf{p} : \mathsf{p}_i \ge \alpha\}$, then $\|\nabla F(\mathsf{p}) - \nabla F(\mathsf{q})\|_\infty = \max_i |\log \mathsf{p}_i - \log \mathsf{q}_i| \le \frac{1}{\alpha}\|\mathsf{p} - \mathsf{q}\|_\infty \le \frac{1}{\alpha}\|\mathsf{p} - \mathsf{q}\|_1$, so F is L-smooth on $\mathcal{S}_\alpha$ with respect to $\|\cdot\|_1$ with $L \le 1/\alpha$, and the third statement applies with $\mu = 1$ and $L = 1/\alpha$.

D.2 Improvement guarantees for the empirical Bregman projection

In many practical settings, we do not have a full description of the underlying distribution $\mathcal{D}_\mathcal{X}$ over prompts, but merely an empirical estimate based on an i.i.d. sample S. This prevents us from directly computing the projection with respect to the expected Bregman divergence, and instead requires us to work with its empirical counterpart. In this section, we establish performance guarantees for the resulting empirical solution. Crucially, these results do not rely on structural assumptions about the objective functional, such as strong convexity.

Theorem D.2 (Guarantees for the empirical Bregman projection)**.** *Let $\mathcal{C}$ be a closed convex set and $\overline{\Pi} = \mathcal{C} \cap \Pi$. Let $S = (x_1, \dots, x_m)$ be a sample of size m drawn i.i.d. from $\mathcal{D}_\mathcal{X}$, and let $\widehat{\mathcal{D}}_\mathcal{X}$ denote the corresponding empirical distribution. Define the function class*

$$\mathcal{G} = \big\{x \mapsto \mathsf{B}_F(\pi(x) \,\|\, \pi_0(x)) : \pi \in \overline{\Pi}\big\} \cup \big\{x \mapsto \mathsf{B}_F(\pi'(x) \,\|\, \pi(x)) : \pi, \pi' \in \overline{\Pi}\big\},$$

where we assume that $\mathsf{B}_F(\pi'(x) \,\|\, \pi(x))$ is well defined and bounded for all $\pi, \pi' \in \overline{\Pi}$ and $x \in \mathcal{X}$ (e.g., $\overline{\Pi} \subseteq \mathcal{S}_\alpha^\mathcal{X}$ with $\mathcal{S}_\alpha = \{\mathsf{p} : \mathsf{p}_i \ge \alpha\}$ for the negative entropy, or F finite and differentiable on a neighborhood of $\Delta(\mathcal{Y})$ as for quadratic generators), and set $\epsilon_m = \sup_{g \in \mathcal{G}} \big|\mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[g(x)] - \mathbb{E}_{x \sim \widehat{\mathcal{D}}_\mathcal{X}}[g(x)]\big|$. Let $\widehat{\pi}$ denote a (population) minimizer of $\pi \mapsto \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi(x) \,\|\, \pi_0(x))]$ over $\overline{\Pi}$ and $\widehat{\pi}_S$ the corresponding empirical minimizer of $\pi \mapsto \mathbb{E}_{x \sim \widehat{\mathcal{D}}_\mathcal{X}}[\mathsf{B}_F(\pi(x) \,\|\, \pi_0(x))]$ over $\overline{\Pi}$. Assume that $\pi^* \in \overline{\Pi}$. Then,

$$\mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_S(x))] \le \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] + 4\epsilon_m + \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) - \mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))].$$

In particular, if $\pi^* = \widehat{\pi}$, then the bound simplifies to $\mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_S(x))] \le \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] + 4\epsilon_m$.

Proof. The proof consists of combining the empirical and population optimality of $\widehat{\pi}$ and $\widehat{\pi}_S$, applying the three-point Bregman identity to relate their gradients via π^*, and bounding the resulting difference using uniform deviations. We seek a bound on the expectation of $\Delta(x) = \mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_S(x)) - \mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))$. By the three-point identity, for any $x \in \mathcal{X}$,

$$\begin{aligned}\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_S(x)) = \mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x)) + \mathsf{B}_F(\widehat{\pi}(x) \,\|\, \widehat{\pi}_S(x)) \\ + \langle \nabla F(\widehat{\pi}(x)) - \nabla F(\widehat{\pi}_S(x)), \pi^*(x) - \widehat{\pi}(x) \rangle.\end{aligned}$$

In view of that, using the identity $\mathsf{B}_F(a \,\|\, b) + \mathsf{B}_F(b \,\|\, a) = \langle \nabla F(a) - \nabla F(b), a - b \rangle$, we can rewrite $\Delta(x)$ as follows:

$$\begin{aligned}\Delta(x) &= \mathsf{B}_F(\widehat{\pi}(x) \,\|\, \widehat{\pi}_S(x)) + \langle \nabla F(\widehat{\pi}(x)) - \nabla F(\widehat{\pi}_S(x)), \pi^*(x) - \widehat{\pi}(x) \rangle \\ &= \langle \nabla F(\widehat{\pi}(x)) - \nabla F(\widehat{\pi}_S(x)), \widehat{\pi}(x) - \widehat{\pi}_S(x) \rangle + \langle \nabla F(\widehat{\pi}(x)) - \nabla F(\widehat{\pi}_S(x)), \pi^*(x) - \widehat{\pi}(x) \rangle \\ &\quad - \mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \widehat{\pi}(x)) \\ &= \langle \nabla F(\widehat{\pi}(x)) - \nabla F(\widehat{\pi}_S(x)), \pi^*(x) - \widehat{\pi}_S(x) \rangle - \mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \widehat{\pi}(x)).\end{aligned}$$

Define $\Psi(x)$ to be the first term on the right-hand side: $\Psi(x) = \langle \nabla F(\widehat{\pi}(x)) - \nabla F(\widehat{\pi}_S(x)), \pi^*(x) - \widehat{\pi}_S(x) \rangle$, so that $\Delta(x) = \Psi(x) - \mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \widehat{\pi}(x))$. We bound $\mathbb{E}[\Delta]$ by decomposing it into the sum of two terms $\mathbb{E}[\Delta] = (\mathbb{E} - \widehat{\mathbb{E}})[\Delta] + \widehat{\mathbb{E}}[\Delta]$, where we use the shorthand $\widehat{\mathbb{E}}$ to denote the empirical expectation. Since $\Delta(x) = \mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_S(x)) - \mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))$ is the difference of two functions in $\mathcal{G}$, by definition of ϵ_m we have $(\mathbb{E} - \widehat{\mathbb{E}})[\Delta] \le 2\epsilon_m$. We now bound $\widehat{\mathbb{E}}[\Delta]$, and start with $\widehat{\mathbb{E}}[\Psi]$. By the optimality of $\widehat{\pi}_S$ as an empirical minimizer over the convex set $\overline{\Pi}$, for any $\pi \in \overline{\Pi}$, we have $\widehat{\mathbb{E}}[\langle \nabla F(\widehat{\pi}_S(x)) - \nabla F(\pi_0(x)), \pi(x) - \widehat{\pi}_S(x) \rangle] \ge 0$. Thus, since π^* is in $\overline{\Pi}$, we can write

$$\begin{aligned}\widehat{\mathbb{E}}[\Psi(x)] &= \widehat{\mathbb{E}}[\langle \nabla F(\widehat{\pi}(x)) - \nabla F(\pi_0(x)), \pi^*(x) - \widehat{\pi}_S(x) \rangle] \\ &\quad - \widehat{\mathbb{E}}[\langle \nabla F(\widehat{\pi}_S(x)) - \nabla F(\pi_0(x)), \pi^*(x) - \widehat{\pi}_S(x) \rangle] \\ &\le \widehat{\mathbb{E}}[\langle \nabla F(\widehat{\pi}(x)) - \nabla F(\pi_0(x)), \pi^*(x) - \widehat{\pi}_S(x) \rangle].\end{aligned}$$

We now write the right-hand side as the difference of two terms and apply the three-point identity to each:

$$\begin{aligned}&\widehat{\mathbb{E}}[\langle \nabla F(\widehat{\pi}(x)) - \nabla F(\pi_0(x)), \pi^*(x) - \widehat{\pi}_S(x) \rangle] \\ &= \widehat{\mathbb{E}}[\langle \nabla F(\widehat{\pi}(x)) - \nabla F(\pi_0(x)), \pi^*(x) - \pi_0(x) \rangle] \\ &\quad - \widehat{\mathbb{E}}[\langle \nabla F(\widehat{\pi}(x)) - \nabla F(\pi_0(x)), \widehat{\pi}_S(x) - \pi_0(x) \rangle] \\ &= \widehat{\mathbb{E}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) + \mathsf{B}_F(\pi_0(x) \,\|\, \widehat{\pi}(x)) - \mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] \\ &\quad - \widehat{\mathbb{E}}[\mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \pi_0(x)) + \mathsf{B}_F(\pi_0(x) \,\|\, \widehat{\pi}(x)) - \mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \widehat{\pi}(x))] \\ &= \widehat{\mathbb{E}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) - \mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \pi_0(x))] \\ &\quad - \widehat{\mathbb{E}}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] + \widehat{\mathbb{E}}[\mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \widehat{\pi}(x))].\end{aligned}$$

Dropping the nonpositive term $-\widehat{\mathbb{E}}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] \le 0$ yields

$$\widehat{\mathbb{E}}[\Psi(x)] \le \widehat{\mathbb{E}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) - \mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \pi_0(x))] + \widehat{\mathbb{E}}[\mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \widehat{\pi}(x))].$$

Subtracting $\widehat{\mathbb{E}}[\mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \widehat{\pi}(x))]$ from both sides, the last term cancels and we obtain

$$\widehat{\mathbb{E}}[\Delta(x)] = \widehat{\mathbb{E}}[\Psi(x)] - \widehat{\mathbb{E}}[\mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \widehat{\pi}(x))] \le \widehat{\mathbb{E}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) - \mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \pi_0(x))].$$

Replacing the empirical difference by population plus deviations gives

$$\widehat{\mathbb{E}}[\Delta(x)] \le \mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) - \mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \pi_0(x))] + 2\epsilon_m,$$

since $\left|(\widehat{\mathbb{E}} - \mathbb{E})[\mathsf{B}_F(\pi(x) \,\|\, \pi_0(x))]\right| \le \epsilon_m$ for all $\pi \in \overline{\Pi}$. Combining with the bound on $(\mathbb{E} - \widehat{\mathbb{E}})[\Delta]$, we have

$$\mathbb{E}[\Delta(x)] \le 4\epsilon_m + \mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) - \mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \pi_0(x))].$$

Finally, since $\widehat{\pi}$ is a population minimizer over $\overline{\Pi} \ni \widehat{\pi}_S$, we have $\mathbb{E}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))] \le \mathbb{E}[\mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \pi_0(x))]$, and therefore

$$\mathbb{E}[\Delta(x)] \le 4\epsilon_m + \mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) - \mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))],$$

which is the statement of the theorem. □

The quantity ϵ_m is precisely the empirical process term that is controlled by generalization bounds. In particular, ϵ_m can be bounded (with high probability over the draw of the sample S) in terms of the Rademacher complexity of the family $\mathcal{G}$, the sample size m, and an upper bound on the function values within that family (see, for example, Koltchinskii and Panchenko 2000, 2002, Bartlett and Mendelson 2002, Mohri et al. 2018). The theorem bounds the excess divergence $\mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_S(x))] - \mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi})]$ by the sum of an estimation error $4\epsilon_m$ and an approximation error $\mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) - \mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))]$. The approximation term is inherent and non-negative by the Bregman Pythagorean theorem. It is small when π^* is close to π_0 or to $\widehat{\pi}$.

Corollary D.3 (Guarantee for empirical improvement)**.** *Under the assumptions of Theorem D.2,*

$$\begin{aligned}\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_S(x))] - \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))] \\ \leq 4\epsilon_m + \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))] - 2\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))].\end{aligned}$$

Proof. The statement follows immediately from Theorem D.2 and the Bregman Pythagorean theorem (Theorem 4.1). □

The corollary shows that the empirical solution $\widehat{\pi}_S$ is a definite improvement over π_0 when the inherent benefit of the projection, captured by $2\,\mathbb{E}[\mathsf{B}_F(\widehat{\pi} \,\|\, \pi_0)]$, is large enough to overcome the statistical cost of finite-sample estimation.

D.3 Discussion of Theorem D.2

The guarantee provided by Theorem D.2 is somewhat subtle, and its proof is correspondingly involved. We first explain why its specific form is necessary, showing that a more direct guarantee on the improvement is not possible without additional assumptions. We then show that under a standard strong-convexity assumption, one obtains the intuitive guarantee one might initially expect, and in fact a more general approximate-projection guarantee.

If we could prove an inequality of the type $|\mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_S(x))] - \mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))]| \leq C\epsilon_m$ for some positive constant C, this would lead directly to an improvement inequality of the form $\mathrm{Improv}_{\pi^*}(\widehat{\pi}_S) \geq \mathrm{Improv}_{\pi^*}(\widehat{\pi}) - C\epsilon_m$. However, this inequality is unlikely to hold in the general case. To see why, consider a standard approach where we bound the population difference by the empirical difference plus the generalization error:

$$\mathbb{E}[\mathsf{B}_F(\pi^* \,\|\, \widehat{\pi}_S)] - \mathbb{E}[\mathsf{B}_F(\pi^* \,\|\, \widehat{\pi})] \leq \left(\mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}_\mathcal{X}}[\mathsf{B}_F(\pi^* \,\|\, \widehat{\pi}_S)] - \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}_\mathcal{X}}[\mathsf{B}_F(\pi^* \,\|\, \widehat{\pi})]\right) + 2\epsilon_m.$$

For the desired inequality to hold, the empirical term in the parenthesis would need to be non-positive. Yet, there is no reason for this to be true. The solution $\widehat{\pi}_S$ was chosen to minimize the distance to π_0 on the sample, not the distance to π^*. We only know that $\mathbb{E}_{x\sim\widehat{\mathcal{D}}_\mathcal{X}}[\mathsf{B}_F(\widehat{\pi}_S \,\|\, \pi_0)] \leq \mathbb{E}_{x\sim\widehat{\mathcal{D}}_\mathcal{X}}[\mathsf{B}_F(\widehat{\pi} \,\|\, \pi_0)]$. It is entirely possible for $\widehat{\pi}_S$ to be closer to π_0 on the sample while simultaneously being farther from π^*. This mismatch is precisely why the proof of the theorem is more involved and must include a *misalignment term*, $\mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) - \mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))]$, which is the price we pay for the difference between our optimization objective (proximity to π_0) and our evaluation goal (proximity to π^*).

D.4 Guarantees under strong convexity: empirical and approximate projections

Lemma D.4 (Two-sided guarantee for the empirical minimizer)**.** *Under the assumptions of Theorem D.2 and using the same notation, the following guarantees hold:*

1. *The population objective value for the empirical solution $\widehat{\pi}_S$ is tightly bounded around that of the population solution $\widehat{\pi}$:*

$$\mathbb{E}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))] \leq \mathbb{E}[\mathsf{B}_F(\widehat{\pi}_S(x) \,\|\, \pi_0(x))] \leq \mathbb{E}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))] + 2\epsilon_m.$$

2. *The improvement of the empirical solution is bounded below as follows:* $\mathrm{Improv}_{\pi^*}(\widehat{\pi}_S) \geq \mathbb{E}[\mathsf{B}_F(\widehat{\pi}(x) \,\|\, \pi_0(x))] - \mathbb{E}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}(x))] - 4\epsilon_m$.

Proof. For the first part, the left-hand side inequality is a direct consequence of the optimality of $\widehat{\pi}$. For the right-hand side, we have the following chain of inequalities:

$$\begin{aligned}
\mathbb{E}[\mathsf{B}_F(\widehat{\pi}_S(x) \parallel \pi_0(x))] &\le \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}_{\mathcal{X}}}[\mathsf{B}_F(\widehat{\pi}_S(x) \parallel \pi_0(x))] + \epsilon_m && \text{(by definition of } \epsilon_m) \\
&\le \mathop{\mathbb{E}}_{x\sim\widehat{\mathcal{D}}_{\mathcal{X}}}[\mathsf{B}_F(\widehat{\pi}(x) \parallel \pi_0(x))] + \epsilon_m && \text{(by optimality of } \widehat{\pi}_S) \\
&\le \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\widehat{\pi}(x) \parallel \pi_0(x))] + 2\epsilon_m. && \text{(by definition of } \epsilon_m)
\end{aligned}$$

For the second part, the proof of Theorem D.2 provides the intermediate inequality $\mathbb{E}[\mathsf{B}_F(\pi^*(x) \parallel \widehat{\pi}_S(x)) - \mathsf{B}_F(\pi^*(x) \parallel \widehat{\pi}(x))] \le 4\epsilon_m + \mathbb{E}[\mathsf{B}_F(\pi^*(x) \parallel \pi_0(x)) - \mathsf{B}_F(\widehat{\pi}_S(x) \parallel \pi_0(x))]$. Rearranging to isolate $\mathrm{Improv}_{\pi^*}(\widehat{\pi}_S)$ gives $\mathrm{Improv}_{\pi^*}(\widehat{\pi}_S) \ge \mathbb{E}[\mathsf{B}_F(\widehat{\pi}_S(x) \parallel \pi_0(x))] - \mathbb{E}[\mathsf{B}_F(\pi^*(x) \parallel \widehat{\pi}(x))] - 4\epsilon_m$, and the first part yields $\mathbb{E}[\mathsf{B}_F(\widehat{\pi}_S(x) \parallel \pi_0(x))] \ge \mathbb{E}[\mathsf{B}_F(\widehat{\pi}(x) \parallel \pi_0(x))]$, which completes the proof. □

The following general statement charges *any* suboptimality in the projection objective, whatever its source (finite sample, model mismatch, inexact or non-convex optimization), to the improvement guarantee.

Theorem D.5 (Improvement guarantee for approximate projections)**.** *Let $\mathcal{Q} \subseteq \Pi_{\mathrm{all}}$ be a closed convex set and let $J\colon \mathcal{Q} \to \mathbb{R}$ be a convex objective of the form $J(\pi) = \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi(x) \parallel \pi_0(x))] + R(\pi)$ with R convex (e.g., $R = 0$ and $\mathcal{Q} = \mathcal{C} \cap \Pi$ for the constrained projection (2), or $\mathcal{Q} = \Pi_{\mathrm{all}}$ and $R = \lambda\mathbb{E}[\mathsf{D}(\pi(x), \pi(\Phi(x)))]$ for the penalized problem (5)). Assume that F is μ-strongly convex with respect to the Euclidean norm, so that J is μ-strongly convex with respect to $\|\pi\|_{L_2} = (\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\|\pi(x)\|_2^2)^{1/2}$, and let $\widehat{\pi}_J = \operatorname{argmin}_{\pi\in\mathcal{Q}} J(\pi)$. Assume that $\pi \mapsto \mathbb{E}[\mathsf{B}_F(\pi^*(x) \parallel \pi(x))]$ is L_{π^*}-Lipschitz with respect to $\|\cdot\|_{L_2}$ on $\mathcal{Q}$. Then, for any $\widetilde{\pi} \in \mathcal{Q}$ with suboptimality $\eta = J(\widetilde{\pi}) - J(\widehat{\pi}_J) \ge 0$,*

$$|\mathbb{E}[\mathsf{B}_F(\pi^*(x) \parallel \widetilde{\pi}(x))] - \mathbb{E}[\mathsf{B}_F(\pi^*(x) \parallel \widehat{\pi}_J(x))]| \le L_{\pi^*}\sqrt{\frac{2\eta}{\mu}},$$

and hence $\mathrm{Improv}_{\pi^*}(\widetilde{\pi}) \ge \mathrm{Improv}_{\pi^*}(\widehat{\pi}_J) - L_{\pi^*}\sqrt{2\eta/\mu}$.

Proof. Since J is μ-strongly convex on the convex set $\mathcal{Q}$ and $\widehat{\pi}_J$ minimizes J over $\mathcal{Q}$, we have, for every $t \in (0,1)$, with $\pi_t = \widehat{\pi}_J + t(\widetilde{\pi} - \widehat{\pi}_J) \in \mathcal{Q}$,

$$J(\widehat{\pi}_J) \le J(\pi_t) \le (1-t)J(\widehat{\pi}_J) + tJ(\widetilde{\pi}) - \frac{\mu}{2}t(1-t)\|\widetilde{\pi} - \widehat{\pi}_J\|_{L_2}^2,$$

hence $\frac{\mu}{2}(1-t)\|\widetilde{\pi} - \widehat{\pi}_J\|_{L_2}^2 \le J(\widetilde{\pi}) - J(\widehat{\pi}_J) = \eta$ after dividing by t; letting $t \to 0$ gives $J(\widetilde{\pi}) \ge J(\widehat{\pi}_J) + \frac{\mu}{2}\|\widetilde{\pi} - \widehat{\pi}_J\|_{L_2}^2$, i.e., $\|\widetilde{\pi} - \widehat{\pi}_J\|_{L_2}^2 \le 2\eta/\mu$ (no differentiability of R or of J at $\widehat{\pi}_J$ is needed). The L_{π^*}-Lipschitz continuity of the error mapping then yields

$$|\mathbb{E}[\mathsf{B}_F(\pi^*(x) \parallel \widetilde{\pi}(x))] - \mathbb{E}[\mathsf{B}_F(\pi^*(x) \parallel \widehat{\pi}_J(x))]| \le L_{\pi^*}\|\widetilde{\pi} - \widehat{\pi}_J\|_{L_2} \le L_{\pi^*}\sqrt{\frac{2\eta}{\mu}}.$$

Since $\mathbb{E}[\mathsf{B}_F(\pi^*(x) \parallel \widetilde{\pi}(x))] - \mathbb{E}[\mathsf{B}_F(\pi^*(x) \parallel \widehat{\pi}_J(x))] = \mathrm{Improv}_{\pi^*}(\widehat{\pi}_J) - \mathrm{Improv}_{\pi^*}(\widetilde{\pi})$ (the term $\mathbb{E}[\mathsf{B}_F(\pi^*(x) \parallel \pi_0(x))]$ cancels in the difference), this yields the result on the improvement. The strong convexity of J follows from that of F since $\pi \mapsto \mathbb{E}[F(\pi(x))]$ is μ-strongly convex with respect to $\|\cdot\|_{L_2}$ and the remaining terms of J are convex. □

Two instances are of particular interest.

- *Finite samples.* For $\mathcal{Q} = \overline{\Pi}$, $R = 0$ and $\widetilde{\pi} = \widehat{\pi}_S$, Lemma D.4 gives $\eta \le 2\epsilon_m$, which yields the following corollary.
- *Parametric fine-tuning.* For the penalized objective J_λ of (5) and $\mathcal{Q} = \Pi_{\mathrm{all}}$, any model $\widetilde{\pi}$, e.g., the model returned by a (non-convex) fine-tuning procedure within a parametric class, satisfies the bound with η equal to its suboptimality in the training objective. The bound is agnostic to the origin of η: it jointly accounts for the fact that the exact projection may not lie in the parametric class (model-approximation error) and for the optimization error of the training procedure. In particular, the improvement guarantee degrades continuously, as $\sqrt{\eta}$, and the projection need not be computed exactly for the guarantee to be meaningful.

Corollary D.6 (Improvement guarantee under strong convexity for the empirical projection)**.** *Under the assumptions of Theorems D.2 and D.5 (with* $\Omega = \overline{\Pi}$ *and* $R = 0$*),* $\mathrm{Improv}_{\pi^*}(\widehat{\pi}_S) \geq \mathrm{Improv}_{\pi^*}(\widehat{\pi}) - \frac{2L_{\pi^*}}{\sqrt{\mu}}\sqrt{\epsilon_m}$.

Proof. By Lemma D.4, the excess population objective of $\widehat{\pi}_S$ satisfies $\eta \leq 2\epsilon_m$; plugging this into Theorem D.5 gives the result. □

Remark D.7 (On the Lipschitz constant L_{π^*})**.** *The Lipschitz assumption of Theorem D.5 concerns the* second *argument of the divergence. For a* C^2 *generator,* $\nabla_{\mathsf{q}} \mathsf{B}_F(\mathsf{p}^* \parallel \mathsf{q}) = -\nabla^2 F(\mathsf{q})(\mathsf{p}^* - \mathsf{q})$*, so the assumption holds whenever* $\nabla^2 F$ *is bounded on the range of the models considered (that is,* ∇F *is Lipschitz there), with* $L_{\pi^*} \leq \sup_x \sup_{\mathsf{q}} \|P_Z \nabla^2 F(\mathsf{q})(\pi^*(x) - \mathsf{q})\|_2$*, where* P_Z *is the orthogonal projection onto the tangent space* $Z = \{v : \sum_k v_k = 0\}$ *of the simplex (only directions in* Z *matter, since* $\pi(x) - \pi'(x) \in Z$ *for* $\pi, \pi' \in \Pi_{\mathrm{all}}$*). For the negative entropy, the computation is explicit and dimension-free. On* Π_{all}*,* $\mathsf{B}_F(\mathsf{p}^* \parallel \mathsf{q}) = \mathsf{D}_{\mathrm{KL}}(\mathsf{p}^* \parallel \mathsf{q})$ *and* $\nabla_{\mathsf{q}} \mathsf{D}_{\mathrm{KL}}(\mathsf{p}^* \parallel \mathsf{q}) = -\mathsf{p}^*/\mathsf{q}$ *(componentwise). Let* Ω *consist of models with* $\pi(y \mid x) \geq \alpha$ *for all* $y \in \operatorname{supp} \pi^*(x)$ *and all* x*, a convex set. For* $\pi, \pi' \in \Omega$ *and* $\pi_t = \pi' + t(\pi - \pi') \in \Omega$*, the mean value theorem and the Cauchy–Schwarz inequality give*

$$\begin{aligned} &|\mathbb{E}[\mathsf{D}_{\mathrm{KL}}(\pi^*(x) \parallel \pi(x))] - \mathbb{E}[\mathsf{D}_{\mathrm{KL}}(\pi^*(x) \parallel \pi'(x))]| \\ &\qquad \leq \sup_{t \in [0,1]} \mathbb{E}\left[\left\| \frac{\pi^*(x)}{\pi_t(x)} \right\|_2 \|\pi(x) - \pi'(x)\|_2\right] \leq \frac{1}{\alpha}\|\pi - \pi'\|_{L_2}, \end{aligned}$$

since $\|\pi^*(x)/\pi_t(x)\|_2 \leq \frac{1}{\alpha}\|\pi^*(x)\|_2 \leq \frac{1}{\alpha}\|\pi^*(x)\|_1 = \frac{1}{\alpha}$ *(the sum defining the ratio runs over* $\operatorname{supp} \pi^*(x)$ *only). Hence* $L_{\pi^*} \leq 1/\alpha$*, with no dependence on* $d = |\mathcal{Y}|$*, and only a lower bound on the model's probabilities* on the support of the ground truth *is needed, not a uniform lower bound over* $\mathcal{Y}$*; this is the relevant form of the condition in sequence-level settings, where* d *is exponentially large and most sequences have negligible probability. Finally, the negative entropy is* 1*-strongly convex with respect to* $\|\cdot\|_2$ *on* $\Delta(\mathcal{Y})$*, since* $\nabla^2 F(\mathsf{q}) = \mathrm{diag}(1/\mathsf{q}_i) \succeq I$ *(or since* $\|\cdot\|_2 \leq \|\cdot\|_1$ *and Pinsker's inequality), so* $\mu = 1$ *in Theorem D.5; note that the norm* $\|\cdot\|_{L_2}$ *of that theorem is built on the Euclidean norm on* $\mathbb{R}^d$*, whereas Theorem 4.3 allows an arbitrary norm.*

E Two-Step Coherence Projection

In Section 4, we studied the *direct projection* method, where the baseline π_0 is projected directly onto the set $(\mathcal{C}_{\mathrm{coh}} \cap \Pi)$ to obtain a coherent conditional distribution $\widehat{\pi} \in \Pi$. Here, we consider an alternative *two-step projection* (or *double-projection*) procedure: first, π_0 is projected onto $\mathcal{C}^{\dagger}_{\mathrm{coh}}$, yielding a coherent function $\overline{\pi}$; then, $\overline{\pi}$ is projected onto $(\mathcal{C}_{\mathrm{coh}} \cap \Pi)$ to produce the final solution $\widehat{\overline{\pi}}$.

For the definition of the two-step projection, we require Bregman generator functions F whose domain $\mathcal{K}$ is a convex set with interior containing the entire positive orthant $\mathbb{R}^d_{++}$. Most commonly used Bregman divergences satisfy this property: the squared Euclidean distance and, more generally, Mahalanobis functions are defined on all of $\mathbb{R}^d$; the negative entropy (which induces the unnormalized relative entropy and yields the D_{KL} divergence by restriction) is defined on the closed orthant $\mathbb{R}^d_{\geq 0}$; the Itakura–Saito generator and many α- and β-entropy functions are defined on the open orthant $\mathbb{R}^d_{++}$. Throughout this section, we assume generator functions F with a domain satisfying this property, and that $\pi_0(x) \in \mathbb{R}^d_{++}$ for all x (so that all divergences from $\pi_0(x)$ are finite, including for generators with $\operatorname{dom} F = \mathbb{R}^d_{++}$).

We begin by proving that, remarkably, the direct and two-step projections *coincide* for a very broad class of Bregman divergences (Section E.2). This class includes not only the squared Euclidean distance and D_{KL} divergence, but all common *separable* (e.g., Itakura-Saito, α-/β-divergences) and *quadratic* (e.g., Mahalanobis) divergences. The equivalence leverages a general property of Bregman projections onto affine sets, which applies directly to the unrestricted coherence set $\mathcal{C}^{\dagger\dagger}_{\mathrm{coh}}$, a linear subspace. A key part of our proof is showing that for this entire class of divergences, the first-step projection $\overline{\pi}$, which projects the non-negative π_0 onto the *linear subspace* $\mathcal{C}^{\dagger\dagger}_{\mathrm{coh}}$, is guaranteed to be non-negative (Section E.1). As a consequence, the two-step projection enjoys the same improvement guarantee as the direct one when they coincide. Nonetheless, we derive an alternative improvement guarantee for $\widehat{\overline{\pi}}$ (Section E.3), based on a closed-form expression for $\overline{\pi}$ (the Bregman centroid), which yields explicit guarantees in terms of the Hellinger distance for the relative entropy (Section E.4). As

before, we assume that Φ is an involution, although all results extend directly to the more general setting in which Φ admits arbitrary finite orbits.

E.1 Pythagorean theorem equality

Our comparison of the direct and two-step projections makes use of the following equality case of the Pythagorean theorem, which holds for any affine subspace, for which we give a general proof using Fenchel duality.

Theorem E.1 (Pythagorean theorem equality)**.** *Let $F:\mathcal{K} \to \mathbb{R}$ be of Legendre type, where $\mathcal{K} = \operatorname{dom} F \subseteq (\mathbb{R}^d)^{\mathcal{X}}$ is its natural domain and $\Omega = \operatorname{int}(\mathcal{K})$ (for instance, $F = \mathsf{F}$ as in Lemma B.6, with $\mathcal{K} = (\mathbb{R}^d)^{\mathcal{X}} = \Omega$ for quadratic generators and $\mathcal{K} = (\mathbb{R}^d_{\geq 0})^{\mathcal{X}}$, $\Omega = (\mathbb{R}^d_{++})^{\mathcal{X}}$ for steep generators on the orthant such as the negative entropy). Let $\mathcal{C} \subseteq (\mathbb{R}^d)^{\mathcal{X}}$ be an affine subspace with $\mathcal{C} \cap \Omega \neq \varnothing$, let $\pi_0 \in \Omega$, and let $\overline{\pi}$ be the B_F-projection of π_0 on $\mathcal{C} \cap \mathcal{K}$. Then, $\overline{\pi} \in \Omega$ and, for any $\pi \in \mathcal{C} \cap \mathcal{K}$,*

$$\mathsf{B}_F(\pi \parallel \pi_0) = \mathsf{B}_F(\pi \parallel \overline{\pi}) + \mathsf{B}_F(\overline{\pi} \parallel \pi_0).$$

It is important that $\mathcal{K}$ be the natural domain of F (i.e., that F not be artificially restricted to a smaller closed set): a quadratic generator is Legendre on $\mathbb{R}^d$ but not on $\mathbb{R}^d_{\geq 0}$, where its restriction has a nonempty subdifferential at boundary points. Gradients, conjugates and orthogonal complements below are taken with respect to the $\mathcal{D}_{\mathcal{X}}$-weighted inner product of Section B.5 (the conclusion does not depend on this choice).

Proof. Since $F(\pi_0)$ and $\langle \nabla F(\pi_0), \pi_0 \rangle$ are constant terms in the definition of the Bregman divergence, the Bregman projection $\overline{\pi}$ is a solution of the convex optimization problem $\min_{\pi \in \mathcal{C} \cap \mathcal{K}} \{F(\pi) - \langle \nabla F(\pi_0), \pi \rangle\}$. Define the indicator function $I_{\mathcal{C}}(\pi)$ of the affine set $\mathcal{C}$ by $I_{\mathcal{C}}(\pi) = 0$ if π is in $\mathcal{C}$, and $+\infty$ otherwise, and let $G(\pi) = F(\pi) - \langle \nabla F(\pi_0), \pi \rangle$, a closed proper convex function with $\operatorname{dom}(G) = \mathcal{K}$ (extended by $+\infty$ outside $\mathcal{K}$). Then, the problem can be rewritten equivalently as $\min_\pi \{G(\pi) + I_{\mathcal{C}}(\pi)\}$. Since $I_{\mathcal{C}}$ is convex and lower semi-continuous and $\operatorname{relint} \operatorname{dom}(G) \cap \operatorname{relint} \operatorname{dom}(I_{\mathcal{C}}) = \Omega \cap \mathcal{C} \neq \varnothing$, the Fenchel duality theorem [Rockafellar, 1997, Theorem 31.1] applies with strong duality and dual attainment:

$$\min_\pi \{G(\pi) + I_{\mathcal{C}}(\pi)\} = \max_{\theta \in (\mathbb{R}^d)^{\mathcal{X}}} \{-G^*(-\theta) - I^*_{\mathcal{C}}(\theta)\},$$

where the conjugate function G^* is given by $G^*(\phi) = F^*(\phi + \nabla F(\pi_0))$ via standard conjugate function calculus rules and the conjugate of the indicator function, $I^*_{\mathcal{C}}$, is the support function $\sigma_{\mathcal{C}}(\theta) = \sup_{\pi \in \mathcal{C}} \langle \theta, \pi \rangle$. Let θ^* denote the solution of the dual problem, then, we can write

$$G(\overline{\pi}) + I_{\mathcal{C}}(\overline{\pi}) = -G^*(-\theta^*) - I^*_{\mathcal{C}}(\theta^*) \Leftrightarrow G(\overline{\pi}) + G^*(-\theta^*) + I^*_{\mathcal{C}}(\theta^*) = 0, \tag{9}$$

since $\overline{\pi}$ is in $\mathcal{C}$ ($I_{\mathcal{C}}(\overline{\pi}) = 0$). Now, by the Fenchel-Young inequality, we have

$$\Delta_G = G(\overline{\pi}) + G^*(-\theta^*) - \langle \overline{\pi}, -\theta^* \rangle \geq 0, \quad \text{and} \quad \Delta_I = I_{\mathcal{C}}(\overline{\pi}) + I^*_{\mathcal{C}}(\theta^*) - \langle \overline{\pi}, \theta^* \rangle \geq 0.$$

Adding up these two inequalities and using $I_{\mathcal{C}}(\overline{\pi}) = 0$ yields:

$$\Delta_G + \Delta_I = [G(\overline{\pi}) + G^*(-\theta^*) + I^*_{\mathcal{C}}(\theta^*)] - [\langle \overline{\pi}, -\theta^* \rangle + \langle \overline{\pi}, \theta^* \rangle] = G(\overline{\pi}) + G^*(-\theta^*) + I^*_{\mathcal{C}}(\theta^*).$$

By (9), the right-hand side equals 0, thus $\Delta_G + \Delta_I = 0$. Since $\Delta_G \geq 0$ and $\Delta_I \geq 0$, this implies $\Delta_G = 0$ and $\Delta_I = 0$. Thus, equality holds in Fenchel-Young for both $(G, \overline{\pi}, -\theta^*)$ and $(I_{\mathcal{C}}, \overline{\pi}, \theta^*)$, which is equivalent to the subgradient conditions $-\theta^* \in \partial G(\overline{\pi})$ and $\theta^* \in \partial I_{\mathcal{C}}(\overline{\pi})$. In particular, $\partial G(\overline{\pi}) = \partial F(\overline{\pi}) - \nabla F(\pi_0) \neq \varnothing$. Since F is Legendre, its subdifferential is empty at every boundary point of $\mathcal{K}$ [Rockafellar, 1997, Theorem 26.1], so $\overline{\pi} \in \Omega$, where G is differentiable; hence $\partial G(\overline{\pi}) = \{\nabla G(\overline{\pi})\} = \{\nabla F(\overline{\pi}) - \nabla F(\pi_0)\}$. Therefore $\theta^* = \nabla F(\pi_0) - \nabla F(\overline{\pi})$. Writing $\mathcal{C} = v + U$ where U is the linear subspace corresponding to $\mathcal{C}$, we have $\partial I_{\mathcal{C}}(\overline{\pi}) = U^\perp$. Thus, $\theta^* = \nabla F(\pi_0) - \nabla F(\overline{\pi}) \in U^\perp$. For any $\pi \in \mathcal{C}$ the difference $\pi - \overline{\pi}$ is in U, therefore $\langle \nabla F(\pi_0) - \nabla F(\overline{\pi}), \pi - \overline{\pi} \rangle = 0$. This is the Bregman orthogonality condition. Expanding the two Bregman divergences and using this equality gives:

$$\begin{aligned}
\mathsf{B}_F(\pi \parallel \overline{\pi}) + \mathsf{B}_F(\overline{\pi} \parallel \pi_0) &= [F(\pi) - F(\overline{\pi}) - \langle \nabla F(\overline{\pi}), \pi - \overline{\pi} \rangle] \\
&\quad + [F(\overline{\pi}) - F(\pi_0) - \langle \nabla F(\pi_0), \overline{\pi} - \pi_0 \rangle] \\
&= F(\pi) - F(\pi_0) - \langle \nabla F(\pi_0), \pi - \overline{\pi} \rangle - \langle \nabla F(\pi_0), \overline{\pi} - \pi_0 \rangle \\
&\qquad \text{(by orthogonality)} \\
&= F(\pi) - F(\pi_0) - \langle \nabla F(\pi_0), \pi - \pi_0 \rangle = \mathsf{B}_F(\pi \parallel \pi_0).
\end{aligned}$$

This completes the proof. □

E.2 Comparison of direct and two-step projections

We cannot apply Theorem E.1 directly in our two-step projection setting as $\mathcal{C}^{\dagger}_{\mathrm{coh}}$, the projection set for the first stage, is a closed convex cone but not an affine set. However, the theorem can be applied to $\mathcal{C}^{\dagger\dagger}_{\mathrm{coh}}$, which is linear. To leverage this, we first show that for a broad family of Bregman divergences, the projection of a point in the simplex over $\mathcal{C}^{\dagger\dagger}_{\mathrm{coh}}$ is always in $\mathcal{C}^{\dagger}_{\mathrm{coh}}$. Thus, projection over $\mathcal{C}^{\dagger\dagger}_{\mathrm{coh}}$ becomes equivalent to projecting on $\mathcal{C}^{\dagger}_{\mathrm{coh}}$.

Theorem E.2 (Non-negativity of the unconstrained projection)**.** *Let $\pi_0 \in \Pi_{\mathrm{all}}$ with $\pi_0(x) \in \mathbb{R}^d_{++}$ for all x, and let B_F be a Bregman divergence generated by a Legendre function F whose domain is a convex set $\mathcal{K} \supseteq \mathbb{R}^d_{++}$ (e.g., $\mathbb{R}^d$, $\mathbb{R}^d_{\geq 0}$ or $\mathbb{R}^d_{++}$). If F is either (1) quadratic, $F(\mathsf{p}) = \frac{1}{2}\mathsf{p}^\top \mathsf{A}\mathsf{p} + b^\top \mathsf{p}$ for some $\mathsf{A} \succ 0$, or (2) separable, $F(\mathsf{p}) = \sum_{i=1}^d f(\mathsf{p}_i)$ for some one-dimensional convex function f, then the expected B_F-projection $\overline{\pi}$ of π_0 onto the linear subspace $\mathcal{C}^{\dagger\dagger}_{\mathrm{coh}}$ is in the non-negative cone $\mathcal{C}^{\dagger}_{\mathrm{coh}}$.*

Proof. The projection $\overline{\pi}$ is the function $\pi \in \mathcal{C}^{\dagger\dagger}_{\mathrm{coh}}$ that minimizes $L(\pi) = \mathbb{E}_{x\sim\mathcal{D}_X}[\mathsf{B}_F(\pi(x) \parallel \pi_0(x))]$. We must show that $\overline{\pi}(x) \geq 0$ for all x. The problem decouples by orbit. For any orbit $\{x, \Phi(x)\}$, the minimizer $\overline{\pi}(x) = \overline{\pi}(\Phi(x)) = c^*$ is the vector $c \in \mathbb{R}^d$ that minimizes the orbit's loss $L_x(c) = \mathbb{P}[x]\mathsf{B}_F(c \parallel \pi_0(x)) + \mathbb{P}[\Phi(x)]\mathsf{B}_F(c \parallel \pi_0(\Phi(x)))$. Let $Z = \mathbb{P}[x] + \mathbb{P}[\Phi(x)]$; minimizing $L_x(c)/Z$ is in the form required by Lemma B.3 with $p = 2$, $\lambda_1 = \mathbb{P}[x]/Z$, $\lambda_2 = \mathbb{P}[\Phi(x)]/Z$, $\mathsf{q}_1 = \pi_0(x) =: p$, and $\mathsf{q}_2 = \pi_0(\Phi(x)) =: q$. By Lemma B.3, the unique minimizer c^* (the right Bregman centroid) satisfies $\nabla F(c^*) = \lambda_1 \nabla F(p) + \lambda_2 \nabla F(q)$. By assumption, $p, q > 0$ (element-wise).

Quadratic generators. Then $\nabla F(\mathsf{p}) = \mathsf{A}\mathsf{p} + b$ and $\mathsf{A}c^* + b = \lambda_1(\mathsf{A}p + b) + \lambda_2(\mathsf{A}q + b)$, i.e., $\mathsf{A}c^* = \mathsf{A}(\lambda_1 p + \lambda_2 q)$. Since A is invertible, $c^* = \lambda_1 p + \lambda_2 q$, a weighted arithmetic mean of p and q, hence positive.

Separable generators. The problem decouples by component i, and $f'(c_i^*) = \lambda_1 f'(p_i) + \lambda_2 f'(q_i)$ with $p_i, q_i > 0$. Since f is Legendre on its one-dimensional domain, f' is continuous and strictly increasing on $\mathrm{int}(\mathrm{dom}\, f) \supseteq (0, \infty)$; the average $\lambda_1 f'(p_i) + \lambda_2 f'(q_i)$ lies between $f'(p_i)$ and $f'(q_i)$, hence in $f'([\min(p_i, q_i), \max(p_i, q_i)])$ by the intermediate value theorem, so that $c_i^* \in [\min(p_i, q_i), \max(p_i, q_i)] \subseteq (0, \infty)$. This covers both generators finite at 0 (e.g., the negative entropy) and generators with $\mathrm{dom}\, f = (0, \infty)$ (e.g., Itakura–Saito, $f(t) = -\log t$, or α-divergences with $\alpha \leq 0$), for which the positivity of π_0 is what makes the divergences finite. In all cases $c^* > 0$. Since $\overline{\pi} \in \mathcal{C}^{\dagger\dagger}_{\mathrm{coh}}$ and $\overline{\pi} \geq 0$, we have $\overline{\pi} \in \mathcal{C}^{\dagger\dagger}_{\mathrm{coh}} \cap \Pi^{\dagger}_{\mathrm{all}} = \mathcal{C}^{\dagger}_{\mathrm{coh}}$. □

The theorem can be generalized to a more abstract rule.

Theorem E.3 (Domain-closure of the unconstrained projection)**.** *Let $\mathcal{K}$ be a convex cone in $\mathbb{R}^d$ (e.g., $\mathbb{R}^d_{\geq 0}$ or $\mathbb{S}^d_{++}$), and let F be a Legendre function whose domain is $\mathcal{K}$. Let $\Pi^{\dagger}_{\mathrm{all}} = (\mathcal{K})^{\mathcal{X}}$ and $\pi_0 \in \Pi^{\dagger}_{\mathrm{all}}$. If B_F has the property that for any $p, q \in \mathcal{K}$ and any weights $\lambda_1, \lambda_2 > 0$ with $\lambda_1 + \lambda_2 = 1$, the right Bregman centroid c^* defined by $\nabla F(c^*) = \lambda_1 \nabla F(p) + \lambda_2 \nabla F(q)$ is also in $\mathcal{K}$, then the expected B_F-projection $\overline{\pi}$ of π_0 onto the linear subspace $\mathcal{C}^{\dagger\dagger}_{\mathrm{coh}}$ is in $\mathcal{C}^{\dagger}_{\mathrm{coh}} = \mathcal{C}^{\dagger\dagger}_{\mathrm{coh}} \cap \Pi^{\dagger}_{\mathrm{all}}$.*

Proof. By definition, $\overline{\pi} \in \mathcal{C}^{\dagger\dagger}_{\mathrm{coh}}$. As established in the proof of Theorem E.2, for any orbit $\{x, \Phi(x)\}$, $\overline{\pi}(x)$ is the right Bregman centroid of $p = \pi_0(x)$ and $q = \pi_0(\Phi(x))$ with weights $\mathbb{P}[x]$ and $\mathbb{P}[\Phi(x)]$. By assumption, $p, q \in \mathcal{K}$ and the centroid is in $\mathcal{K}$. Since this holds for all x, $\overline{\pi} \in (\mathcal{K})^{\mathcal{X}} = \Pi^{\dagger}_{\mathrm{all}}$. □

This theorem covers the vast majority of divergences used in machine learning. The quadratic class (squared Euclidean and Mahalanobis distances) is covered because the centroid is the weighted arithmetic mean. The separable class is also covered, including the D_{KL} divergence, Itakura-Saito divergence, and most α- and β-divergences: their centroids are component-wise generalized means (e.g., the geometric mean for KL) or are forced to 0 by the steep boundary of the generator. The principle also extends to other domains, such as the cone of positive-definite matrices ($\mathcal{K} = \mathbb{S}^d_{++}$): divergences like the log-determinant (Stein's loss) and von Neumann entropy satisfy this property, as their centroids, which are forms of harmonic or geometric matrix means, remain within the positive-definite cone. The primary requirement is the ability to use Lemma B.3, which requires F to be Legendre; Bregman divergences generated by functions that are not strictly convex are not covered.

Theorem E.4 (Equivalence of direct and two-step projections)**.** *Assume that the B_F-projection of π_0 onto the linear subspace $\mathcal{C}_{\mathrm{coh}}^{\dagger\dagger}$ (that is, onto $\mathcal{C}_{\mathrm{coh}}^{\dagger\dagger} \cap \mathrm{dom}(\mathsf{F})$) lies in $\mathcal{C}_{\mathrm{coh}}^{\dagger}$, as is the case under the hypotheses of Theorem E.2 (quadratic or separable generators) or of Theorem E.3 with $\mathcal{K} = \mathbb{R}^d_{\geq 0}$. Then, for any $\pi_0 \in \Pi_{\mathrm{all}}$ with $\pi_0(x) \in \mathbb{R}^d_{++}$ for all x, the direct and two-step projections of π_0 on $\Pi \cap \mathcal{C}_{\mathrm{coh}}$ coincide.*

Proof. Let $\overline{\pi} = \mathrm{Proj}_{\Pi \cap \mathcal{C}_{\mathrm{coh}}}(\pi_0)$ be the direct projection. The two-step projection is $\widehat{\overline{\pi}} = \mathrm{Proj}_{\Pi \cap \mathcal{C}_{\mathrm{coh}}}(\overline{\pi})$, where $\overline{\pi} = \mathrm{Proj}_{\mathcal{C}_{\mathrm{coh}}^{\dagger}}(\pi_0)$. By assumption, the projection onto the linear subspace $\mathcal{C}_{\mathrm{coh}}^{\dagger\dagger}$ lies in $\mathcal{C}_{\mathrm{coh}}^{\dagger} \subseteq \mathcal{C}_{\mathrm{coh}}^{\dagger\dagger}$, hence it is also the projection onto $\mathcal{C}_{\mathrm{coh}}^{\dagger}$: $\overline{\pi} = \mathrm{Proj}_{\mathcal{C}_{\mathrm{coh}}^{\dagger}}(\pi_0) = \mathrm{Proj}_{\mathcal{C}_{\mathrm{coh}}^{\dagger\dagger}}(\pi_0)$. Since $\overline{\pi}$ is the projection onto the *affine* set $\mathcal{C}_{\mathrm{coh}}^{\dagger\dagger}$, which meets $(\mathrm{int}\,\mathrm{dom}\,F)^{\mathcal{X}}$ (it contains the constant positive functions), Theorem E.1 gives, for all $\pi \in \Pi \cap \mathcal{C}_{\mathrm{coh}} \subseteq \mathcal{C}_{\mathrm{coh}}^{\dagger\dagger}$, $\mathsf{B}_F(\pi \parallel \pi_0) = \mathsf{B}_F(\pi \parallel \overline{\pi}) + \mathsf{B}_F(\overline{\pi} \parallel \pi_0)$. Since $\mathsf{B}_F(\overline{\pi} \parallel \pi_0)$ is a constant, minimizing the left-hand side over $\pi \in \Pi \cap \mathcal{C}_{\mathrm{coh}}$ is equivalent to minimizing the first term, so $\widehat{\overline{\pi}} = \overline{\pi}$. □

E.3 Improvement guarantees via two-step projection

Theorem E.5 (Coherent two-step Bregman projection improves the baseline)**.** *Let $\Phi : \mathcal{X} \to \mathcal{X}$ be an involution and let Π be a closed convex family of models $\pi : \mathcal{X} \to \Delta(\mathcal{Y})$. Assume that F is a Legendre function and, as throughout this section, that $\pi_0(x) \in \mathbb{R}^d_{++}$ for all x. Let $\overline{\pi} \in \Pi_{\mathrm{all}}^{\dagger}$ be the Bregman projection of π_0 onto $\mathcal{C}_{\mathrm{coh}}^{\dagger}$, $\overline{\pi} \in \mathrm{argmin}_{\pi \in \mathcal{C}_{\mathrm{coh}}^{\dagger}} \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi(x) \parallel \pi_0(x))]$, and $\widehat{\overline{\pi}}$ the Bregman projection of $\overline{\pi}$ onto $\mathcal{C}_{\mathrm{coh}} \cap \Pi$. Then, for any $\pi^* \in \mathcal{C}_{\mathrm{coh}} \cap \Pi$, $\widehat{\overline{\pi}}$ satisfies:*

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \parallel \widehat{\overline{\pi}}(x))] \leq \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \parallel \pi_0(x))] - \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\widehat{\overline{\pi}}(x) \parallel \overline{\pi}(x))] + \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\overline{\pi}(x) \parallel \pi_0(x))]\right].$$

Furthermore, if the right Bregman centroid of any two points of $\mathbb{R}^d_{++}$ lies in $\mathbb{R}^d_{++}$ (which holds under the hypotheses of Theorem E.2, or whenever $\mathrm{int}(\mathrm{dom}\,F) = \mathbb{R}^d_{++}$, since then ∇F is a bijection from $\mathbb{R}^d_{++}$ onto the convex set $\mathrm{int}(\mathrm{dom}\,F^*)$), the following upper bound holds:

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \parallel \widehat{\overline{\pi}}(x))] \leq \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \parallel \pi_0(x))] - \left[\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\widehat{\overline{\pi}}(x) \parallel \overline{\pi}(x))] + \delta\right],$$

where $\delta = \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[[\lambda(x)F^*(u_0) + (1-\lambda(x))F^*(u_1)] - F^*(\lambda(x)u_0 + (1-\lambda(x))u_1)] \geq 0$, with $\lambda(x) = \frac{\mathbb{P}[x]}{\mathbb{P}[x]+\mathbb{P}[\Phi(x)]}$, $u_0 = \nabla F(\pi_0(x))$ and $u_1 = \nabla F(\pi_0(\Phi(x)))$.

Proof. We first prove the first statement, which only uses the Pythagorean inequality. By definition, $\widehat{\overline{\pi}}$ is a Bregman projection of $\overline{\pi}$ onto $(\mathcal{C}_{\mathrm{coh}} \cap \Pi)$. By Lemma B.6, the expectation of the Bregman divergence B_F is also a Bregman divergence. Thus, since π^* is in $(\mathcal{C}_{\mathrm{coh}} \cap \Pi)$, by the Pythagorean theorem, we can write

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \parallel \widehat{\overline{\pi}}(x))] \leq \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \parallel \overline{\pi}(x))] - \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\widehat{\overline{\pi}}(x) \parallel \overline{\pi}(x))]. \tag{10}$$

By the Pythagorean theorem, considering the projection of π_0 onto the closed convex cone $\mathcal{C}_{\mathrm{coh}}^{\dagger} \ni \pi^*$, we can also write $\mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \parallel \overline{\pi}(x))] \leq \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \parallel \pi_0(x))] - \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\overline{\pi}(x) \parallel \pi_0(x))]$. Combining these inequalities yields the first statement. For the second statement, we first compute $\overline{\pi}$ in closed form. Since Φ is an involution, the expectation minimized to obtain $\overline{\pi}$ can be written as follows for a coherent π:

$$\mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi(x) \parallel \pi_0(x))] = \frac{1}{2}\sum_{x \in \mathcal{X}}\Big[\mathbb{P}[x]\,\mathsf{B}_F(\pi(x) \parallel \pi_0(x)) + \mathbb{P}[\Phi(x)]\,\mathsf{B}_F(\pi(x) \parallel \pi_0(\Phi(x)))\Big].$$

Since $\overline{\pi}$ is only restricted to be in $\mathcal{C}_{\mathrm{coh}}^{\dagger}$, the minimization of this objective decouples into the following problem for each orbit $\{x, \Phi(x)\}$: $\min_{\pi(x) \in \mathbb{R}^d_+} \lambda(x)\,\mathsf{B}_F(\pi(x) \parallel \pi_0(x)) + (1-\lambda(x))\,\mathsf{B}_F(\pi(x) \parallel \pi_0(\Phi(x)))$. By Lemma B.3, the minimizer of this objective over $\mathrm{int}(\mathrm{dom}\,F)$ is the right Bregman centroid $(\nabla F)^{-1}(\lambda(x)\nabla F(\pi_0(x)) + (1-\lambda(x))\nabla F(\pi_0(\Phi(x))))$, which lies in $\mathbb{R}^d_{++}$ by assumption (recall that $\pi_0(x) \in \mathbb{R}^d_{++}$ for all x); by Lemma B.3, it is in fact the unique

minimizer over all of $\operatorname{dom} F \supseteq \mathbb{R}^d_+ \cap \operatorname{dom} F$, hence the minimizer of the constrained problem. Thus, for any $x \in \mathcal{X}$,

$$\overline{\pi}(x) = (\nabla F)^{-1}(\lambda(x)\nabla F(\pi_0(x)) + (1-\lambda(x))\nabla F(\pi_0(\Phi(x)))).$$

We now analyze more precisely the first term on the right-hand side of (10). Assume first that $\pi^*(x) \in \operatorname{int}(\operatorname{dom} F)$ for all x. Using duality and Lemma B.2, we can write:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x) \,\|\, \overline{\pi}(x))] \\
&= \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_{F^*}(\nabla F(\overline{\pi}(x)) \,\|\, \nabla F(\pi^*(x)))] && (F \text{ Legendre and duality property}) \\
&= -\delta + \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\big[\lambda(x)\,\mathsf{B}_{F^*}(\nabla F(\pi_0(x)) \,\|\, \nabla F(\pi^*(x))) \\
&\quad + (1-\lambda(x))\,\mathsf{B}_{F^*}(\nabla F(\pi_0(\Phi(x))) \,\|\, \nabla F(\pi^*(x)))\big] && (\text{Lemma B.2}) \\
&= -\delta + \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\lambda(x)\,\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) + (1-\lambda(x))\,\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(\Phi(x)))]. && (\text{duality})
\end{aligned}$$

The resulting pointwise identity, $\lambda\mathsf{B}_F(\mathsf{p} \,\|\, \pi_0(x)) + (1-\lambda)\mathsf{B}_F(\mathsf{p} \,\|\, \pi_0(\Phi(x))) - \mathsf{B}_F(\mathsf{p} \,\|\, \overline{\pi}(x)) = \lambda F^*(u_0) + (1-\lambda)F^*(u_1) - F^*(\lambda u_0 + (1-\lambda)u_1)$ with $\lambda = \lambda(x)$, in fact holds for *every* $\mathsf{p} \in \operatorname{dom} F$, including boundary points: expanding the three divergences, the terms in $F(\mathsf{p})$ cancel, the terms linear in p cancel since $\nabla F(\overline{\pi}(x)) = \lambda u_0 + (1-\lambda)u_1$, and the remaining constants equal $\lambda[\langle u_0, \pi_0(x)\rangle - F(\pi_0(x))] + (1-\lambda)[\langle u_1, \pi_0(\Phi(x))\rangle - F(\pi_0(\Phi(x)))] - [\langle \nabla F(\overline{\pi}(x)), \overline{\pi}(x)\rangle - F(\overline{\pi}(x))]$, which is the right-hand side by the Fenchel–Young equality $F^*(\nabla F(\mathsf{q})) = \langle \nabla F(\mathsf{q}), \mathsf{q}\rangle - F(\mathsf{q})$. Hence the last display holds without the interiority assumption on π^*. We show that the last expectation coincides with $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))]$. Define $b(x) = \mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))$ and note that $1-\lambda(x) = \lambda(\Phi(x))$ and, by coherence of π^*, $\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(\Phi(x))) = b(\Phi(x))$. Then:

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[\lambda(x)\,b(x) + \lambda(\Phi(x))\,b(\Phi(x))] \\
&= \frac{1}{2}\sum_{x\in\mathcal{X}}\big[\mathbb{P}[x]\lambda(x)b(x) \\
&\quad + \mathbb{P}[x]\lambda(\Phi(x))b(\Phi(x)) + \mathbb{P}[\Phi(x)]\lambda(x)b(x) + \mathbb{P}[\Phi(x)]\lambda(\Phi(x))b(\Phi(x))\big] \\
&= \frac{1}{2}\sum_{x\in\mathcal{X}}[\mathbb{P}[x]\,b(x) + \mathbb{P}[\Phi(x)]\,b(\Phi(x))] = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}[b(x)].
\end{aligned}$$

Thus, $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x) \,\|\, \overline{\pi}(x))] = \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))] - \delta$. Substituting this expression in (10) yields the second statement and completes the proof. □

When $\widehat{\overline{\pi}} = \widehat{\pi}$, the theorem establishes an alternative guarantee for $\widehat{\pi}$. We briefly compare the improvement guarantees provided by our theorems for $\widehat{\pi}$ and $\widehat{\overline{\pi}}$: $\mathsf{B}_F(\pi^* \,\|\, \pi_0) - \mathsf{B}_F(\pi^* \,\|\, \widehat{\pi}) \geq \mathsf{B}_F(\widehat{\pi} \,\|\, \pi_0)$ and $\mathsf{B}_F(\pi^* \,\|\, \pi_0) - \mathsf{B}_F(\pi^* \,\|\, \widehat{\overline{\pi}}) \geq \mathsf{B}_F(\overline{\pi} \,\|\, \pi_0) + \mathsf{B}_F(\widehat{\overline{\pi}} \,\|\, \overline{\pi})$ (with expectations omitted). Since $\widehat{\pi}$ is in $\Pi \cap \mathcal{C}_{\mathrm{coh}}$ and thus also in $\mathcal{C}^\dagger_{\mathrm{coh}}$, by the Pythagorean theorem applied to the projection of π_0 onto $\mathcal{C}^\dagger_{\mathrm{coh}}$ and of $\overline{\pi}$ onto $\Pi \cap \mathcal{C}_{\mathrm{coh}}$, we have $\mathsf{B}_F(\widehat{\pi} \,\|\, \pi_0) \geq \mathsf{B}_F(\widehat{\pi} \,\|\, \overline{\pi}) + \mathsf{B}_F(\overline{\pi} \,\|\, \pi_0) \geq \mathsf{B}_F(\widehat{\pi} \,\|\, \widehat{\overline{\pi}}) + \mathsf{B}_F(\widehat{\overline{\pi}} \,\|\, \overline{\pi}) + \mathsf{B}_F(\overline{\pi} \,\|\, \pi_0)$. Thus, the improvement guarantee for $\widehat{\pi}$ is always at least as favorable as the one for $\widehat{\overline{\pi}}$.

E.4 Improvement guarantees for the relative entropy two-step projection

When F^* is μ-strongly convex, which implies the $1/\mu$-smoothness of F, the quantity δ in the second statement of the theorem admits the explicit lower bound $\frac{\mu}{2}\,\mathbb{E}[\lambda(x)(1-\lambda(x))\|\nabla F(\pi_0(x)) - \nabla F(\pi_0(\Phi(x)))\|^2]$. In the following, we present an explicit bound in the case of the relative entropy, $\mathsf{B}_F = \mathsf{D}_{\mathrm{KL}}$, in terms of the Hellinger distance. Recall that the *Hellinger distance* between two distributions p and q in Δ_d is defined by $\mathsf{D}_{\mathrm{Hell}}(\mathsf{p} \,\|\, \mathsf{q}) = \frac{1}{\sqrt{2}}\sqrt{\sum_{k=1}^d\big[\sqrt{\mathsf{p}_k} - \sqrt{\mathsf{q}_k}\big]^2}$, so that $\mathsf{D}^2_{\mathrm{Hell}}(\mathsf{p} \,\|\, \mathsf{q}) = 1 - \sum_{k=1}^d \sqrt{\mathsf{p}_k\mathsf{q}_k}$. Both quantities take values in $[0,1]$.

Corollary E.6. *Let F be the negative entropy, so that $\mathsf{B}_F = \mathsf{D}_{\mathrm{KL}}$ is the (unnormalized) relative entropy. Then, under the assumptions of Theorem E.5, the* log*-loss improvement of $\widehat{\overline{\pi}}$ over the baseline π_0 is at least*

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_\mathcal{X}}\big[2\min\{\lambda(x), 1-\lambda(x)\}\,\mathsf{D}^2_{\mathrm{Hell}}(\pi_0(x) \,\|\, \pi_0(\Phi(x)))\big],$$

where $\lambda(x) = \frac{\mathbb{P}[x]}{\mathbb{P}[x]+\mathbb{P}[\Phi(x)]}$.

Proof. Let $\mathsf{p} = \pi_0(x)$ and $\mathsf{q} = \pi_0(\Phi(x))$, and define $u_0 = \nabla F(\mathsf{p})$ and $u_1 = \nabla F(\mathsf{q})$ where $F\colon \mathbb{R}^d_+ \to \mathbb{R}$ is the negative entropy function defined by $F(\mathsf{r}) = \sum_k \mathsf{r}_k \log \mathsf{r}_k$, which is differentiable over $\mathbb{R}^d_{++}$ with $(\nabla F(\mathsf{r}))_k = \log \mathsf{r}_k + 1$. The conjugate function is $F^*(y) = \sum_{k=1}^d e^{y_k - 1}$ for all $y \in \mathbb{R}^d$, so that $F^*(\nabla F(\mathsf{r})) = \sum_k \mathsf{r}_k = 1$ for $\mathsf{r} \in \Delta_d$. Writing $\lambda = \lambda(x)$, the quantity whose expectation defines δ is therefore $-\Delta(\lambda)$, where

$$\Delta(\lambda) = F^*(\lambda u_0 + (1-\lambda)u_1) - 1 = \sum_{k=1}^d e^{\lambda(\log \mathsf{p}_k + 1) + (1-\lambda)(\log \mathsf{q}_k + 1) - 1} - 1 = \sum_{k=1}^d \mathsf{p}_k^{\lambda}\, \mathsf{q}_k^{1-\lambda} - 1.$$

The function $\lambda \mapsto \Delta(\lambda)$ is convex on $[0,1]$, as a sum of exponentials of affine functions of λ (each term $\mathsf{p}_k^{\lambda}\mathsf{q}_k^{1-\lambda} = \mathsf{q}_k e^{\lambda \log(\mathsf{p}_k/\mathsf{q}_k)}$ is convex in λ; terms with $\mathsf{p}_k\mathsf{q}_k = 0$ vanish on $(0,1)$ and do not affect convexity on $[0,1]$ after taking limits). Moreover, $\Delta(0) = \Delta(1) = 0$ and $\Delta(1/2) = \sum_k \sqrt{\mathsf{p}_k \mathsf{q}_k} - 1 = -\mathsf{D}^2_{\mathrm{Hell}}(\mathsf{p} \,\|\, \mathsf{q})$. By convexity, for $\lambda \in [0, 1/2]$, writing $\lambda = (1-2\lambda)\cdot 0 + 2\lambda \cdot \frac{1}{2}$,

$$\Delta(\lambda) \le (1-2\lambda)\Delta(0) + 2\lambda\Delta(1/2) = -2\lambda\, \mathsf{D}^2_{\mathrm{Hell}}(\mathsf{p} \,\|\, \mathsf{q}),$$

and symmetrically, for $\lambda \in [1/2, 1]$, $\Delta(\lambda) \le -2(1-\lambda)\mathsf{D}^2_{\mathrm{Hell}}(\mathsf{p} \,\|\, \mathsf{q})$. Hence, in all cases, $-\Delta(\lambda) \ge 2\min\{\lambda, 1-\lambda\}\, \mathsf{D}^2_{\mathrm{Hell}}(\mathsf{p} \,\|\, \mathsf{q})$, and the claim follows by taking expectations and using the second statement of Theorem E.5, together with the fact that the log-loss improvement equals the improvement in expected D_{KL}. □

Equivalently, $\Delta(\lambda) = \exp((\lambda-1)\mathsf{D}_\lambda(\mathsf{p} \,\|\, \mathsf{q})) - 1$, where D_λ is the Rényi divergence of order λ [van Erven and Harremoës, 2014]; alternative bounds in terms of the total variation distance can be derived from the inequality $\mathsf{D}_\lambda(\mathsf{p} \,\|\, \mathsf{q}) \ge \frac{\lambda}{2} V^2(\mathsf{p}, \mathsf{q})$, valid for all $\lambda \in (0, 1]$ [van Erven and Harremoës, 2014], where $V(\mathsf{p}, \mathsf{q})$ is the total variation distance.

F Proofs for Section 5

F.1 Proof of Theorem 5.1

Theorem 5.1 (Failure of minimax projections)**.** *There exist finite sets $\mathcal{X}, \mathcal{Y}$ with $|\mathcal{Y}| = 4$, a prompt distribution $\mathcal{D}_{\mathcal{X}}$, a closed convex model class $\Pi \subseteq \Pi_{\mathrm{all}}$, an involution Φ on $\mathcal{X}$ with coherence set $\mathcal{C}_{\mathrm{coh}}$, a source model $\pi_0 \notin \Pi$, and a convex family $\mathcal{F} = \mathrm{conv}\{F_1, F_2\}$ of two quadratic Legendre generators such that the unique minimax solution*

$$\pi_{mm} = \operatorname*{argmin}_{\pi \in \Pi \cap \mathcal{C}_{\mathrm{coh}}} \max_{F \in \mathcal{F}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{\mathcal{X}}} [\mathsf{B}_F(\pi(x) \,\|\, \pi_0(x))]$$

violates the improvement condition: there exist $F \in \mathcal{F}$ and a feasible reference $\pi^* \in \Pi \cap \mathcal{C}_{\mathrm{coh}}$ with $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_{mm}(x))] > \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))]$.

Proof. We give an explicit construction depending on a parameter $M > 0$, later chosen large enough.

Let $\mathcal{X} = \{x_1, x_2\}$ with $\mathcal{D}_{\mathcal{X}}$ uniform, let $\mathcal{Y} = \{1,2,3,4\}$, and let Φ be the involution swapping x_1 and x_2, so that $\mathcal{C}_{\mathrm{coh}} = \{\pi : \pi(x_1) = \pi(x_2)\}$. We identify a coherent model with the common value $q = \pi(x_1) = \pi(x_2) \in \Delta_4$. Define

$$\Pi = \left\{(q,q) : q \in \Delta_4,\ q_1 + q_2 \le \tfrac{1}{2}\right\},$$

which is closed and convex, and $\Pi \cap \mathcal{C}_{\mathrm{coh}} = \Pi$. Define π_0 by $\pi_0(x_1) = p_1 = (1,0,0,0)$ and $\pi_0(x_2) = p_2 = (0,1,0,0)$; π_0 lies outside Π and outside $\mathcal{C}_{\mathrm{coh}}$. For a positive definite matrix A, write $\|v\|^2_{\mathsf{A}} = v^\top \mathsf{A} v$. Let $\mathsf{A}_1 = \mathrm{diag}(1,1,M,1)$ and $\mathsf{A}_2 = \mathrm{diag}(1,1,1,M)$, and define the quadratic Legendre generators $F_k(\mathsf{p}) = \frac{1}{2}\mathsf{p}^\top \mathsf{A}_k \mathsf{p}$ on $\mathbb{R}^4$, so that $\mathsf{B}_{F_k}(\mathsf{p} \,\|\, \mathsf{q}) = \frac{1}{2}\|\mathsf{p} - \mathsf{q}\|^2_{\mathsf{A}_k}$, and let $\mathcal{F} = \mathrm{conv}(\{F_1, F_2\})$, a convex family of (quadratic) Legendre generators.

Computation of the minimax solution. For a coherent $\pi = (q,q)$, define

$$f_k(q) = \mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}} [\mathsf{B}_{F_k}(\pi(x) \,\|\, \pi_0(x))] = \tfrac{1}{4}\left[\|q - p_1\|^2_{\mathsf{A}_k} + \|q - p_2\|^2_{\mathsf{A}_k}\right].$$

Since $\mathsf{B}_{\alpha F_1+(1-\alpha)F_2} = \alpha\mathsf{B}_{F_1}+(1-\alpha)\mathsf{B}_{F_2}$, we have $\max_{F\in\mathcal{F}} \mathbb{E}[\mathsf{B}_F(\pi \parallel \pi_0)] = \max(f_1(q), f_2(q)) =: g(q)$. Each f_k is a strictly convex quadratic, so g is strictly convex and admits a unique minimizer q_{mm} over the compact convex set $Q = \{q \in \Delta_4 : q_1 + q_2 \le 1/2\}$. Let τ_{34} be the permutation of coordinates swapping 3 and 4, and τ_{12} the one swapping 1 and 2. Since τ_{34} fixes p_1, p_2 and exchanges A_1 and A_2, we have $f_1(\tau_{34}q) = f_2(q)$, hence $g(\tau_{34}q) = g(q)$; since τ_{12} exchanges p_1 and p_2 and commutes with $\mathsf{A}_1, \mathsf{A}_2$, we have $f_k(\tau_{12}q) = f_k(q)$, hence $g(\tau_{12}q) = g(q)$. The set Q is invariant under both permutations. By uniqueness of the minimizer, q_{mm} is invariant under τ_{12} and τ_{34}, hence of the form $q_{mm} = (t, t, s, s)$ with $2t + 2s = 1$ and $2t \le 1/2$, i.e., $s \in [1/4, 1/2]$. On this segment, with $t = 1/2 - s$,

$$\begin{aligned} f_1(q) &= \tfrac{1}{4}\big[(t-1)^2 + t^2 + Ms^2 + s^2 + t^2 + (t-1)^2 + Ms^2 + s^2\big] \\ &= \tfrac{1}{2}\big[(t-1)^2 + t^2 + (M+1)s^2\big] = \tfrac{1}{2}\big[\tfrac{1}{2} + (M+3)s^2\big], \end{aligned}$$

which is increasing in s. Thus, $s = 1/4$ and $q_{mm} = (\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$, i.e., $\pi_{mm} = (q_{mm}, q_{mm})$.

Violation of the improvement condition. Choose the feasible reference $\pi^* = (q^*, q^*) \in \Pi \cap \mathcal{C}_{\mathrm{coh}}$ with $q^* = (\frac{1}{4}, \frac{1}{4}, \frac{1}{2}, 0)$. Since both π^* and π_{mm} are coherent, with $q^* - q_{mm} = (0, 0, \frac{1}{4}, -\frac{1}{4})$,

$$\mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_{F_2}(\pi^*(x) \parallel \pi_{mm}(x))] = \tfrac{1}{2}\|q^* - q_{mm}\|^2_{\mathsf{A}_2} = \tfrac{1}{2}\big(\tfrac{1}{16} + \tfrac{M}{16}\big) = \tfrac{1+M}{32}.$$

On the other hand, $q^* - p_1 = (-\frac{3}{4}, \frac{1}{4}, \frac{1}{2}, 0)$ and $q^* - p_2 = (\frac{1}{4}, -\frac{3}{4}, \frac{1}{2}, 0)$ have vanishing fourth coordinate, so

$$\begin{aligned} &\mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_{F_2}(\pi^*(x) \parallel \pi_0(x))] \\ &\qquad = \tfrac{1}{4}\big[\|q^* - p_1\|^2_{\mathsf{A}_2} + \|q^* - p_2\|^2_{\mathsf{A}_2}\big] = \tfrac{1}{4}\big[\tfrac{9}{16} + \tfrac{1}{16} + \tfrac{1}{4} + \tfrac{1}{16} + \tfrac{9}{16} + \tfrac{1}{4}\big] = \tfrac{7}{16} = \tfrac{14}{32}. \end{aligned}$$

Therefore, for any $M > 13$ (e.g., $M = 16$), $\mathbb{E}[\mathsf{B}_{F_2}(\pi^* \parallel \pi_{mm})] = \frac{1+M}{32} > \frac{14}{32} = \mathbb{E}[\mathsf{B}_{F_2}(\pi^* \parallel \pi_0)]$: the minimax solution π_{mm} does not satisfy the improvement inequality for the divergence $F_2 \in \mathcal{F}$ and the feasible reference π^*, which completes the proof. □

The mechanism behind the counterexample is transparent. The F_2-projection of π_0 onto $\Pi \cap \mathcal{C}_{\mathrm{coh}}$ is obtained by first averaging p_1 and p_2 into $\bar{p} = (\frac{1}{2}, \frac{1}{2}, 0, 0)$ (the projection onto $\mathcal{C}^{\dagger}_{\mathrm{coh}}$, Section E) and then moving the excess mass $\frac{1}{2}$ from coordinates $1, 2$ to coordinates $3, 4$ at minimal A_2-cost, i.e., almost entirely to coordinate 3: $\widehat{q}_2 \approx q^*$. Symmetrically, $\widehat{q}_1 \approx (\frac{1}{4}, \frac{1}{4}, 0, \frac{1}{2})$. The minimax solution hedges between the two and places mass $\frac{1}{4}$ on coordinate 4, which is expensive in the F_2 geometry; a reference at $\widehat{q}_2$ is then farther from π_{mm} than from π_0, which agrees with q^* on coordinate 4. The same computation shows that for $M > 14$ *no* point of $\Pi \cap \mathcal{C}_{\mathrm{coh}}$ improves over π_0 for both F_1 and F_2 simultaneously: improvement for F_2 against the reference q^* forces $Mq_4^2 \le 7/8$, improvement for F_1 against the symmetric reference $(\frac{1}{4}, \frac{1}{4}, 0, \frac{1}{2})$ forces $Mq_3^2 \le 7/8$, and both are incompatible with $q_3 + q_4 \ge 1/2$ when $2\sqrt{7/(8M)} < 1/2$.

F.2 Proof of Theorem 5.2

Theorem 5.2 (Improvement via orbitwise averaging for jointly convex divergences)**.** *Let Π be a closed convex family of models, Φ an involution, $\pi_0 \in \Pi$ a baseline and $\pi^* \in \Pi \cap \mathcal{C}_{\mathrm{coh}}$ a coherent reference. Let $\mathcal{F}$ be a family of Legendre generators such that, for every $F \in \mathcal{F}$ and every $\mathsf{p} \in \Delta(\mathcal{Y})$, the map $\mathsf{q} \mapsto \mathsf{B}_F(\mathsf{p} \parallel \mathsf{q})$ is convex on $\Delta(\mathcal{Y})$; this holds in particular when B_F is jointly convex, as for the squared Euclidean, Mahalanobis and relative entropy divergences. Define the orbitwise averaged (coherent) model*

$$\widehat{\pi}_{\mathrm{orbit}}(x) = \frac{\mathbb{P}[x]\pi_0(x) + \mathbb{P}[\Phi(x)]\pi_0(\Phi(x))}{\mathbb{P}[x] + \mathbb{P}[\Phi(x)]} = \lambda(x)\pi_0(x) + (1 - \lambda(x))\pi_0(\Phi(x)).$$

Then, for all $F \in \mathcal{F}$, $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \parallel \widehat{\pi}_{\mathrm{orbit}}(x))] \le \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \parallel \pi_0(x))]$. Moreover, $\widehat{\pi}_{\mathrm{orbit}} \in \Pi$ whenever Π is closed under the swap $\pi \mapsto \pi \circ \Phi$ and $\mathcal{D}_{\mathcal{X}}$ is Φ-invariant (so that $\lambda \equiv 1/2$).

Proof. Partition $\mathcal{X}$ into orbits under Φ, i.e., sets $\{x, \Phi(x)\}$ (of size one or two). Fix an orbit, with weights $w_x = \mathbb{P}[x]$ and $w_{\Phi(x)} = \mathbb{P}[\Phi(x)]$. Since π^* is coherent, $\pi^*(x) = \pi^*(\Phi(x))$ on the orbit. By

convexity of $\mathsf{B}_F(\pi^*(x) \,\|\, \cdot)$, we have

$$\mathsf{B}_F\Big(\pi^*(x) \,\Big\|\, \widehat{\pi}_{\mathrm{orbit}}(x)\Big) = \mathsf{B}_F\left(\frac{w_x + w_{\Phi(x)}}{w_x + w_{\Phi(x)}}\pi^*(x) \,\middle\|\, \frac{w_x\pi_0(x) + w_{\Phi(x)}\pi_0(\Phi(x))}{w_x + w_{\Phi(x)}}\right)$$
$$\leq \frac{w_x}{w_x + w_{\Phi(x)}}\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) + \frac{w_{\Phi(x)}}{w_x + w_{\Phi(x)}}\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(\Phi(x))).$$

Multiplying both sides by $(w_x + w_{\Phi(x)})$, using the coherence of π^* and $\widehat{\pi}_{\mathrm{orbit}}$, and summing over all orbits yields

$$\begin{aligned}
&\mathop{\mathbb{E}}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_{\mathrm{orbit}}(x))] \\
&= \sum_{\text{orbits}} \big(w_x + w_{\Phi(x)}\big)\,\mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_{\mathrm{orbit}}(x)) \\
&\leq \sum_{\text{orbits}} \big[w_x\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) + w_{\Phi(x)}\mathsf{B}_F(\pi^*(\Phi(x)) \,\|\, \pi_0(\Phi(x)))\big],
\end{aligned}$$

and the right-hand side is $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))]$ (for a fixed point $x = \Phi(x)$, the orbit contributes $w_x\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x))$ on both sides). This proves the inequality. Finally, if $\mathcal{D}_{\mathcal{X}}$ is Φ-invariant, then $\lambda(x) = 1/2$ for all x, so $\widehat{\pi}_{\mathrm{orbit}} = \frac{1}{2}(\pi_0 + \pi_0 \circ \Phi)$ is a convex combination of $\pi_0 \in \Pi$ and $\pi_0 \circ \Phi \in \Pi$ (by the swap-closure assumption), hence belongs to the convex set Π. □

We note that when $\mathcal{D}_{\mathcal{X}}$ is not Φ-invariant, the weights $\lambda(x)$ vary with x, and the swap-closure and convexity of Π alone do not guarantee $\widehat{\pi}_{\mathrm{orbit}} \in \Pi$; feasibility must then be checked directly (it holds, e.g., when Π is a product of convex sets over $\mathcal{X}$, or when $\Pi = \Pi_{\mathrm{all}}$).

F.3 Universal improvement for the Mahalanobis family

Corollary F.1 (Improvement for the Mahalanobis family)**.** *Let Π, Φ, π_0, and π^* be as in Theorem 5.2. Let $\mathcal{M}$ be a set of symmetric positive semidefinite matrices and, for each $\mathsf{M} \in \mathcal{M}$, define the quadratic (Mahalanobis) Bregman divergence $\mathsf{B}_{\mathsf{M}}(\mathsf{p} \,\|\, \mathsf{q}) = \frac{1}{2}(\mathsf{p}-\mathsf{q})^\top\mathsf{M}(\mathsf{p}-\mathsf{q})$. Then, the orbitwise averaged coherent model $\widehat{\pi}_{\mathrm{orbit}}$ satisfies $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathsf{B}_{\mathsf{M}}(\pi^*(x) \,\|\, \widehat{\pi}_{\mathrm{orbit}}(x))\right] \leq \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}\left[\mathsf{B}_{\mathsf{M}}(\pi^*(x) \,\|\, \pi_0(x))\right]$ for all $\mathsf{M} \in \mathcal{M}$.*

Proof. For each fixed $\mathsf{M} \succeq 0$, the Mahalanobis divergence $\mathsf{B}_{\mathsf{M}}(\mathsf{p}\|\mathsf{q}) = \frac{1}{2}(\mathsf{p}-\mathsf{q})^\top\mathsf{M}(\mathsf{p}-\mathsf{q})$ is jointly convex in (p,q), as the composition of a convex quadratic form with the linear map $(\mathsf{p},\mathsf{q}) \mapsto \mathsf{p}-\mathsf{q}$; in particular it is convex in q. The result then follows directly from the inequality of Theorem 5.2. □

F.4 Proof of Theorem 5.3

Theorem 5.3 (Universal improvement over quadratic divergences forces orbit averaging)**.** *Let Φ be an involution and let $\mathcal{F}$ contain the quadratic generators $F_{\mathsf{M}}(\mathsf{p}) = \frac{1}{2}\mathsf{p}^\top\mathsf{M}\mathsf{p}$ for all positive definite M. Let $\widehat{\pi} \in \mathcal{C}_{\mathrm{coh}}$ satisfy (6), and let $u(x) = \widehat{\pi}(x) - \widehat{\pi}_{\mathrm{orbit}}(x)$ and $\delta_0(x) = \pi_0(x) - \pi_0(\Phi(x))$. Then:*

(a) $\mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[u(x)u(x)^\top] \preceq \mathbb{E}_{x\sim\mathcal{D}_{\mathcal{X}}}[\lambda(x)(1-\lambda(x))\delta_0(x)\delta_0(x)^\top]$*: the deviation from $\widehat{\pi}_{\mathrm{orbit}}$ is dominated by the incoherence of π_0.*

(b) *If $|\mathcal{Y}| \geq 3$ and $\mathcal{X}$ consists of a single orbit $\{x, \Phi(x)\}$, then $\widehat{\pi} = \widehat{\pi}_{\mathrm{orbit}}$. In particular, if $\widehat{\pi}_{\mathrm{orbit}} \notin \Pi$, no model in $\Pi \cap \mathcal{C}_{\mathrm{coh}}$ has the universal improvement property.*

Proof. Throughout, let $\mathsf{M} \succ 0$ and write $\mathsf{B}_{\mathsf{M}}(\mathsf{p} \,\|\, \mathsf{q}) = \frac{1}{2}(\mathsf{p}-\mathsf{q})^\top\mathsf{M}(\mathsf{p}-\mathsf{q})$ for the divergence of F_{M}. Consider an orbit $O = \{x, \Phi(x)\}$ with weights $w_1 = \mathbb{P}[x]$, $w_2 = \mathbb{P}[\Phi(x)]$, $W_O = w_1 + w_2$ and $\lambda = \lambda(x) = w_1/W_O$; let $a = \pi_0(x)$, $b = \pi_0(\Phi(x))$, $\bar{p}_O = \widehat{\pi}_{\mathrm{orbit}}(x) = \lambda a + (1-\lambda)b$, $q_O = \widehat{\pi}(x)$ (which is constant on the orbit by coherence of $\widehat{\pi}$), $u_O = q_O - \bar{p}_O$, and $p_O = \pi^*(x)$ for a coherent π^*. For a fixed point $x = \Phi(x)$, the orbit has a single element and we set $W_O = w_1$, $\lambda = 1$, $b = a$.

Step 1: an exact decomposition. For quadratic divergences, the variance decomposition

$$w_1\mathsf{B}_{\mathsf{M}}(p \,\|\, a) + w_2\mathsf{B}_{\mathsf{M}}(p \,\|\, b) = W_O[\mathsf{B}_{\mathsf{M}}(p \,\|\, \bar{p}_O) + \lambda(1-\lambda)\mathsf{B}_{\mathsf{M}}(a \,\|\, b)]$$

holds for every p (expand the squares and use $a - \bar{p}_O = (1-\lambda)(a-b)$, $b - \bar{p}_O = -\lambda(a-b)$). Moreover, $\mathsf{B}_{\mathsf{M}}(p \,\|\, q_O) - \mathsf{B}_{\mathsf{M}}(p \,\|\, \bar{p}_O) = \frac{1}{2}\|p - \bar{p}_O - u_O\|_{\mathsf{M}}^2 - \frac{1}{2}\|p - \bar{p}_O\|_{\mathsf{M}}^2 = -(p-\bar{p}_O)^\top\mathsf{M}u_O + \frac{1}{2}u_O^\top\mathsf{M}u_O$.

Summing over orbits, the universal improvement inequality (6) for $F = F_{\mathsf{M}}$ and the coherent π^* with values $(p_O)_O$ reads

$$\sum_O W_O\big[\tfrac{1}{2}u_O^\top \mathsf{M} u_O - (p_O - \bar{p}_O)^\top \mathsf{M} u_O\big] \le \sum_O W_O\, \lambda_O(1-\lambda_O)\, \tfrac{1}{2}(a_O - b_O)^\top \mathsf{M}(a_O - b_O), \qquad (11)$$

and this must hold for every $\mathsf{M} \succ 0$ and every choice of $p_O \in \Delta(\mathcal{Y})$ for each orbit. Since both sides are linear in M and continuous, (11) extends to all positive semidefinite M.

Step 2: proof of (a). Choose $p_O = \bar{p}_O \in \Delta(\mathcal{Y})$ for every orbit (a convex combination of $a_O, b_O \in \Delta(\mathcal{Y})$). Then (11) becomes $\operatorname{tr}\big(\mathsf{M}\sum_O W_O u_O u_O^\top\big) \le \operatorname{tr}\big(\mathsf{M}\sum_O W_O \lambda_O(1-\lambda_O)(a_O-b_O)(a_O-b_O)^\top\big)$ for all $\mathsf{M} \succeq 0$, which is equivalent to the Loewner inequality $\sum_O W_O u_O u_O^\top \preceq \sum_O W_O \lambda_O(1-\lambda_O)(a_O-b_O)(a_O-b_O)^\top$. Since the functions $x \mapsto u(x)u(x)^\top$ and $x \mapsto \lambda(x)(1-\lambda(x))\delta_0(x)\delta_0(x)^\top$ are constant on orbits (note that $\lambda(\Phi(x)) = 1-\lambda(x)$ and $\delta_0(\Phi(x)) = -\delta_0(x)$), the two sums are exactly $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[u(x)u(x)^\top]$ and $\mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\lambda(x)(1-\lambda(x))\delta_0(x)\delta_0(x)^\top]$, which proves (a).

Step 3: proof of (b). Assume now that $\mathcal{X} = O$ is a single orbit with $a \neq b$ (if $a = b$, then (a) gives $W_O u_O u_O^\top \preceq 0$, hence $u_O = 0$), and that $|\mathcal{Y}| \ge 3$. Write $u = u_O$, $\bar{p} = \bar{p}_O$ and $\delta = a - b \neq 0$. By (a), $uu^\top \preceq \lambda(1-\lambda)\delta\delta^\top$, so the range of $uu^\top$ is contained in that of $\delta\delta^\top$: $u = s\delta$ for some $s \in \mathbb{R}$. Suppose, for the sake of contradiction, that $s \neq 0$. Let $Z = \{v \in \mathbb{R}^d : \sum_k v_k = 0\}$ be the zero-sum subspace; $\dim Z = d-1 \ge 2$, $\delta \in Z$, and $\Delta(\mathcal{Y}) - \bar{p}$ spans Z (its span contains all differences $e_k - e_j = (e_k - \bar{p}) - (e_j - \bar{p})$). Hence there exists $p \in \Delta(\mathcal{Y})$ such that $d_p = p - \bar{p} \notin \operatorname{span}(\delta)$. Then there exists $v_0 \in Z$ with $v_0 \perp \delta$ and $c_0 = v_0^\top d_p \neq 0$: otherwise d_p would be orthogonal to the orthogonal complement of δ in Z, i.e., $d_p \in \operatorname{span}(\delta)$. For $\epsilon \neq 0$, set $v_\epsilon = v_0 + \epsilon\,\delta/\|\delta\|_2^2$ and $\mathsf{M}_\epsilon = v_\epsilon v_\epsilon^\top \succeq 0$. Then $v_\epsilon^\top \delta = \epsilon$, $v_\epsilon^\top u = s\epsilon$ and $v_\epsilon^\top d_p = c_0 + \epsilon c_1$ with $c_1 = \delta^\top d_p/\|\delta\|_2^2$. Applying (11) with $\mathsf{M} = \mathsf{M}_\epsilon$ and $p_O = p$ gives

$$W_O\big[\tfrac{1}{2}s^2\epsilon^2 - (c_0 + \epsilon c_1)s\epsilon\big] \le W_O \lambda(1-\lambda)\tfrac{1}{2}\epsilon^2.$$

Dividing by $W_O\epsilon^2 > 0$ yields $\frac{1}{2}s^2 - \frac{c_0 s}{\epsilon} - c_1 s \le \frac{1}{2}\lambda(1-\lambda)$ for every $\epsilon \neq 0$. Choosing ϵ with the sign of $-c_0 s$ and letting $|\epsilon| \to 0$, the left-hand side tends to $+\infty$, a contradiction. Hence $s = 0$, i.e., $u = 0$ and $\widehat{\pi} = \widehat{\pi}_{\mathrm{orbit}}$. The last claim follows: if $\widehat{\pi}_{\mathrm{orbit}} \notin \Pi$, no $\widehat{\pi} \in \Pi \cap \mathcal{C}_{\mathrm{coh}}$ can satisfy (6). □

Remark F.2 (Several orbits and binary outputs)**.** *Two features of the statement are sharp. First, with several orbits, the exact conclusion $\widehat{\pi} = \widehat{\pi}_{\mathrm{orbit}}$ does not follow from (6) in general: if the matrix $S = \mathbb{E}[\lambda(1-\lambda)\delta_0\delta_0^\top]$ is positive definite on the zero-sum subspace Z (which happens as soon as the incoherence vectors δ_0 of the orbits span Z), then any coherent $\widehat{\pi}$ that coincides with $\widehat{\pi}_{\mathrm{orbit}}$ except on one orbit O_1, where u_{O_1} is small enough, satisfies (11) for all $\mathsf{M} \succeq 0$ and all (p_O): indeed, the left-hand side is $\operatorname{tr}(\mathsf{M}T)$ with $T = W_{O_1}(\frac{1}{2}uu^\top - \frac{1}{2}(du^\top + ud^\top))$, $\|d\|_2 \le 2$, whose operator norm on Z is at most $W_{O_1}(\frac{1}{2}\|u\|_2^2 + 2\|u\|_2)$, while the right-hand side is $\operatorname{tr}(\mathsf{M}S)$, so the inequality holds whenever $W_{O_1}(\frac{1}{2}\|u\|_2^2 + 2\|u\|_2) \le \lambda_{\min}^Z(S)$. The incoherence of the other orbits provides slack that absorbs a small deviation; part (a) is precisely the quantitative statement that survives. Second, for binary outputs ($|\mathcal{Y}| = 2$), part (b) fails: with a single orbit, $a = (1,0)$, $b = (0,1)$ and $\lambda = 1/2$, one checks from (11) that $q = \bar{p} - s(a-b)$ has the universal improvement property over all quadratic generators iff $|s| + s^2 \le 1/4$, since all vectors $p - \bar{p}$ are then collinear with $a - b$ and (11) reduces to $\frac{1}{2}s^2 + \alpha s \le \frac{1}{8}$ for all $\alpha \in [-\frac{1}{2}, \frac{1}{2}]$.*

F.5 Proof of Theorem 5.4

Theorem 5.4 (Impossibility without convexity in the second argument)**.** *Let $|\mathcal{Y}| \ge 3$ and let $\mathcal{F}$ be a convex family of Legendre generators containing all quadratic generators F_{M}, $\mathsf{M} \succ 0$, and at least one generator F whose Bregman divergence is not convex in its second argument on $\Delta(\mathcal{Y})$, i.e., for which $\mathsf{q} \mapsto \mathsf{B}_F(\mathsf{p}^* \parallel \mathsf{q})$ is not convex on $\operatorname{relint}\Delta(\mathcal{Y})$ for some $\mathsf{p}^* \in \Delta(\mathcal{Y})$. Then, there exists an instance, with $\mathcal{X}$ a single orbit $\{x, \Phi(x)\}$, a distribution $\mathcal{D}_\mathcal{X}$ and a baseline π_0, on which* no *coherent model $\widehat{\pi} \in \mathcal{C}_{\mathrm{coh}}$ has the universal improvement property (6) over $\mathcal{F}$.*

Proof. By assumption, $\mathsf{q} \mapsto \mathsf{B}_F(\mathsf{p}^* \parallel \mathsf{q})$ is not convex on $\operatorname{relint}\Delta(\mathcal{Y})$ for some $\mathsf{p}^* \in \Delta(\mathcal{Y})$, that is, there exist points $\mathsf{q}_1, \mathsf{q}_2 \in \operatorname{relint}\Delta(\mathcal{Y})$ and a weight $\lambda \in (0,1)$ such that Jensen's inequality is reversed:

$$\mathsf{B}_F(\mathsf{p}^* \parallel \lambda\mathsf{q}_1 + (1-\lambda)\mathsf{q}_2) > \lambda\mathsf{B}_F(\mathsf{p}^* \parallel \mathsf{q}_1) + (1-\lambda)\mathsf{B}_F(\mathsf{p}^* \parallel \mathsf{q}_2). \qquad (12)$$

We build the instance. Let $\mathcal{X} = \{x, \Phi(x)\}$ be a single orbit with $x \neq \Phi(x)$, let $\mathcal{D}_\mathcal{X}$ place masses λ on x and $1-\lambda$ on $\Phi(x)$, and define the baseline by $\pi_0(x) = \mathsf{q}_1$ and $\pi_0(\Phi(x)) = \mathsf{q}_2$. Then $\widehat{\pi}_{\mathrm{orbit}}(x) = \widehat{\pi}_{\mathrm{orbit}}(\Phi(x)) = \lambda\mathsf{q}_1 + (1-\lambda)\mathsf{q}_2$. Let $\widehat{\pi} \in \mathcal{C}_{\mathrm{coh}}$ be any coherent model with the universal improvement property (6) over $\mathcal{F}$. Since $\mathcal{F}$ contains all quadratic generators and $|\mathcal{Y}| \geq 3$, Theorem 5.3(b) yields $\widehat{\pi} = \widehat{\pi}_{\mathrm{orbit}}$. Now take the coherent target $\pi^*(x) = \pi^*(\Phi(x)) = \mathsf{p}^*$. Then

$$\mathbb{E}[\mathsf{B}_F(\pi^* \,\|\, \widehat{\pi})] = \mathsf{B}_F(\mathsf{p}^* \,\|\, \lambda\mathsf{q}_1 + (1-\lambda)\mathsf{q}_2) > \lambda\mathsf{B}_F(\mathsf{p}^* \,\|\, \mathsf{q}_1) + (1-\lambda)\mathsf{B}_F(\mathsf{p}^* \,\|\, \mathsf{q}_2) = \mathbb{E}[\mathsf{B}_F(\pi^* \,\|\, \pi_0)],$$

by (12), which contradicts (6) for $F \in \mathcal{F}$. Hence no coherent $\widehat{\pi}$ has the universal improvement property on this instance. □

G Characterization and Rigidity of Improvement Mechanisms

This appendix develops the results announced in Section 6. Throughout, $\Pi \subseteq \Pi_{\mathrm{all}}$ is a closed convex set, $\pi_0 \in \Pi_{\mathrm{all}}$ is a baseline, and $\mathcal{D}_\mathcal{X}$ has full support. We equip the space $(\mathbb{R}^d)^\mathcal{X}$ of all functions $\mathcal{X} \to \mathbb{R}^d$ with the $\mathcal{D}_\mathcal{X}$-weighted inner product

$$\langle u, v \rangle_\mathcal{D} = \mathop{\mathbb{E}}_{x \sim \mathcal{D}_\mathcal{X}} [\langle u(x), v(x) \rangle],$$

and write $\|\cdot\|_\mathcal{D}$ for the associated norm. A *mechanism* assigns to each Legendre generator F in a family $\mathcal{F}$ an output $\widehat{\pi}_F \in \Pi$; we always assume that $\pi_0(x)$ and $\widehat{\pi}_F(x)$ belong to $\mathrm{int}(\mathrm{dom}\, F)$ for all x, so that the dual residual

$$V_F(x) = \nabla F(\widehat{\pi}_F(x)) - \nabla F(\pi_0(x)), \qquad x \in \mathcal{X},$$

is well defined. The mechanism satisfies the strong improvement inequality (SII) at F if (7) holds for all $\pi^* \in \Pi$.

G.1 Characterization: provable improvement is Bregman projection

Theorem 6.1 (Characterization of improvement mechanisms)**.** *Let F be Legendre and assume $\pi_0(x)$ and $\widehat{\pi}_F(x)$ lie in $\mathrm{int}(\mathrm{dom}\, F)$ for all x. The following are equivalent: (i) $\widehat{\pi}_F$ satisfies the SII (7) for all $\pi^* \in \Pi$; (ii) $\widehat{\pi}_F$ satisfies the variational inequality $\mathbb{E}[\langle \nabla F(\widehat{\pi}_F(x)) - \nabla F(\pi_0(x)), \pi(x) - \widehat{\pi}_F(x) \rangle] \geq 0$ for all $\pi \in \Pi$; (iii) $\widehat{\pi}_F$ is the (unique) B_F-projection of π_0 onto Π.*

Proof. (i) $\Leftrightarrow$ (ii). For any $\pi^* \in \Pi$ and $x \in \mathcal{X}$, the three-point identity with $a = \pi^*(x)$, $b = \widehat{\pi}_F(x)$ and $c = \pi_0(x)$ reads

$$\mathsf{B}_F(\pi^*(x) \,\|\, \pi_0(x)) = \mathsf{B}_F(\pi^*(x) \,\|\, \widehat{\pi}_F(x)) + \mathsf{B}_F(\widehat{\pi}_F(x) \,\|\, \pi_0(x)) + \langle V_F(x),\ \pi^*(x) - \widehat{\pi}_F(x) \rangle.$$

Taking expectations, the SII (7) for π^* is *equivalent* to $\langle V_F, \pi^* - \widehat{\pi}_F \rangle_\mathcal{D} \geq 0$. Hence (i) holds for all $\pi^* \in \Pi$ iff (ii) holds for all $\pi \in \Pi$.

(ii) $\Rightarrow$ (iii). Let $J(\pi) = \mathbb{E}_{x \sim \mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi(x) \,\|\, \pi_0(x))]$. By Lemma B.6, J is convex and differentiable at $\widehat{\pi}_F$ with gradient $\nabla J(\widehat{\pi}_F) = V_F$ with respect to $\langle \cdot, \cdot \rangle_\mathcal{D}$. Condition (ii) is the first-order optimality condition $\langle \nabla J(\widehat{\pi}_F), \pi - \widehat{\pi}_F \rangle_\mathcal{D} \geq 0$ for all $\pi \in \Pi$, which is sufficient for $\widehat{\pi}_F$ to minimize the convex function J over the convex set Π: by convexity of F and differentiability at the interior points $\widehat{\pi}_F(x)$, $F(\pi(x)) \geq F(\widehat{\pi}_F(x)) + \langle \nabla F(\widehat{\pi}_F(x)), \pi(x) - \widehat{\pi}_F(x) \rangle$ for all $\pi \in \Pi$, so $J(\pi) \geq J(\widehat{\pi}_F) + \langle V_F, \pi - \widehat{\pi}_F \rangle_\mathcal{D} \geq J(\widehat{\pi}_F)$. The minimizer is unique: if $\pi' \in \Pi$ were another minimizer, then J would be constant on the segment $[\widehat{\pi}_F, \pi']$ (by convexity and minimality), whereas for $t \in (0,1)$ the points $\pi_t = \widehat{\pi}_F + t(\pi' - \widehat{\pi}_F)$ satisfy $\pi_t(x) \in \mathrm{int}(\mathrm{dom}\, F)$ (an open segment from an interior point of a convex set to one of its points lies in the interior), where F is strictly convex and $\mathcal{D}_\mathcal{X}$ has full support, so $t \mapsto J(\pi_t)$ is strictly convex on $(0,1)$, a contradiction. Hence $\widehat{\pi}_F$ is the B_F-projection of π_0 onto Π.

(iii) $\Rightarrow$ (ii). If $\widehat{\pi}_F$ minimizes J over Π, then for any $\pi \in \Pi$ and $t \in (0,1]$, $\pi_t = \widehat{\pi}_F + t(\pi - \widehat{\pi}_F) \in \Pi$ and $J(\pi_t) \geq J(\widehat{\pi}_F)$; dividing by t and letting $t \to 0$ gives $\langle \nabla J(\widehat{\pi}_F), \pi - \widehat{\pi}_F \rangle_\mathcal{D} \geq 0$, which is (ii). □

Theorem 6.1 identifies the Bregman projection as the *only* mechanism with the Pythagorean improvement guarantee over a convex set Π. Its proof also shows that the SII is equivalent to a *variational inequality*, which is the form used in all the results below. When $\Pi = \mathcal{C} \cap \Pi'$ for a coherence set $\mathcal{C}$, this says that any mechanism with the guarantee of Theorem 4.1 is a coherence projection. Note that the theorem does not require any continuity assumption on the mechanism.

G.2 Symmetry rigidity

Theorem 6.2 (Symmetry rigidity)**.** *Let $\sigma: \mathcal{X} \to \mathcal{X}$ be a bijection such that Π is σ-invariant ($\pi \in \Pi \Rightarrow \pi \circ \sigma \in \Pi$), $\pi_0 \circ \sigma = \pi_0$, and $\mathcal{D}_\mathcal{X} \circ \sigma = \mathcal{D}_\mathcal{X}$. Then, for every Legendre generator F for which $\widehat{\pi}_F$ satisfies the SII, $\widehat{\pi}_F(\sigma(x)) = \widehat{\pi}_F(x)$ for all $x \in \mathcal{X}$.*

Proof. By Theorem 6.1, $\widehat{\pi}_F$ is the unique minimizer of $J(\pi) = \mathbb{E}_{x\sim\mathcal{D}_\mathcal{X}}[\mathsf{B}_F(\pi(x) \parallel \pi_0(x))]$ over Π. Define $\widetilde{\pi} = \widehat{\pi}_F \circ \sigma$, i.e., $\widetilde{\pi}(x) = \widehat{\pi}_F(\sigma(x))$. Since $\widehat{\pi}_F \in \Pi$ and Π is σ-invariant, $\widetilde{\pi} \in \Pi$ is feasible. Furthermore,

$$
\begin{aligned}
J(\widetilde{\pi}) &= \sum_{x\in\mathcal{X}} \mathcal{D}_\mathcal{X}(x)\, \mathsf{B}_F(\widehat{\pi}_F(\sigma(x)) \parallel \pi_0(x)) = \sum_{x\in\mathcal{X}} \mathcal{D}_\mathcal{X}(x)\, \mathsf{B}_F(\widehat{\pi}_F(\sigma(x)) \parallel \pi_0(\sigma(x))) \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (\pi_0 \circ \sigma = \pi_0) \\
&= \sum_{y\in\mathcal{X}} \mathcal{D}_\mathcal{X}(\sigma^{-1}(y))\, \mathsf{B}_F(\widehat{\pi}_F(y) \parallel \pi_0(y)) \qquad\qquad\qquad\qquad (y = \sigma(x)) \\
&= \sum_{y\in\mathcal{X}} \mathcal{D}_\mathcal{X}(y)\, \mathsf{B}_F(\widehat{\pi}_F(y) \parallel \pi_0(y)) = J(\widehat{\pi}_F). \qquad\qquad\qquad\qquad (\mathcal{D}_\mathcal{X} \circ \sigma^{-1} = \mathcal{D}_\mathcal{X})
\end{aligned}
$$

Thus, $\widetilde{\pi}$ achieves the same objective value as $\widehat{\pi}_F$. By uniqueness of the minimizer, $\widetilde{\pi} = \widehat{\pi}_F$, i.e., $\widehat{\pi}_F(\sigma(x)) = \widehat{\pi}_F(x)$ for all $x \in \mathcal{X}$. □

The theorem establishes that the symmetries of the problem $(\Pi, \pi_0, \mathcal{D}_\mathcal{X})$ are faithfully reflected in the mechanism output, regardless of the generator. It is unconditional and global: it holds for every Legendre function, for any closed convex Π, with no continuity, strictness, or affineness assumption.

When do the symmetry conditions hold? The three invariance conditions arise naturally in several settings. In *fairness-constrained* problems, if Π encodes equal-treatment constraints across demographic groups and the groups have equal representation ($\mathcal{D}_\mathcal{X}$-symmetry) with equal baselines (π_0-symmetry), then the permutation swapping two groups satisfies all three conditions. In *exchangeable* or *group-invariant* problems (e.g., $\pi(x)$ depends on x only through its equivalence class under a group action), the group elements provide the required bijections. The conditions are also met in multi-task settings where tasks share identical structure and are weighted equally. In the coherence setting, the invariance mapping Φ itself is such a σ whenever π_0 is already coherent on an orbit and $\mathcal{D}_\mathcal{X}$ is Φ-invariant; the theorem then shows that no generator can *break* an existing coherence.

Symmetry-forced versus accidental ties. For x, x' in the same orbit of σ, the tie $\widehat{\pi}_F(x) = \widehat{\pi}_F(x')$ holds for every generator F. For x, x' in distinct orbits, a tie can still hold for specific generators (an "accidental" tie), but one expects such ties to be non-generic: they are destroyed by small perturbations of the generator, whereas symmetry-forced ties are structurally robust. The affine results below make this precise in the affine setting (Theorem G.4).

G.3 Affine structure

Let $Z = \{v \in \mathbb{R}^d : \sum_k v_k = 0\}$ denote the zero-sum subspace, $P_Z v = v - (\frac{1}{d}\sum_k v_k)\mathbf{1}$ the orthogonal projection onto Z, and note that $\mathrm{aff}(\Pi_{\mathrm{all}}) - \mathrm{aff}(\Pi_{\mathrm{all}}) = Z^\mathcal{X}$. In this subsection, we assume that Π is *affinely constrained*:

$$
\Pi = A \cap \Pi_{\mathrm{all}}, \qquad A = v_0 + \Pi_0 \subseteq \mathrm{aff}(\Pi_{\mathrm{all}}) \text{ an affine subspace with direction space } \Pi_0 \subseteq Z^\mathcal{X}
$$

(the inclusion $A \subseteq \mathrm{aff}(\Pi_{\mathrm{all}})$ is without loss of generality, since replacing A by $A \cap \mathrm{aff}(\Pi_{\mathrm{all}})$ does not change Π). Such sets arise from linear equality constraints: calibration or moment conditions ($\mathbb{E}[\pi(x)] = \mathbb{E}[\pi_0(x)]$), linear fairness constraints ($\sum_x w_x \pi(x) = c$), or block-equality constraints such as coherence itself. Since Π_{all} is bounded, Π is never affine, but its *directions* are. We write

$$
\Pi_0^\perp = \left\{ w \in Z^\mathcal{X} : \langle w, v \rangle_\mathcal{D} = 0 \text{ for all } v \in \Pi_0 \right\}
$$

for the $\mathcal{D}_\mathcal{X}$-weighted orthogonal complement of Π_0 *within* $Z^\mathcal{X}$, and $r = \dim(\Pi_0^\perp) = n(d-1) - \dim(\Pi_0)$. We say that the mechanism output is *relatively interior* at F, written (RI), if $\widehat{\pi}_F \in \mathrm{relint}(\Pi)$, the relative interior being taken within A; equivalently, for every $v \in \Pi_0$ there exists $\epsilon > 0$ with $\widehat{\pi}_F \pm \epsilon v \in \Pi$.

Remark G.1 (On the relative interior assumption)**.** *Condition (RI) holds as soon as* $\widehat{\pi}_F(x) \in \operatorname{relint}\Delta(\mathcal{Y})$ *for every* x*, since then* $\widehat{\pi}_F \pm \epsilon v \in \Pi_{\mathrm{all}}$ *for* $v \in \Pi_0 \subseteq Z^{\mathcal{X}}$ *and* ϵ *small, and* $\widehat{\pi}_F \pm \epsilon v \in A$ *by affineness. In particular, (RI) is automatic when* F *is steep at the boundary of the simplex (e.g., for the negative entropy,* $\operatorname{dom} F = \mathbb{R}^d_+$*) and* Π *meets* $\operatorname{relint}\Pi_{\mathrm{all}}$*: a boundary point cannot satisfy the first-order optimality condition (ii) of Theorem 6.1, because the gradient of* F *blows up there. For non-steep generators (e.g., quadratic ones), (RI) requires that no simplex constraint be active at* $\widehat{\pi}_F$*; when some are active,* Π *should be treated as a polytope and the results of Section G.4 apply instead.*

The following lemma isolates the key structural property of affine constraints: the dual residual of the mechanism output is orthogonal to the constraint directions.

Lemma G.2 (Affine orthogonality)**.** *Let* $\Pi = A \cap \Pi_{\mathrm{all}}$ *be affinely constrained, let* F *be Legendre, and suppose that* $\widehat{\pi}_F \in \Pi$ *satisfies the SII and (RI). Then,*

$$\langle V_F,\, v\rangle_{\mathcal{D}} = 0 \qquad \textit{for all } v \in \Pi_0, \qquad \textit{equivalently} \qquad w_F := P_Z V_F \in \Pi_0^{\perp},$$

where $(P_Z V_F)(x) = P_Z(V_F(x))$.

Proof. By Theorem 6.1, the SII is equivalent to the variational inequality $\langle V_F, \pi - \widehat{\pi}_F\rangle_{\mathcal{D}} \geq 0$ for all $\pi \in \Pi$. Fix $v \in \Pi_0$. By (RI), there exists $\epsilon > 0$ such that $\pi^{\pm} = \widehat{\pi}_F \pm \epsilon v \in \Pi$. Testing the variational inequality with π^+ and π^- gives $\pm\epsilon\langle V_F, v\rangle_{\mathcal{D}} \geq 0$, hence $\langle V_F, v\rangle_{\mathcal{D}} = 0$. Finally, since $v(x) \in Z$ for all x, $\langle V_F(x), v(x)\rangle = \langle P_Z V_F(x), v(x)\rangle$, so the condition is equivalent to $P_Z V_F \in \Pi_0^{\perp}$. □

The orthogonality $w_F \perp \Pi_0$ says that the mechanism's correction in dual space, the shift from $\nabla F(\pi_0)$ to $\nabla F(\widehat{\pi}_F)$, lives (up to normalization) in the r-dimensional subspace $\Pi_0^{\perp}$ determined by the constraint geometry $(\Pi, \mathcal{D}_{\mathcal{X}})$ alone, independently of both the generator F and the baseline π_0. This single fact has several consequences.

G.3.1 Finite-dimensional rigidity

Theorem 6.3 (Finite-dimensional rigidity for affine constraints)**.** *Let* $\Pi = A \cap \Pi_{\mathrm{all}}$ *be affinely constrained and let* $\mathcal{F}_{\mathrm{RI}}$ *be the set of Legendre generators* F *for which* $\widehat{\pi}_F$ *satisfies the SII and (RI). Then: (a) for every* $F \in \mathcal{F}_{\mathrm{RI}}$*, there exist* $w_F \in \Pi_0^{\perp}$ *and a scalar function* $c_F \colon \mathcal{X} \to \mathbb{R}$ *such that*

$$\nabla F(\widehat{\pi}_F(x)) = \nabla F(\pi_0(x)) + w_F(x) + c_F(x)\mathbf{1}, \qquad x \in \mathcal{X}; \tag{8}$$

(b) for a fixed F*, the dual residual* w_F *determines* $\widehat{\pi}_F$*: if* $\pi, \pi' \in \Pi_{\mathrm{all}}$ *with values in* $\operatorname{int}(\operatorname{dom} F)$ *satisfy* $P_Z(\nabla F(\pi(x)) - \nabla F(\pi_0(x))) = P_Z(\nabla F(\pi'(x)) - \nabla F(\pi_0(x))) = w(x)$ *for all* x*, then* $\pi = \pi'$*; (c) consequently, the family* $\{\widehat{\pi}_F \colon F \in \mathcal{F}_{\mathrm{RI}}\}$ *is parametrized by the* r*-dimensional space* $\Pi_0^{\perp}$*, however large the family of generators; in particular, if* $r = 0$ *(i.e.,* $\Pi = \Pi_{\mathrm{all}}$*), then* $\widehat{\pi}_F = \pi_0$ *for all* F.

Proof. (a) By Lemma G.2, $w_F = P_Z V_F \in \Pi_0^{\perp}$, and $V_F(x) - P_Z V_F(x) = c_F(x)\mathbf{1}$ with $c_F(x) = \frac{1}{d}\sum_k V_F(x)_k$. (b) If $P_Z \nabla F(\pi(x)) = P_Z \nabla F(\pi'(x))$, then $\nabla F(\pi(x)) - \nabla F(\pi'(x)) = c\mathbf{1}$ for some $c \in \mathbb{R}$. Since $\pi(x), \pi'(x) \in \Delta(\mathcal{Y})$, their difference lies in Z, so $\langle \nabla F(\pi(x)) - \nabla F(\pi'(x)), \pi(x) - \pi'(x)\rangle = c\langle \mathbf{1}, \pi(x) - \pi'(x)\rangle = 0$. By strict convexity of F, ∇F is strictly monotone, so $\pi(x) = \pi'(x)$. (c) The map $F \mapsto w_F \in \Pi_0^{\perp}$ composed with the injective correspondence of (b) recovers $\widehat{\pi}_F$; different generators yield different correspondences, but all share the domain $\Pi_0^{\perp}$. If $r = 0$, then $w_F = 0$, so $\nabla F(\widehat{\pi}_F(x)) - \nabla F(\pi_0(x)) \in \operatorname{span}(\mathbf{1})$, and the argument of (b) gives $\widehat{\pi}_F = \pi_0$. □

The *rigidity dimension* r quantifies the degrees of freedom of the mechanism: more constraints (a larger $\dim \Pi_0$) yield a smaller r and a more constrained output. The theorem separates the role of the generator into two components: the choice of the element $w_F \in \Pi_0^{\perp}$ (the *direction* of the dual correction), and the nonlinear distortion $(\nabla F)^{-1}$ (the *scale* of the correction in each coordinate). For problems with few constraints, only a small number of parameters distinguish one generator's output from another's, which limits the practical impact of generator selection.

The hyperplane case ($r = 1$). Suppose Π is defined by a single linear constraint $\sum_{z\in\mathcal{X}}\langle a(z), \pi(z)\rangle = c$ with $a(z) \in Z$ (which is without loss of generality, since $\sum_k \pi(z)_k = 1$) and $a \neq 0$. Then $\Pi_0 = \{v \in Z^{\mathcal{X}} : \sum_z \langle a(z), v(z)\rangle = 0\}$ and $\Pi_0^\perp = \operatorname{span}\{z \mapsto a(z)/\mathcal{D}_{\mathcal{X}}(z)\}$, so $w_F = \lambda_F\, w_0$ for the fixed direction $w_0(z) = a(z)/\mathcal{D}_{\mathcal{X}}(z)$ and a single scalar $\lambda_F \in \mathbb{R}$: the outputs of all generators trace a *one-parameter family* in Π. As an illustration, consider a mean-score constraint $\sum_x \mathcal{D}_{\mathcal{X}}(x)\langle a, \pi(x)\rangle = c$ for a fixed vector $a \in Z$; then $a(x) = \mathcal{D}_{\mathcal{X}}(x)a$ and $w_0 \equiv a$ is constant. For any constant baseline $\pi_0(x) = p$, the representation (8) and Theorem 6.3(b) force the output to be constant as well, $\widehat{\pi}_F(x) = p'_F$ for all x and every Legendre F; the single scalar λ_F only shifts this constant value along the constraint hyperplane.

G.3.2 Universal tie criterion

The symmetry rigidity theorem requires a global bijection σ preserving the entire problem. The following result identifies universal ties at a *pointwise* level, directly from the structure of the constraint. A key observation is that the SII and Lemma G.2 hold for every baseline π_0, so the equivalence relation below, which depends only on $(\Pi, \mathcal{D}_{\mathcal{X}})$, is a *structural* property of the pair (x, x').

Theorem 6.4 (Universal tie criterion for affine constraints)**.** *Let $\Pi = A \cap \Pi_{\text{all}}$ be affinely constrained and define the* structural equivalence *on $\mathcal{X}$ by $x \approx x'$ iff $w(x) = w(x')$ for every $w \in \Pi_0^\perp$. Then, for every baseline $\pi_0 \in \Pi_{\text{all}}$, every Legendre generator F such that $\widehat{\pi}_F$ satisfies the SII and (RI), and every pair $x, x' \in \mathcal{X}$: if $x \approx x'$ and $\pi_0(x) = \pi_0(x')$, then $\widehat{\pi}_F(x) = \widehat{\pi}_F(x')$. In particular, $\widehat{\pi}_F$ belongs to the coherence set $\mathcal{C}_{\approx} = \{\pi \in \Pi : \pi(x) = \pi(x') \text{ for all } x \approx x' \text{ with } \pi_0(x) = \pi_0(x')\}$ for every such F and every π_0.*

Proof. By Lemma G.2, $w_F = P_Z V_F \in \Pi_0^\perp$. Let $x \approx x'$ with $\pi_0(x) = \pi_0(x')$. Then $w_F(x) = w_F(x')$, and by (8),

$$P_Z \nabla F(\widehat{\pi}_F(x)) = P_Z \nabla F(\pi_0(x)) + w_F(x) = P_Z \nabla F(\pi_0(x')) + w_F(x') = P_Z \nabla F(\widehat{\pi}_F(x')).$$

By Theorem 6.3(b) (applied pointwise), $\widehat{\pi}_F(x) = \widehat{\pi}_F(x')$. □

When Π is defined by r explicit linear constraints $\sum_{z\in\mathcal{X}}\langle a_k(z), \pi(z)\rangle = c_k$, $k = 1, \dots, r$, with $a_k(z) \in Z$ (again without loss of generality) and $a_1, \dots, a_r$ linearly independent, the orthogonal complement is $\Pi_0^\perp = \operatorname{span}\{z \mapsto a_k(z)/\mathcal{D}_{\mathcal{X}}(z) : k = 1, \dots, r\}$, and the structural equivalence reduces to the verifiable criterion

$$\frac{a_k(x)}{\mathcal{D}_{\mathcal{X}}(x)} = \frac{a_k(x')}{\mathcal{D}_{\mathcal{X}}(x')} \qquad \text{for all } k = 1, \dots, r,$$

i.e., the $\mathcal{D}_{\mathcal{X}}$-normalized constraint coefficients at x and x' coincide across every constraint. This condition detects *weighted symmetries*, in which the constraint structure and the input distribution compensate each other.

Relation to symmetry rigidity. Theorem 6.2 requires a global bijection preserving $(\Pi, \pi_0, \mathcal{D}_{\mathcal{X}})$ simultaneously; the universal tie criterion operates pointwise and separates the roles of the data: the structural equivalence $\approx$, a property of $(\Pi, \mathcal{D}_{\mathcal{X}})$ alone, determines which pairs *can* be tied, and the baseline π_0 *activates* the tie by satisfying $\pi_0(x) = \pi_0(x')$. As a concrete example, let $\mathcal{X} = \{1, 2\}$, $\mathcal{Y} = \{0, 1\}$, identify a model with $p(x) = \pi(x)(1) = \mathbb{P}[y = 1 \mid x]$, let $\mathcal{D}_{\mathcal{X}} = (1/3, 2/3)$ and impose the affine constraint $p(1) + 2p(2) = c$. In vector form, $a(1) = P_Z(0, 1) = (-\frac{1}{2}, \frac{1}{2})$ and $a(2) = P_Z(0, 2) = (-1, 1)$, so that $a(1)/\mathcal{D}_{\mathcal{X}}(1) = (-\frac{3}{2}, \frac{3}{2}) = a(2)/\mathcal{D}_{\mathcal{X}}(2)$ and $1 \approx 2$. Hence for *any* baseline with $\pi_0(1) = \pi_0(2)$ and any Legendre F (with (RI), e.g., any steep F), the tie $\widehat{\pi}_F(1) = \widehat{\pi}_F(2)$ holds. Yet the transposition $1 \leftrightarrow 2$ does not preserve $\mathcal{D}_{\mathcal{X}}$, so Theorem 6.2 does not apply: the universal tie criterion captures structural ties that are invisible to permutation symmetry. (One can verify the tie directly: with f the one-dimensional generator, the first-order conditions read $\frac{1}{3}(f'(p(1)) - f'(p_0)) = \lambda$ and $\frac{2}{3}(f'(p(2)) - f'(p_0)) = 2\lambda$, which force $p(1) = p(2)$.)

G.3.3 Exact universal partition

Theorem 6.4 provides a *sufficient* condition for ties to hold universally. Under a mild richness assumption on the generator family, the condition is also *necessary*.

Definition G.3 (Generator richness)**.** *A family $\mathcal{F} \subseteq \mathcal{F}_{\mathrm{RI}}$ is* rich *relative to $(\Pi, \pi_0, \mathcal{D}_{\mathcal{X}})$ if the dual residuals $\{w_F : F \in \mathcal{F}\}$ span $\Pi_0^{\perp}$.*

Richness asks that the generators probe all directions of the constraint geometry. For instance, varying the matrix M in the quadratic family $F_{\mathsf{M}}(p) = \frac{1}{2} p^{\top} \mathsf{M} p$ rotates the M-weighted projection and one expects the corresponding residuals to span $\Pi_0^{\perp}$ for generic $(\Pi, \pi_0, \mathcal{D}_{\mathcal{X}})$; in the hyperplane case, richness reduces to the requirement that some F has $\lambda_F \neq 0$, i.e., that the mechanism actually moves away from the baseline.

Theorem G.4 (Exact universal partition for affine constraints)**.** *Let $\Pi = A \cap \Pi_{\mathrm{all}}$ be affinely constrained and let $\mathcal{F} \subseteq \mathcal{F}_{\mathrm{RI}}$ be rich. Define $x \sim x'$ iff $\pi_0(x) = \pi_0(x')$ and $x \approx x'$. Then, for every pair $x, x' \in \mathcal{X}$ with $\pi_0(x) = \pi_0(x')$,*

$$\widehat{\pi}_F(x) = \widehat{\pi}_F(x') \text{ for all } F \in \mathcal{F} \quad \Longleftrightarrow \quad x \sim x'. \tag{13}$$

Thus, within each level set of π_0, the relation "tied by every generator of $\mathcal{F}$" coincides with the structural relation $\approx$, which is computable from $(\Pi, \mathcal{D}_{\mathcal{X}})$ alone; its classes form the exact universal partition*.*

Proof. The implication $\Leftarrow$ is Theorem 6.4. For $\Rightarrow$, we prove the contrapositive. Suppose $x \not\sim x'$ with $\pi_0(x) = \pi_0(x')$. Then there exists $w_0 \in \Pi_0^{\perp}$ with $w_0(x) \neq w_0(x')$. The linear map $L : \Pi_0^{\perp} \to Z$, $L(w) = w(x) - w(x')$, is therefore nonzero. By richness, $\{w_F : F \in \mathcal{F}\}$ spans $\Pi_0^{\perp}$, so L cannot vanish on all w_F: there exists $F_* \in \mathcal{F}$ with $w_{F_*}(x) \neq w_{F_*}(x')$. Since $\pi_0(x) = \pi_0(x')$, (8) gives $P_Z \nabla F_*(\widehat{\pi}_{F_*}(x)) - P_Z \nabla F_*(\widehat{\pi}_{F_*}(x')) = w_{F_*}(x) - w_{F_*}(x') \neq 0$, hence $\widehat{\pi}_{F_*}(x) \neq \widehat{\pi}_{F_*}(x')$, and the tie does not hold for all $F \in \mathcal{F}$. □

Unequal baselines. Theorem G.4 addresses pairs with $\pi_0(x) = \pi_0(x')$. When $\pi_0(x) \neq \pi_0(x')$, a universal tie can still occur if the constraint *forces* it (i.e., $\pi(x) = \pi(x')$ for every $\pi \in \Pi$). Outside this case, under richness, one expects the universal tie to fail generically: by (8), a tie for F_{M} requires $P_Z \mathsf{M}(\pi_0(x) - \pi_0(x')) = w_{F_{\mathsf{M}}}(x') - w_{F_{\mathsf{M}}}(x) \in L(\Pi_0^{\perp})$, and the left-hand side sweeps a set of dimension $d - 1$ as M ranges over positive definite matrices, which cannot be contained in $L(\Pi_0^{\perp})$ when $\dim L(\Pi_0^{\perp}) < d - 1$.

G.3.4 Bregman Pythagorean identity and cross-generator gap

Theorem G.5 (Bregman Pythagorean identity for affine constraints)**.** *Let $\Pi = A \cap \Pi_{\mathrm{all}}$ be affinely constrained and let F be Legendre such that $\widehat{\pi}_F$ satisfies the SII and (RI). Then, for any $\pi \in \Pi$,*

$$\mathbb{E}[\mathsf{B}_F(\pi(x) \parallel \pi_0(x))] = \mathbb{E}[\mathsf{B}_F(\pi(x) \parallel \widehat{\pi}_F(x))] + \mathbb{E}[\mathsf{B}_F(\widehat{\pi}_F(x) \parallel \pi_0(x))]. \tag{14}$$

In particular, for the output $\widehat{\pi}_G$ of a second generator G, the cross-generator gap is

$$\mathbb{E}[\mathsf{B}_F(\widehat{\pi}_G(x) \parallel \widehat{\pi}_F(x))] = \mathbb{E}[\mathsf{B}_F(\widehat{\pi}_G(x) \parallel \pi_0(x))] - \mathbb{E}[\mathsf{B}_F(\widehat{\pi}_F(x) \parallel \pi_0(x))] \geq 0, \tag{15}$$

with equality if and only if $\widehat{\pi}_G = \widehat{\pi}_F$.

Proof. The three-point identity applied pointwise with $a = \pi(x)$, $b = \widehat{\pi}_F(x)$, $c = \pi_0(x)$ and integrated gives, for any $\pi \in \Pi$, $\mathbb{E}[\mathsf{B}_F(\pi \parallel \pi_0)] = \mathbb{E}[\mathsf{B}_F(\pi \parallel \widehat{\pi}_F)] + \mathbb{E}[\mathsf{B}_F(\widehat{\pi}_F \parallel \pi_0)] + \langle V_F,\ \pi - \widehat{\pi}_F \rangle_{\mathcal{D}}$. Since $\pi, \widehat{\pi}_F \in \Pi \subseteq A$, the difference $\pi - \widehat{\pi}_F$ lies in Π_0, and by Lemma G.2 the cross-term vanishes, which yields (14). The identity (15) is the special case $\pi = \widehat{\pi}_G \in \Pi$; non-negativity follows from $\mathsf{B}_F \geq 0$, with equality iff $\widehat{\pi}_G(x) = \widehat{\pi}_F(x)$ for all x, by strict convexity of F. □

The identity (14) is the Bregman analogue of the Pythagorean theorem: π_0, $\widehat{\pi}_F$ and π form a right triangle in the Bregman geometry, with the right angle at $\widehat{\pi}_F$. This is strictly stronger than what holds for general convex Π, where only the inequality $\geq$ is available. The cross-generator gap (15) quantifies the exact cost of using one generator's output in place of another's.

Corollary G.6 (Cross-generator proximity)**.** *Under the hypotheses of Theorem G.5, assume further that F is m-strongly convex on $\Delta(\mathcal{Y})$ with respect to the Euclidean norm for some $m > 0$. Then, for any generator G with $\widehat{\pi}_G \in \Pi$,*

$$\frac{m}{2} \sum_{x \in \mathcal{X}} \mathcal{D}_{\mathcal{X}}(x) \|\widehat{\pi}_G(x) - \widehat{\pi}_F(x)\|_2^2 \leq \mathbb{E}[\mathsf{B}_F(\widehat{\pi}_G(x) \parallel \pi_0(x))] - \mathbb{E}[\mathsf{B}_F(\widehat{\pi}_F(x) \parallel \pi_0(x))],$$

and in particular, for each $x \in \mathcal{X}$,

$$\|\widehat{\pi}_G(x) - \widehat{\pi}_F(x)\|_2^2 \le \frac{2}{m\,\mathcal{D}_{\mathcal{X}}(x)} \left(\mathbb{E}[\mathsf{B}_F(\widehat{\pi}_G(x) \,\|\, \pi_0(x))] - \mathbb{E}[\mathsf{B}_F(\widehat{\pi}_F(x) \,\|\, \pi_0(x))]\right). \tag{16}$$

Proof. The m-strong convexity of F implies $\mathsf{B}_F(q \,\|\, p) \ge \frac{m}{2}\|q - p\|_2^2$ for $q, p \in \Delta(\mathcal{Y})$. Applying this pointwise and taking expectations bounds $\mathbb{E}[\mathsf{B}_F(\widehat{\pi}_G(x) \,\|\, \widehat{\pi}_F(x))]$ from below; the left-hand side equals the cross-generator gap by Theorem G.5. The pointwise bound (16) follows by isolating the single term x from the sum. □

The m-strong convexity is satisfied by many standard generators: the L^2 generator $F(p) = \frac{1}{2}\|p\|_2^2$ has $m = 1$, and the negative entropy has $\nabla^2 F(p) = \mathrm{diag}(1/p_i) \succeq I$ on $\Delta(\mathcal{Y})$, giving $m = 1$. The pointwise bound (16) has a direct consequence for the stability of the level-set partition. If all distinct block values of $\widehat{\pi}_F$ are δ-separated ($\|\widehat{\pi}_F(x) - \widehat{\pi}_F(x')\|_2 \ge \delta$ for x, x' in distinct blocks) and the cross-generator gap satisfies $\mathbb{E}[\mathsf{B}_F(\widehat{\pi}_G \,\|\, \pi_0)] - \mathbb{E}[\mathsf{B}_F(\widehat{\pi}_F \,\|\, \pi_0)] < \frac{m\,\delta^2\,\mu_{\min}}{8}$ with $\mu_{\min} = \min_x \mathcal{D}_{\mathcal{X}}(x)$, then $\|\widehat{\pi}_G(x) - \widehat{\pi}_F(x)\|_2 < \delta/2$ for all x, and by the triangle inequality no two distinct blocks of $\widehat{\pi}_F$ can merge in $\widehat{\pi}_G$. Thus, for affinely constrained Π and strongly convex generators, nearby generators (in the cross-gap sense) cannot merge distinct blocks, complementing the symmetry rigidity, which prevents block splitting on orbits.

G.4 Local rigidity for polyhedral constraints

We now consider a closed convex Π that is a *polytope*,

$$\Pi = \left\{\pi \in (\mathbb{R}^d)^{\mathcal{X}} : \langle g_j, \pi\rangle_{\mathcal{D}} \le c_j,\ j = 1, \dots, m\right\} \subseteq \Pi_{\mathrm{all}},$$

where we may include among the m constraints those defining Π_{all}. Polytopes cover all constraint sets defined by finitely many linear equalities and inequalities (equalities being pairs of opposite inequalities): coherence sets, order constraints $\pi(y \mid x) \ge \pi(y' \mid x)$, calibration constraints, and their intersections with Π_{all}. For $y \in (\mathbb{R}^d)^{\mathcal{X}}$, let $E(y) = \mathrm{argmax}_{\pi \in \Pi} \langle y, \pi\rangle_{\mathcal{D}}$ denote the face of Π *exposed* by y; it is a nonempty face of Π, and a polytope has finitely many faces. For $\pi \in \Pi$, let $J(\pi) = \{j : \langle g_j, \pi\rangle_{\mathcal{D}} = c_j\}$ be the set of active constraints, $N_\Pi(\pi) = \mathrm{cone}\{g_j : j \in J(\pi)\}$ the normal cone of Π at π, and $\mathcal{F}(\pi) = \{\pi' \in \Pi : J(\pi') \supseteq J(\pi)\}$ the minimal face of Π containing π.

In this subsection, $\mathcal{F}$ is a family of Legendre generators whose domains contain a fixed open convex neighborhood of $\Delta(\mathcal{Y})$, equipped with the *generator topology* of uniform convergence of F and ∇F on $\Delta(\mathcal{Y})$. In this setting, the mechanism is automatically *stable*.

Lemma G.7 (Stability)**.** *If the mechanism satisfies the SII at every $F \in \mathcal{F}$, then the map $F \mapsto \widehat{\pi}_F$ is continuous from $\mathcal{F}$ into $(\mathbb{R}^d)^{\mathcal{X}}$, and so is $F \mapsto V_F$.*

Proof. By Theorem 6.1, $\widehat{\pi}_F$ is the unique minimizer over the compact set Π of $J_F(\pi) = \mathbb{E}_{x \sim \mathcal{D}_{\mathcal{X}}}[F(\pi(x)) - F(\pi_0(x)) - \langle \nabla F(\pi_0(x)), \pi(x) - \pi_0(x)\rangle]$. The map $(F, \pi) \mapsto J_F(\pi)$ is jointly continuous on $\mathcal{F} \times \Pi$: uniform convergence of F and ∇F on $\Delta(\mathcal{Y})$ together with the continuity of each F and ∇F on the compact set $\Delta(\mathcal{Y})$ give $J_{F_n}(\pi_n) \to J_F(\pi)$ whenever $F_n \to F$ and $\pi_n \to \pi$. By Berge's maximum theorem [Berge, 1963, Aliprantis and Border, 2006], the argmin correspondence is upper hemicontinuous, and since the minimizer is unique, it is a continuous function of F. Continuity of $F \mapsto V_F$ follows: for $F_n \to F_0$, $\|\nabla F_n(\widehat{\pi}_{F_n}(x)) - \nabla F_0(\widehat{\pi}_{F_0}(x))\| \le \sup_{\Delta(\mathcal{Y})} \|\nabla F_n - \nabla F_0\| + \|\nabla F_0(\widehat{\pi}_{F_n}(x)) - \nabla F_0(\widehat{\pi}_{F_0}(x))\| \to 0$ by uniform continuity of ∇F_0 on $\Delta(\mathcal{Y})$, and similarly $\nabla F_n(\pi_0(x)) \to \nabla F_0(\pi_0(x))$. □

Theorem G.8 (Local face rigidity for polyhedral constraints)**.** *Let Π be a polytope and let the mechanism satisfy the SII at every $F \in \mathcal{F}$. For $F \in \mathcal{F}$, let $G_F = E(-V_F)$ be the face of Π exposed by $-V_F$. Then:*

(a) *$\widehat{\pi}_F \in G_F$ for every $F \in \mathcal{F}$.*

(b) *For every $F_0 \in \mathcal{F}$, there is a neighborhood $\mathcal{U}$ of F_0 in the generator topology such that $G_F \subseteq G_{F_0}$ for all $F \in \mathcal{U}$; in particular, $\widehat{\pi}_F \in G_{F_0}$ for all $F \in \mathcal{U}$.*

(c) *Consequently, every affine identity that holds on G_{F_0}, in particular every tie $\pi(x) = \pi(x')$ or block-equality satisfied by all $\pi \in G_{F_0}$, is satisfied* exactly *by $\widehat{\pi}_F$ for all $F \in \mathcal{U}$.*

(d) *If, moreover,* strict complementarity *holds at F_0, i.e., $-V_{F_0} \in \operatorname{relint} N_\Pi(\widehat{\pi}_{F_0})$, then $G_{F_0} = \mathcal{F}(\widehat{\pi}_{F_0})$ is the minimal face of Π containing $\widehat{\pi}_{F_0}$, so that for all $F \in \mathcal{U}$ the output $\widehat{\pi}_F$ lies in the minimal face of $\widehat{\pi}_{F_0}$ and $J(\widehat{\pi}_F) \supseteq J(\widehat{\pi}_{F_0})$. Finally, G_{F_0} is a proper face of Π unless V_{F_0} is $\langle\cdot,\cdot\rangle_{\mathcal{D}}$-orthogonal to the direction space of* $\operatorname{aff}\Pi$.

Proof. (a) By Theorem 6.1, the SII at F is equivalent to $\langle V_F, \pi - \widehat{\pi}_F\rangle_{\mathcal{D}} \geq 0$ for all $\pi \in \Pi$, i.e., $\langle -V_F, \pi\rangle_{\mathcal{D}} \leq \langle -V_F, \widehat{\pi}_F\rangle_{\mathcal{D}}$ for all $\pi \in \Pi$, which says exactly that $\widehat{\pi}_F \in E(-V_F) = G_F$.

(b) By Lemma G.7, $V_{F_n} \to V_{F_0}$ whenever $F_n \to F_0$ in the generator topology. Next, we show that the exposed-face map is upper semicontinuous at $-V_{F_0}$: if $y_n \to y$, then $E(y_n) \subseteq E(y)$ for all n large enough. Suppose not. Since Π has finitely many faces, passing to a subsequence we may assume that $E(y_n) = G_1$ is constant with $G_1 \not\subseteq E(y)$. For any $\pi_1 \in G_1$ and any $\pi \in \Pi$, $\langle y_n, \pi\rangle_{\mathcal{D}} \leq \langle y_n, \pi_1\rangle_{\mathcal{D}}$ for all n; letting $n \to \infty$ gives $\langle y, \pi\rangle_{\mathcal{D}} \leq \langle y, \pi_1\rangle_{\mathcal{D}}$, so $\pi_1 \in E(y)$. Hence $G_1 \subseteq E(y)$, a contradiction. Applying this with $y_n = -V_{F_n}$ and $y = -V_{F_0}$ shows that there is a neighborhood $\mathcal{U}$ of F_0 on which $G_F \subseteq G_{F_0}$, and (a) gives $\widehat{\pi}_F \in G_F \subseteq G_{F_0}$ for $F \in \mathcal{U}$.

(c) is immediate from (b).

(d) Write $\pi_\infty = \widehat{\pi}_{F_0}$, $y = -V_{F_0}$ and $N = N_\Pi(\pi_\infty)$. Since G_{F_0} is a face of Π containing π_∞, it contains the minimal such face: $\mathcal{F}(\pi_\infty) \subseteq G_{F_0}$. Conversely, let $\pi' \in G_{F_0}$. Since both π' and π_∞ maximize $\langle y, \cdot\rangle_{\mathcal{D}}$ over Π, we have $\langle y, \pi' - \pi_\infty\rangle_{\mathcal{D}} = 0$. Fix $y' \in N$. Since $y \in \operatorname{relint} N$ and N is convex, there exists $t > 0$ such that $y_- = y + t(y - y') \in N$ [Rockafellar, 1997, Theorem 6.4]. Both y_- and y' belong to the normal cone at π_∞, so $\langle y_-, \pi' - \pi_\infty\rangle_{\mathcal{D}} \leq 0$ and $\langle y', \pi' - \pi_\infty\rangle_{\mathcal{D}} \leq 0$. Since $(1+t)y = y_- + ty'$, we get $0 = (1+t)\langle y, \pi' - \pi_\infty\rangle_{\mathcal{D}} = \langle y_-, \pi' - \pi_\infty\rangle_{\mathcal{D}} + t\langle y', \pi' - \pi_\infty\rangle_{\mathcal{D}}$, a sum of two non-positive terms, so both vanish: $\langle y', \pi' - \pi_\infty\rangle_{\mathcal{D}} = 0$ for all $y' \in N$. Taking $y' = g_j$ for $j \in J(\pi_\infty)$ gives $\langle g_j, \pi'\rangle_{\mathcal{D}} = \langle g_j, \pi_\infty\rangle_{\mathcal{D}} = c_j$, so $J(\pi') \supseteq J(\pi_\infty)$ and $\pi' \in \mathcal{F}(\pi_\infty)$. Hence $G_{F_0} = \mathcal{F}(\pi_\infty)$, and the inclusion $J(\widehat{\pi}_F) \supseteq J(\widehat{\pi}_{F_0})$ for $F \in \mathcal{U}$ follows from (b). Finally, if $G_{F_0} = \Pi$, then $\langle -V_{F_0}, \cdot\rangle_{\mathcal{D}}$ is constant on Π, i.e., V_{F_0} is orthogonal to the direction space of $\operatorname{aff}\Pi$. □

The theorem is a genuine rigidity statement: the face of Π on which the mechanism lands is *locally frozen*, and so is every coherence relation (tie) that this face encodes. Plain continuity (Lemma G.7) would only give $\widehat{\pi}_F \approx \widehat{\pi}_{F_0}$, which does not preserve ties; the conclusion of (c) is an exact equality. The strict complementarity condition in (d) is the standard non-degeneracy assumption of sensitivity analysis in polyhedral optimization and holds generically; when it fails, (b) still applies to the possibly larger exposed face G_{F_0}. For affinely constrained Π with (RI), the exposed face is Π itself and the theorem reduces to the affine orthogonality of Lemma G.2.

Remark G.9 (Why a uniform strictness gap cannot be assumed)**.** *One might attempt to obtain local rigidity directly from a quantitative version of the variational inequality: a uniform gap $\inf_{\pi \in \Pi \setminus \mathcal{C}} \langle V_{F_0}, \pi - \widehat{\pi}_{F_0}\rangle_{\mathcal{D}} \geq \gamma > 0$ over the competitors outside a coherence set $\mathcal{C} \ni \widehat{\pi}_{F_0}$, combined with continuity. Such a gap can never hold unless $\mathcal{C} = \Pi$: the map $\pi \mapsto \langle V_{F_0}, \pi - \widehat{\pi}_{F_0}\rangle_{\mathcal{D}}$ is affine and vanishes at $\widehat{\pi}_{F_0}$, and if $\pi' \in \Pi \setminus \mathcal{C}$, then by convexity $(1-s)\widehat{\pi}_{F_0} + s\pi' \in \Pi \setminus \mathcal{C}$ for all $s \in (0,1]$ (a violated tie of π' remains violated), with value $s\langle V_{F_0}, \pi' - \widehat{\pi}_{F_0}\rangle_{\mathcal{D}} \to 0$ as $s \to 0$. The face-based formulation of Theorem G.8 replaces the (unavailable) gap by the finiteness of the face lattice of a polytope. Moreover, exact preservation of* all *ties of $\widehat{\pi}_{F_0}$ cannot hold for general convex Π: if Π is a Euclidean ball and $\widehat{\pi}_{F_0}$ lies on the sphere with an accidental tie $\widehat{\pi}_{F_0}(x) = \widehat{\pi}_{F_0}(x')$, perturbing the generator slides the projection along the sphere and breaks the tie. Only ties encoded by the geometry of Π (faces, symmetries, affine directions) are rigid, which is exactly what the results of this appendix capture.*

The right general-convex substitute for the gap is a *conic* (sharp-minimum) condition, under which a quantitative form of rigidity holds for any closed convex Π; the proof is self-contained and does not use stability.

Proposition G.10 (Quantitative rigidity under conic strictness)**.** *Let Π be closed convex, let F_0 be a Legendre generator at which the mechanism satisfies the SII, and let $\mathcal{C} \subseteq \Pi$ be a closed convex set containing $\widehat{\pi}_{F_0}$ (for instance the level-set coherence set $\mathcal{C}_{F_0}$ of $\widehat{\pi}_{F_0}$). Assume* conic strictness at F_0*: there is $\gamma > 0$ such that*

$$\langle V_{F_0}, \pi - \widehat{\pi}_{F_0}\rangle_{\mathcal{D}} \geq \gamma \operatorname{dist}_{\mathcal{D}}(\pi, \mathcal{C}) \qquad \text{for all } \pi \in \Pi.$$

Then, for every Legendre F at which the mechanism satisfies the SII, with $\epsilon_F = \sup_{\mathsf{p}\in\Delta(\mathcal{Y})} \|\nabla F(\mathsf{p}) - \nabla F_0(\mathsf{p})\|_2$:

(a) $\gamma \operatorname{dist}_{\mathcal{D}}(\widehat{\pi}_F, \mathcal{C}) \le 2\epsilon_F \|\widehat{\pi}_F - \widehat{\pi}_{F_0}\|_{\mathcal{D}}$*;*

(b) *if, in addition, F_0 is μ-strongly convex on $\Delta(\mathcal{Y})$ with respect to the Euclidean norm, then $\|\widehat{\pi}_F - \widehat{\pi}_{F_0}\|_{\mathcal{D}} \le 2\epsilon_F/\mu$ and $\operatorname{dist}_{\mathcal{D}}(\widehat{\pi}_F, \mathcal{C}) \le \epsilon_F^2/(\mu\gamma)$.*

Thus, the output approaches $\mathcal{C}$ to second order*: its distance to the coherence set is negligible compared with its distance to $\widehat{\pi}_{F_0}$.*

Proof. Let $\Delta = \widehat{\pi}_F - \widehat{\pi}_{F_0}$ and $d = \operatorname{dist}_{\mathcal{D}}(\widehat{\pi}_F, \mathcal{C})$. By Theorem 6.1, the SII at F_0 and at F give the two variational inequalities $\langle V_{F_0}, \widehat{\pi}_F - \widehat{\pi}_{F_0}\rangle_{\mathcal{D}} \ge \gamma d$ (using conic strictness with the competitor $\widehat{\pi}_F \in \Pi$) and $\langle V_F, \widehat{\pi}_{F_0} - \widehat{\pi}_F\rangle_{\mathcal{D}} \ge 0$ (competitor $\widehat{\pi}_{F_0} \in \Pi$). Adding them, $\langle V_{F_0} - V_F, \Delta\rangle_{\mathcal{D}} \ge \gamma d$. Now,

$$V_{F_0} - V_F = -[\nabla F_0(\widehat{\pi}_F) - \nabla F_0(\widehat{\pi}_{F_0})] + [(\nabla F_0 - \nabla F)(\widehat{\pi}_F)] + [(\nabla F - \nabla F_0)(\pi_0)].$$

By monotonicity of ∇F_0, the first bracket contributes $-\langle \nabla F_0(\widehat{\pi}_F) - \nabla F_0(\widehat{\pi}_{F_0}), \Delta\rangle_{\mathcal{D}} = -\mathbb{E}[\mathsf{B}_{F_0}(\widehat{\pi}_F \parallel \widehat{\pi}_{F_0}) + \mathsf{B}_{F_0}(\widehat{\pi}_{F_0} \parallel \widehat{\pi}_F)] \le 0$, and $\le -\mu\|\Delta\|_{\mathcal{D}}^2$ under μ-strong convexity. The last two brackets are bounded pointwise by ϵ_F in Euclidean norm, so by Cauchy–Schwarz they contribute at most $2\epsilon_F\|\Delta\|_{\mathcal{D}}$. Hence $\gamma d \le 2\epsilon_F\|\Delta\|_{\mathcal{D}}$, which is (a), and under strong convexity $\gamma d + \mu\|\Delta\|_{\mathcal{D}}^2 \le 2\epsilon_F\|\Delta\|_{\mathcal{D}}$. Since $\gamma d \ge 0$, this gives $\|\Delta\|_{\mathcal{D}} \le 2\epsilon_F/\mu$, and maximizing $t \mapsto 2\epsilon_F t - \mu t^2$ gives $\gamma d \le \epsilon_F^2/\mu$, which is (b). □

Conic strictness is satisfiable, unlike the uniform gap: it holds trivially when $\mathcal{C} = \Pi$, it fails on the ball example above (there the left-hand side grows only quadratically in the tangential displacement, consistently with the loss of rigidity), and for a polytope with $\mathcal{C}$ the minimal face of $\widehat{\pi}_{F_0}$ it is implied by strict complementarity (by Hoffman's error bound for the polyhedral cone of feasible directions). In the polyhedral case, the second-order closeness of Proposition G.10 is upgraded to *exact* membership by Theorem G.8; the two results together are the correct form of the "potential well" picture: the geometry of Π creates a well around the coherence set, and continuity in the generator cannot lift the output out of it.

Corollary G.11 (Global face rigidity along connected families)**.** *Under the hypotheses of Theorem G.8, let $\widetilde{\mathcal{F}} \subseteq \mathcal{F}$ be a path-connected subfamily such that strict complementarity holds at every $F \in \widetilde{\mathcal{F}}$. Then, the minimal face $\mathcal{F}(\widehat{\pi}_F)$ is the same for all $F \in \widetilde{\mathcal{F}}$: all outputs lie in the relative interior of a single face $\mathcal{F}_0$ of Π, and every tie encoded by $\mathcal{F}_0$ holds for every $F \in \widetilde{\mathcal{F}}$.*

Proof. The argument is topological: we show that the set of generators whose output lies in a given face is both open and closed. For each face H of Π, let $S_H = \{F \in \widetilde{\mathcal{F}} : \mathcal{F}(\widehat{\pi}_F) = H\}$. The sets S_H are finitely many, pairwise disjoint, and cover $\widetilde{\mathcal{F}}$. We claim that each S_H is open. Let $G \in S_H$, so that $\widehat{\pi}_G \in \operatorname{relint} H$ and, by Theorem G.8(d) at G, $G_G = H$. By Theorem G.8(b), there is a neighborhood of G on which $\widehat{\pi}_F \in H$; by Lemma G.7, $\widehat{\pi}_F \to \widehat{\pi}_G$ as $F \to G$, and since $\operatorname{relint} H$ is relatively open in H and contains $\widehat{\pi}_G$, we have $\widehat{\pi}_F \in \operatorname{relint} H$, i.e., $\mathcal{F}(\widehat{\pi}_F) = H$, for F close enough to G. Thus, S_H is open, and so is its complement $\bigcup_{H' \ne H} S_{H'}$, i.e., S_H is also closed. A connected space has no nontrivial subset that is both open and closed, so $\widetilde{\mathcal{F}} = S_{\mathcal{F}_0}$ for the single face $\mathcal{F}_0 = \mathcal{F}(\widehat{\pi}_{G_0})$, where G_0 is any element of $\widetilde{\mathcal{F}}$. □

G.5 Example: an emergent tie in the hyperplane case

This toy example shows how a nontrivial tie can be *emergent*, i.e., not explicitly enforced by the constraint equations, yet hold for *every* Legendre generator, exactly as predicted by the results above. Let $\mathcal{X} = \{1, 2, 3\}$ and $\mathcal{Y} = \{0, 1\}$, and identify a model π with $p \in [0, 1]^3$ where $p(k) = \mathbb{P}(y = 1 \mid x = k)$. Take $\mathcal{D}_{\mathcal{X}}$ uniform on $\mathcal{X}$. Fix $S \in (0, 3)$ and define the feasible set by a single linear equality constraint, $\Pi = \{\pi \in \Pi_{\text{all}} : p(1) + p(2) + p(3) = S\}$, which does *not* force $p(1) = p(2)$. Choose a baseline that is symmetric in coordinates 1 and 2 but infeasible, for example $\pi_0 = (a, a, b)$ with $a, b \in (0, 1)$ and $2a + b \ne S$.

The tie is universal but not forced. The transposition τ swapping inputs $1 \leftrightarrow 2$ is a symmetry of $(\Pi, \pi_0, \mathcal{D}_\mathcal{X})$: Π is invariant under τ, $\tau\pi_0 = \pi_0$, and $\mathcal{D}_\mathcal{X}$ is uniform. Therefore, Theorem 6.2 implies $\widehat{\pi}_F(1) = \widehat{\pi}_F(2)$ for every Legendre generator F (no interiority assumption is needed here). This is an emergent tie: it comes from the interaction of the constraint geometry with the symmetric baseline and uniform $\mathcal{D}_\mathcal{X}$, not from an explicit constraint equation. The actual values differ across generators: for the squared Euclidean generator the common value is obtained by shifting (a, a, b) uniformly, while for the negative entropy it is obtained by an exponential tilt, but the tie itself is rigid.

Connection to $r = 1$ and the affine characterization. Here Π is a hyperplane section of Π_{all}, with $\Pi_0^\perp$ one-dimensional and spanned by the constant function $z \mapsto a(z)/\mathcal{D}_\mathcal{X}(z)$ (all normalized coefficients are equal), so $r = 1$ and every pair of inputs is structurally equivalent. By Theorem 6.4, for any interior output, *every* tie of the baseline is preserved: $\pi_0(1) = \pi_0(2)$ implies $\widehat{\pi}_F(1) = \widehat{\pi}_F(2)$, recovering the conclusion of the symmetry argument, and by Theorem 6.3 the entire family $\{\widehat{\pi}_F\}$ is parametrized by a single scalar λ_F.

G.6 Discussion: rigidity versus characterization

The term *rigidity*, borrowed from geometry and topology, is most fitting for the results showing that the mechanism *cannot* deviate from a predetermined structure. The symmetry rigidity (Theorem 6.2) is a genuine rigidity statement: every symmetry of $(\Pi, \pi_0, \mathcal{D}_\mathcal{X})$ is rigidly inherited by the mechanism, for every generator. The local face rigidity (Theorem G.8) is a rigidity result in the classical sense: the face of the polytope on which the mechanism lands, hence every tie it encodes, is locally frozen; small perturbations of the generator cannot alter it, and Proposition G.10 is its quantitative counterpart for general convex constraints. The affine results are more accurately described as *structural characterizations*: the finite-dimensional rigidity (Theorem 6.3) and the exact universal partition (Theorem G.4) provide complete, computable descriptions of the mechanism's degrees of freedom (the r-dimensional subspace $\Pi_0^\perp$) and of its tie structure (the equivalence $\approx$), and the Bregman Pythagorean identity (Theorem G.5) complements these with a quantitative geometric identity. Taken together with the characterization theorem (Theorem 6.1), these results show that any provable improvement mechanism is a Bregman projection whose behavior is highly constrained by the problem geometry $(\Pi, \mathcal{D}_\mathcal{X})$, limiting the impact of generator selection and providing a principled foundation for the coherence-based self-improvement framework.

H Empirical Evaluations: Details

We empirically evaluate whether coherence training improves both downstream question-answering performance and semantic consistency across paraphrases. The goal of these experiments is not to optimize a full training recipe, but to test whether the proposed coherence-based objective can improve model behavior in a realistic language-model setting.

Models and dataset. We evaluate three instruction-tuned language models: `Phi-3-mini-4k-instruct` [Abdin et al., 2024], released under the MIT License;[2] `Qwen2.5-3B-Instruct` [Yang et al., 2024], released under the Qwen Research License;[3] and `Llama-3.2-3B-Instruct` [Dubey et al., 2024], released under the Llama 3.2 Community License.[4] All models are used through their public HuggingFace checkpoints. We use the `rc.nocontext` configuration of TriviaQA [Joshi et al., 2017] through the HuggingFace `datasets` package. TriviaQA is released under the Apache License 2.0, which applies to both the code and the data according to the official TriviaQA repository. The validation set contains 17,944 examples. The training split contains 138,384 usable question classes after preprocessing. For coherence training, each of the 200 batches is formed by sampling 32 classes uniformly at random without replacement from these 138,384 classes (Python's `random.sample`), for a total of 6,400 class draws; batches are drawn independently, so a class may recur across batches. The first 10,000 classes of the training split are also used for the coherence evaluation described below.

[2] Model-specific license: https://huggingface.co/microsoft/Phi-3-mini-4k-instruct/blob/main/LICENSE.

[3] Model-specific license: https://huggingface.co/Qwen/Qwen2.5-3B-Instruct/blob/main/LICENSE.

[4] Model-specific license: https://huggingface.co/meta-llama/Llama-3.2-3B-Instruct/blob/main/LICENSE.txt.

Paraphrase generation. For each training question, we construct a paraphrase class consisting of the original question and one generated paraphrase. Paraphrases are generated using `Qwen2.5-3B-Instruct` with a chat prompt instructing the model to rephrase the question without changing its meaning, preserving names and numbers, and returning only the rephrased question. Paraphrase generation uses greedy decoding, batch size 64, and maximum generation length 32 tokens. The same paraphrase set is used for all three models. Consequently, the setup is strictly single-model self-improvement for `Qwen2.5-3B-Instruct`, whose own paraphrasing ability defines the invariance mapping Φ, whereas for `Phi-3-mini-4k-instruct` and `Llama-3.2-3B-Instruct` the paraphrases are produced by a peer model of the same scale. We made this choice deliberately, so as to hold Φ, and hence the coherence set $\mathcal{C}_{\mathrm{coh}}$, fixed across architectures and isolate the effect of the training objective; the theory places no restriction on how Φ is obtained, as long as it is task-preserving. We return to the implications of this choice when discussing the results below.

Paraphrasing prompt template. The paraphraser is prompted with the following system and user messages, rendered with its chat template and an assistant-generation prefix. Here and in the templates below, braced fields denote substitutions and line breaks within a message are as shown.

System message:

```
You paraphrase questions.  Rephrase without changing meaning.  Do
not add or remove facts.  Preserve names and numbers.  Return only
the rephrased question.  Keep it short and fluent.
```

User message:

```
Paraphrase the following question without changing meaning.
{question}
```

Training objective and hyperparameters. Starting from each baseline model, we perform completion-level coherence training. The training objective combines a KL-divergence term that keeps the trained model close to the frozen baseline with a pairwise Jensen–Shannon coherence divergence between the model's output distributions on original and paraphrased questions. We set the coherence weight to $\lambda = 1$, giving the coherence-divergence term equal weight to the KL-divergence term. We do not tune λ; this value is fixed across all experiments. This objective is exactly the penalized form (5) of the relaxed coherence projection analyzed in Section 4.3, with F the negative entropy (so that B_F is the relative entropy) and $\mathsf{D} = \mathsf{D}_{\mathrm{JS}}$; it is thus derived from, rather than postulated independently of, the theory.

Training uses QLoRA with 4-bit NF4 quantization, double quantization, and bfloat16 computation. We use LoRA rank $r = 8$, LoRA scaling $\alpha = 16$, and LoRA dropout 0.05. The LoRA target modules are the standard projection and MLP modules detected in the model, including `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj` when present. We train using AdamW with learning rate 2×10^{-5}, weight decay 0.01, and gradient clipping at norm 1.0. The random seed is set to 42. During training, the sequence-level objective is estimated with REINFORCE-style samples; we use $n = 20$ samples per prompt, sampling temperature 0.7, top-$p = 0.9$, top-$k = 50$, and maximum generation length 48 tokens.

Sequence-level estimator. We now make the completion-level objective and its gradient estimator explicit. Let π_θ denote the trained model (LoRA parameters θ) and π_0 the frozen baseline; both are distributions over complete answer strings y given a prompt, so that all divergences below are *sequence-level*. A paraphrase class $\{x, x'\}$, $x' = \Phi(x)$, contributes the loss

$$\ell_\theta(x, x') = \tfrac{1}{2}[\mathsf{D}_{\mathrm{KL}}(\pi_\theta(x) \,\|\, \pi_0(x)) + \mathsf{D}_{\mathrm{KL}}(\pi_\theta(x') \,\|\, \pi_0(x'))] + \lambda\, \mathsf{D}_{\mathrm{JS}}(\pi_\theta(x), \pi_\theta(x')), \tag{17}$$

which is the empirical counterpart of (5) when $\mathcal{D}_{\mathcal{X}}$ places equal mass on the two prompts of a class: the trained model appears in the *first* argument of the KL term (as $\mathsf{B}_F(\pi \,\|\, \pi_0)$ with F the negative entropy), which is what makes on-policy sampling natural, and D_{JS} is symmetric. Writing $m_\theta(y) = \tfrac{1}{2}[\pi_\theta(y \mid x) + \pi_\theta(y \mid x')]$ for the mixture, every term in (17) is the expectation, under a

distribution of the model, of a log-ratio:

$$\begin{aligned}
\mathsf{D}_{\mathrm{KL}}(\pi_\theta(x) \,\|\, \pi_0(x)) &= \mathop{\mathbb{E}}_{y\sim\pi_\theta(\cdot|x)}\big[r_\theta^{\mathrm{KL}}(y;x)\big],\\
r_\theta^{\mathrm{KL}}(y;x) &= \log\pi_\theta(y\mid x) - \log\pi_0(y\mid x),\\
\mathsf{D}_{\mathrm{JS}}(\pi_\theta(x),\pi_\theta(x')) &= \tfrac{1}{2}\mathop{\mathbb{E}}_{y\sim\pi_\theta(\cdot|x)}\big[r_\theta^{\mathrm{JS}}(y;x,x')\big] + \tfrac{1}{2}\mathop{\mathbb{E}}_{y\sim\pi_\theta(\cdot|x')}\big[r_\theta^{\mathrm{JS}}(y;x',x)\big],\\
r_\theta^{\mathrm{JS}}(y;x,x') &= \log\pi_\theta(y\mid x) - \log m_\theta(y),
\end{aligned}$$

with $\log m_\theta(y) = \mathrm{logsumexp}(\log\pi_\theta(y\mid x), \log\pi_\theta(y\mid x')) - \log 2$. Given n completions $y_1,\dots,y_n$ sampled from $\pi_\theta(\cdot\mid x)$ and $y'_1,\dots,y'_n$ from $\pi_\theta(\cdot\mid x')$, each expectation is replaced by the corresponding sample average. The sequence log-probabilities are sums of token log-probabilities under teacher forcing: $\log\pi_0(y_j\mid x)$ is obtained from the frozen baseline, and $\log m_\theta(y_j)$ requires, in addition to $\log\pi_\theta(y_j\mid x)$, the *cross-prompt* score $\log\pi_\theta(y_j\mid x')$ of the same completion under the paraphrased prompt (and symmetrically for the y'_j); no clustering or normalization of the completions is involved. For a term of the form $\mathbb{E}_{y\sim\pi_\theta(\cdot|x)}[r_\theta(y)]$, in which both the sampling distribution and the integrand depend on θ, the gradient is

$$\nabla_\theta \mathop{\mathbb{E}}_{y\sim\pi_\theta(\cdot|x)}\big[r_\theta(y)\big] = \mathop{\mathbb{E}}_{y\sim\pi_\theta(\cdot|x)}\big[(r_\theta(y)-b)\,\nabla_\theta\log\pi_\theta(y\mid x) + \nabla_\theta r_\theta(y)\big], \tag{18}$$

for any constant b (score-function term plus pathwise term). For r^{KL} the pathwise term $\nabla_\theta\log\pi_\theta(y\mid x)$ has zero mean and the first term is the usual REINFORCE estimator of the KL to the baseline; for r^{JS} the pathwise term is the gradient through $\log\pi_\theta(y\mid x) - \log m_\theta(y)$, which involves both prompts. We estimate (18) by the average over the n completions of the prompt, with b the average of r_θ over these same completions, a standard variance-reduction baseline. The completions are drawn with the truncated, tempered sampler described above (temperature 0.7, top-p = 0.9, top-k = 50), while the log-probabilities entering the log-ratios and the score function are those of the model itself; the sampler thus acts as a proposal that concentrates the n samples on high-probability completions, a standard approximation in sequence-level fine-tuning. Finally, the per-class losses are averaged over the 32 classes of a batch, and one AdamW step is taken per batch.

Evaluation protocol. We evaluate both the baseline and coherence-trained models on the TriviaQA validation set using two standard answer-level metrics: exact match (EM) and token-level F1. We consider two decoding settings: greedy decoding and sampling. Sampling uses temperature 0.7 and top-p = 0.9. For both decoding settings, answers are generated with maximum length 24 tokens. We use a concise instruction-style prompt asking the model to answer with a short factual answer, such as a word, phrase, date, or name.

Exact match is a strict metric: a prediction receives score 1 if it exactly matches one of the ground-truth answers after normalization, and score 0 otherwise. Token-level F1 is a more lenient overlap-based metric: it computes precision and recall between the predicted answer and the ground-truth answer as bags of tokens, and reports their harmonic mean. We use SQuAD-style normalization, including lowercasing, punctuation removal, article removal, and whitespace normalization. For each example, we take the maximum EM and F1 score over all TriviaQA aliases.

QA prompt template. Questions are presented to the models with the following messages, rendered with each model's tokenizer chat template and an assistant-generation prefix.

> **System message:**
>
> ```
> You answer questions.  Give a concise factual answer (a
> word/phrase/date/name).
> ```
>
> **User message:**
>
> ```
> {question}
> ```

Coherence evaluation. We also directly evaluate semantic-level pairwise coherence. For each original question and its paraphrase, we decode an answer with greedy decoding and judge whether the two answers express the same factual answer. The judge is the corresponding frozen baseline model, so that it is identical for the baseline and the trained model and is not itself affected by training; it is prompted with the question and the two answers and must output either `same` or `different`.

Pairwise coherence is the fraction of answer pairs judged to be equivalent. We evaluate it on the first $10{,}000$ paraphrase classes of the training split. Two caveats are in order. First, these classes are drawn from the training split. Under the batch-sampling scheme described above, the expected number of evaluation classes selected at least once during training is

$$10{,}000\left[1-\left(1-\frac{32}{138{,}384}\right)^{200}\right] \simeq 452,$$

that is, 4.52% of the coherence-evaluation classes (an expectation under the sampling scheme rather than a measured overlap). For this small fraction of classes, the reported coherence gains combine in-sample reduction of the training penalty with generalization to unseen paraphrase pairs; in contrast, the accuracy results of Tables 1 and 2 are measured on the held-out validation split, which contains no paraphrases and was never used for training, and therefore measure out-of-sample improvement. Reporting pairwise coherence on paraphrases of the validation questions, together with a deterministic (SQuAD-normalized exact-agreement) coherence metric, is a natural complement that we leave to follow-up work. Second, the judge is a language model, and its errors are shared by the baseline and trained conditions; the Wilson intervals account for sampling variability but not for judge error.

Equivalence-judge prompt template. The frozen baseline judges a pair of answers using the following user message, greedy decoding, and a maximum of four generated tokens.

User message:

```
Question: {question}
Answer A: {answer_a}
Answer B: {answer_b}
Are Answer A and Answer B expressing the same answer?
Reply with exactly one word: same or different.
```

A response is counted as agreement if its lowercased text contains `same` and does not contain `different`; all other responses are counted as disagreement.

Statistical uncertainty. For F1, we report the mean together with the half-width of a 95% confidence interval. The confidence interval is computed using a normal approximation over examples: we estimate the standard error from the sample standard deviation of per-example F1 scores and use $z = 1.96$. For pairwise coherence, we report Wilson 95% confidence intervals for the binomial pairwise agreement rate.

Model	Decoding	Δ EM	Rel. EM (%)	Δ F1	Rel. F1 (%)
Phi-3	greedy	3.66	11.43	2.59	6.06
Phi-3	sampling	1.75	5.93	0.95	2.37
Qwen2.5	greedy	1.04	3.17	0.75	1.89
Qwen2.5	sampling	0.81	2.54	0.58	1.50
Llama-3.2	greedy	4.41	11.27	2.54	5.23
Llama-3.2	sampling	3.16	8.54	1.79	3.89

Table 2: Absolute and relative improvements of coherence-trained models over their corresponding baselines on TriviaQA. Relative improvements are computed with respect to the baseline score in the same model and decoding setting.

Coherence results. Table 3 shows that coherence training increases semantic pairwise agreement for all three models. The largest improvement is observed for `Phi-3-mini-4k-instruct`, where pairwise coherence increases from 0.533 to 0.608, a relative improvement of 14.07%. These results support the intended effect of the training objective: the trained models become more consistent across task-preserving paraphrases. Combined with the TriviaQA results in Tables 1 and 2, the results suggest that increasing coherence can improve consistency without sacrificing downstream accuracy, and in these experiments improves both.

On the smaller gains for `Qwen2.5`. `Qwen2.5-3B-Instruct` is the model with the smallest accuracy gain ($+1.04$ EM under greedy decoding) and the smallest coherence gain ($+0.019$); it is also the paraphrase generator, and the model with the lowest baseline coherence on these paraphrases

Model	Training	Pairwise coherence (95% CI)	Δ	Rel. improve (%)
Phi-3	baseline	0.533 ± 0.010	–	–
	trained	0.608 ± 0.010	0.075	14.07
Qwen2.5	baseline	0.472 ± 0.010	–	–
	trained	0.491 ± 0.010	0.019	4.03
Llama-3.2	baseline	0.655 ± 0.010	–	–
	trained	0.690 ± 0.010	0.035	5.34

Table 3: Semantic-level pairwise coherence before and after coherence training. Pairwise coherence is measured as the average agreement between answers to paraphrased versions of the same question, judged by the frozen baseline model. Results are computed on the first $10{,}000$ paraphrase classes from the training set using greedy decoding (see the caveats in the text). Confidence intervals are Wilson 95% intervals for the pairwise agreement rate.

(0.472). Several factors may contribute, which the present experiments cannot disentangle. When the paraphraser and the trained model coincide, the paraphrases are precisely those that the model itself produces as equivalent, and the incoherence they reveal may be concentrated on pairs that are harder to reconcile; for `Phi-3` and `Llama-3.2`, paraphrases produced by a different model may introduce surface variation that these models had not been exposed to and that is comparatively easy to make coherent, yielding larger measured gains. The three models also differ in their baseline accuracy and in how far they are from the coherent set, and Theorem 4.3 predicts a guaranteed gain that grows with the incoherence gap Δ_{coh} of the baseline as measured by the *training* divergence, not by the semantic agreement rate. A controlled comparison in which each model is trained both on its own paraphrases and on the shared set would isolate the paraphrase-source effect; we leave it to future work.

Baselines and ablations. We discuss the natural baselines and ablations of the training objective, several of which are degenerate in our label-free setting, and the non-degenerate label-free alternatives, in the light of the theory. (i) *KL-only training* ($\lambda = 0$) has π_0 itself as its minimizer and returns the baseline unchanged: the coherence term is the sole source of improvement, and the baseline row of each table is exactly this ablation. (ii) *Supervised fine-tuning* and *paraphrase data augmentation with labels* both require ground-truth answers and are therefore not comparable to a mechanism whose defining feature is that it uses no labels; standard supervised fine-tuning also contains no mechanism that reduces incoherence, and the substantial baseline incoherence in Table 3 (pairwise agreement between 47% and 66%) shows that the phenomenon addressed here is prevalent in instruction-trained models. (iii) *Unanchored consistency minimization* ($\lambda \to \infty$, i.e., minimizing the pairwise D_{JS} penalty without the KL term) is a non-degenerate label-free alternative operating on the same unlabeled pairs. The theory explains why the anchor matters: a model that outputs the same distribution for every prompt is perfectly coherent, so the penalty alone has many minimizers, most of them useless, and the improvement guarantee of Theorem 4.1 is a statement about the *projection* of π_0, i.e., about the coherent model closest to the baseline; the characterization of Theorem 6.1 shows that no other selection among coherent models enjoys the Pythagorean guarantee. In the penalized form, Remark C.1 shows that λ parametrizes the tolerance $\Lambda(\lambda)$ of the relaxed constraint, with the KL term enforcing proximity to π_0 for every λ; the unanchored limit corresponds to $\Lambda = 0$ with the proximity term dropped, for which collapse toward prompt-independent answers is the expected failure mode in practice. (iv) *Majority-vote pseudo-labeling* [Huang et al., 2023, Wang et al., 2023a], which samples answers across x and $\Phi(x)$ and fine-tunes on the consensus answer, is a hard-label, label-free alternative. In our terms, it replaces the projection onto $\mathcal{C}_{\mathrm{coh}}$ by a move toward a point mass on the majority answer; it is not a Bregman projection onto a convex set containing the ground truth, and it offers no improvement guarantee when the consensus is wrong, whereas the coherence projection never moves the model farther from any coherent ground truth (Theorem 4.1). The two mechanisms are complementary and can be combined. (v) *Inference-time orbit averaging*, $\widehat{\pi}_{\mathrm{orbit}}(x) = \frac{1}{2}[\pi_0(x) + \pi_0(\Phi(x))]$, is the mechanism analyzed in Theorem 5.2: it is the Euclidean projection onto $\mathcal{C}_{\mathrm{coh}}$, improves simultaneously for all divergences convex in their second argument, and requires no training. It is, however, an inference-time ensemble: it requires generating paraphrases and running the model on each of them at test time, multiplying inference cost by the orbit size, and it does not modify the model, whose incoherence persists on prompts for which no paraphrase is available. Coherence training amortizes this cost by moving the parameters themselves. Comparing the two on matched inference budgets is a natural experiment. Within the

proof-of-concept scope of this section, we report the comparison to the baseline, which is the quantity our guarantees bound. Empirical comparisons with (iii)–(v), matched-compute comparisons, larger models and more complex (e.g., reasoning) tasks are natural extensions; see also the larger-scale follow-up study of Mohri et al. [2026].

Compute resources. The experiments were run in Google Colab on a single NVIDIA A100-SXM4 GPU with 40GB of memory. The software stack consists of PyTorch (BSD-style license), 4-bit quantization through `bitsandbytes` (MIT License), PEFT/LoRA for parameter-efficient fine-tuning (Apache License 2.0), HuggingFace `transformers` and `datasets` (Apache License 2.0), and FlashAttention-2 (BSD 3-Clause License) when available; the timed run described next used Python 3.12.12 and PyTorch 2.9.0 with CUDA 12.6.

Execution time. To document execution time, we timed a rerun of the coherence training of `Qwen2.5-3B-Instruct` with the released code on the runtime described above. The rerun loaded the existing cache of 138,384 paraphrase classes and completed the 200 batches of 32 classes with 20 samples per prompt, but with a maximum generation length of 24 tokens instead of the 48 tokens used for the reported results. Loading the tokenizer and model and initializing QLoRA took about 40 seconds, loading TriviaQA about 10 seconds, and loading the paraphrase cache under one second. The training loop took 32,946 seconds, about 9.2 hours; this figure includes the training steps, batch sampling and timing-log writes, and excludes paraphrase generation, periodic validation, checkpoint writing and all evaluations. Since the sampling cost grows with the maximum generation length, these measurements provide a lower estimate, of roughly 9 A100 GPU-hours, of the training time of the reported 48-token `Qwen2.5` run; the two other models are of comparable size.

I Limitations

This work is mostly theoretical, with empirical evaluations intended as a proof of concept rather than a large-scale empirical study. Follow-up work [Mohri et al., 2026] has conducted larger-scale experiments on GSM8K and TriviaQA and also finds that coherence-based self-improvement can improve model performance. The main idea in both works is to use a lower-level model ability, such as recognizing or enforcing coherence across task-preserving transformations, to improve the model itself. We discuss the main limitations of the present framework.

Availability and quality of task-preserving transformations. Our guarantees require access to a coherence set that contains, or approximately contains, the ground-truth model. In applications, such a set must be constructed from task-preserving transformations, such as paraphrases, symmetries or logical-consistency constraints, and the quality of these transformations may vary across domains. If the transformations fail to preserve the task-relevant output distribution, enforcing coherence may introduce bias rather than improve the model. Theorem D.1 quantifies this trade-off, improvement persisting as long as the ground truth is closer to coherent than the displacement it induces, and Theorem 4.3 shows that relaxed constraints tolerate Λ-approximately coherent ground truths, but neither result removes the need for transformations of reasonable quality. In our experiments, the transformations are produced by a language model (the trained model itself for `Qwen2.5`, a peer model of the same scale for the other two models), which is practical but not infallible; moreover, our proof-of-concept evaluation measures pairwise coherence on paraphrases of training questions rather than on a held-out paraphrase set (Section H). Constructing transformations for tasks such as safety or multi-step reasoning (e.g., via monotonicity or logical-consistency constraints, which are convex) is a natural direction that we have not explored empirically.

Output-level analysis versus parametric fine-tuning. The improvement guarantees are stated for Bregman projections at the level of output distributions. The connection to parametric fine-tuning is made in two steps: the relaxed projection of Section 4.3 coincides, by Lagrangian duality, with the penalized training objective used in practice, and Theorem D.5 charges any suboptimality in this objective, whether due to the parametric class not containing the exact projection or to inexact, non-convex optimization, to the improvement guarantee with an explicit $\sqrt{\eta}$ dependence. What the theory does not provide is an analysis of the optimization dynamics themselves, i.e., of how small η can be made by a given training procedure; this remains an open and important question.

Scope of the divergences and of the guarantees. The guarantees are stated with respect to Bregman divergences and convex projection geometry. This is a broad and tractable framework, but it does not cover every evaluation metric used for language models or downstream decision systems. Our universal-improvement results require additional assumptions, such as convexity of the divergences in their second argument and feasibility of the orbitwise averaged model; when these fail, we show that simultaneous improvement across divergences can be impossible, so the choice of divergence remains important in general. Likewise, the exact rigidity results of Section G.3 require affine constraints and interior solutions, and the local rigidity results of Section G.4 require polyhedral constraints; extending them to general convex constraint sets is open.

Finite-sample and computational aspects. While the paper includes finite-sample and approximate-projection extensions, these results depend on generalization and suboptimality terms that may be difficult to estimate tightly in large-scale neural-network settings. The empirical projection problem may also be computationally challenging when the coherence constraints are numerous. The results should thus be viewed as providing a formal foundation for coherence-based self-improvement mechanisms rather than an immediately scalable training recipe for arbitrary large language models.