Lemma 2 proof

Credit: Claude-3-Opus-200k, GPT4

Explanation on $\pi+$:

$\pi+$ is defined as the solution to an optimization problem:

$\pi+$ = arg maxπ∈Δ(Y) {$\sum y$ $\pi(y)\delta(y)$ – $D\varphi(\pi,\pi-)$}.

Here's what each part of the equation means:

1. "arg max" stands for "arguments of the maximum." It means that $\pi+$ is the value of π that maximizes the expression inside the curly braces {}.
2. "π∈Δ(Y)" indicates that we are searching for $\pi+$ within the set Δ(Y), which is the set of all probability distributions over the set Y. In other words, $\pi+$ is a probability distribution that satisfies the constraints:
   - $\sum y$ $\pi+(y)$ = 1 (i.e., the probabilities sum up to 1)
   - $\pi+(y)$ ≥ 0 for all y ∈ Y (i.e., probabilities are non-negative)
3. The expression being maximized is "$\sum y$ $\pi(y)\delta(y)$ – $D\varphi(\pi,\pi-)$". It consists of two parts:
   - "$\sum y$ $\pi(y)\delta(y)$" is the dot product between the probability distribution π and a given vector δ. This term encourages $\pi+$ to assign high probabilities to the elements y where $\delta(y)$ is large.
   - "$D\varphi(\pi,\pi-)$" is the Bregman divergence between π and another fixed probability distribution $\pi-$. This term penalizes $\pi+$ for being too different from $\pi-$.

In summary, $\pi+$ is the probability distribution that maximizes the dot product with δ while staying close to $\pi-$ in terms of the Bregman divergence. The balance between these two objectives is determined by the relative magnitudes of δ and the Bregman divergence.

The definition of $\pi+$ is used in the lemma to prove an upper bound on the Bregman divergence between any probability distribution π and the optimum $\pi+$, in terms of the divergence between π and $\pi-$, the dot product of $\pi- - \pi$ with δ, and the ℓq norm of δ.

Explanation on *Lemma 7* in Fast Computation of Nash Equilibria in Imperfect Information Games:

ChatGPT

The statement that $-\delta + \nabla\Phi(\pi^{t+1}) - \nabla\Phi(\pi^t)$ forms an acute angle with $(\pi - \pi^{t+1})$ means that the inner product of these two vectors is non-negative:

$$[-\delta + \nabla\Phi(\pi^{t+1}) - \nabla\Phi(\pi^t)] \cdot (\pi - \pi^{t+1}) \geq 0$$

This condition arises from the optimality condition of the maximization problem that defines π^{t+1}. Specifically, π^{t+1} is chosen to maximize the following quantity:

$$\pi^{t+1} = \arg\max_{\pi \in \Delta(Y)} \left\{ \sum_{y \in Y} \pi(y)\delta(y) - D_\Phi(\pi, \pi^t) \right\}$$

The term $-\delta + \nabla\Phi(\pi^{t+1}) - \nabla\Phi(\pi^t)$ can be interpreted as the gradient of the objective function with respect to π at π^{t+1}, including the reward signal and the Bregman divergence terms. The fact that π^{t+1} maximizes the objective function implies that any move away from π^{t+1} within the feasible set $\Delta(Y)$ will result in a decrease in the objective function value, or equivalently, that the gradient of the objective function at π^{t+1} points in a direction that is "opposite" to the direction of $(\pi - \pi^{t+1})$.

Mathematically, an acute angle between two vectors a and b means that their dot product $a \cdot b$ is positive. If the vectors are orthogonal (form a right angle), their dot product is zero. If they form an obtuse angle, the dot product is negative. So, when we say the vectors form an acute angle, we mean that as we move from π^{t+1} in the direction of π, the objective function increases, which is true for a maximizer by the definition of convexity.

Therefore, the condition

$$[-\delta + \nabla\Phi(\pi^{t+1}) - \nabla\Phi(\pi^t)] \cdot (\pi - \pi^{t+1}) \geq 0$$

ensures that π^{t+1} is indeed a maximizer (or at least not worse than π^t) with respect to the objective function given by the rewards δ and the Bregman divergence from π^t. This is consistent with the properties of the Bregman divergence in a convex optimization setting, where the Bregman divergence from a point π^{t+1} is always non-negative, reaching its minimum (zero) at π^{t+1} itself.

GPT4 mentioned on the use of
https://en.wikipedia.org/wiki/Karush%E2%80%93Kuhn%E2%80%93Tucker_conditions

Note: In both *IPO-MD* and *Nash-MD-PG* , there is no requirement for *g(x)* inequality constraint.

Nonlinear optimization problem [edit]

Consider the following nonlinear optimization problem in standard form:

minimize $f(\mathbf{x})$

subject to

$g_i(\mathbf{x}) \leq 0,$

$h_j(\mathbf{x}) = 0.$

where $\mathbf{x} \in \mathbf{X}$ is the optimization variable chosen from a convex subset of $\mathbb{R}^n$, f is the objective or utility function, $g_i \ (i = 1, \ldots, m)$ are the inequality constraint functions and $h_j \ (j = 1, \ldots, \ell)$ are the equality constraint functions. The numbers of inequalities and equalities are denoted by m and ℓ respectively. Corresponding to the constrained optimization problem one can form the Lagrangian function

$$\mathcal{L}(\mathbf{x}, \mu, \lambda) = f(\mathbf{x}) + \mu^\top \mathbf{g}(\mathbf{x}) + \lambda^\top \mathbf{h}(\mathbf{x}) = L(\mathbf{x}, \alpha) = f(\mathbf{x}) + \alpha^\top \begin{pmatrix} \mathbf{g}(\mathbf{x}) \\ \mathbf{h}(\mathbf{x}) \end{pmatrix}$$

where

$$\mathbf{g}(\mathbf{x}) = \begin{bmatrix} g_1(\mathbf{x}) \\ \vdots \\ g_i(\mathbf{x}) \\ \vdots \\ g_m(\mathbf{x}) \end{bmatrix}, \quad \mathbf{h}(\mathbf{x}) = \begin{bmatrix} h_1(\mathbf{x}) \\ \vdots \\ h_j(\mathbf{x}) \\ \vdots \\ h_\ell(\mathbf{x}) \end{bmatrix}, \quad \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_m \end{bmatrix}, \quad \lambda = \begin{bmatrix} \lambda_1 \\ \vdots \\ \lambda_j \\ \vdots \\ \lambda_\ell \end{bmatrix} \quad \text{and} \quad \alpha = \begin{bmatrix} \mu \\ \lambda \end{bmatrix}.$$

The **Karush–Kuhn–Tucker theorem** then states the following.

Theorem — (sufficiency) If $(\mathbf{x}^*, \alpha^*)$ is a saddle point of $L(\mathbf{x}, \alpha)$ in $\mathbf{x} \in \mathbf{X}$, $\mu \geq \mathbf{0}$, then $\mathbf{x}^*$ is an optimal vector for the above optimization problem.

(necessity) Suppose that $f(\mathbf{x})$ and $g_i(\mathbf{x})$, $i = 1, \ldots, m$, are convex in $\mathbf{X}$ and that there exists $\mathbf{x}_0 \in \operatorname{relint}(\mathbf{X})$ such that $\mathbf{g}(\mathbf{x}_0) < \mathbf{0}$ (i.e., Slater's condition holds). Then with an optimal vector $\mathbf{x}^*$ for the above optimization problem there is associated a vector $\alpha^* = \begin{bmatrix} \mu^* \\ \lambda^* \end{bmatrix}$ satisfying $\mu^* \geq \mathbf{0}$ such that $(\mathbf{x}^*, \alpha^*)$ is a saddle point of $L(\mathbf{x}, \alpha)$.[5]

Since the idea of this approach is to find a supporting hyperplane on the feasible set $\mathbf{\Gamma} = \{\mathbf{x} \in \mathbf{X} : g_i(\mathbf{x}) \leq 0, i = 1, \ldots, m\}$, the proof of the Karush–Kuhn–Tucker theorem makes use of the hyperplane separation theorem.[6]

The system of equations and inequalities corresponding to the KKT conditions is usually not solved directly, except in the few special cases where a closed-form solution can be derived analytically. In general, many optimization algorithms can be interpreted as methods for numerically solving the KKT system of equations and inequalities.[7]

$\pi^+ = \arg\max_{\pi \in \Delta(Y)} \left\{ \sum_{y \in Y} \pi(y)\delta(y) - D_\Phi(\pi, \pi^-) \right\}$

The Lagrangian for this optimization problem that incorporates the simplex constraint would indeed typically take the form:

$$L(\pi, \lambda) = \sum_{y \in Y} \pi(y)\delta(y) - D_\Phi(\pi, \pi^-) - \lambda \left(\sum_{y \in Y} \pi(y) - 1 \right)$$

Here, $\delta(y)$ is the immediate reward associated with action y, and λ is the Lagrange multiplier enforcing the probability simplex constraint.

When we take the gradient of $L(\pi, \lambda)$ with respect to π, we would consider the gradient of each component of the Lagrangian:

- The gradient of $\sum_{y \in Y} \pi(y)\delta(y)$ with respect to π is simply δ.
- The gradient of $D_\Phi(\pi, \pi^-)$ with respect to π is $\nabla\Phi(\pi) - \nabla\Phi(\pi^-)$, as previously discussed.
- The gradient of the term involving the equality constraint $\lambda \left(\sum_{y \in Y} \pi(y) - 1 \right)$ is $-\lambda\mathbf{1}$.

Putting it together, the gradient of $L(\pi, \lambda)$ is:

$$\nabla_\pi L(\pi, \lambda) = \delta - (\nabla\Phi(\pi) - \nabla\Phi(\pi^-)) - \lambda\mathbf{1}$$

For the optimal policy π^+, the stationarity condition of the KKT requirements dictates:

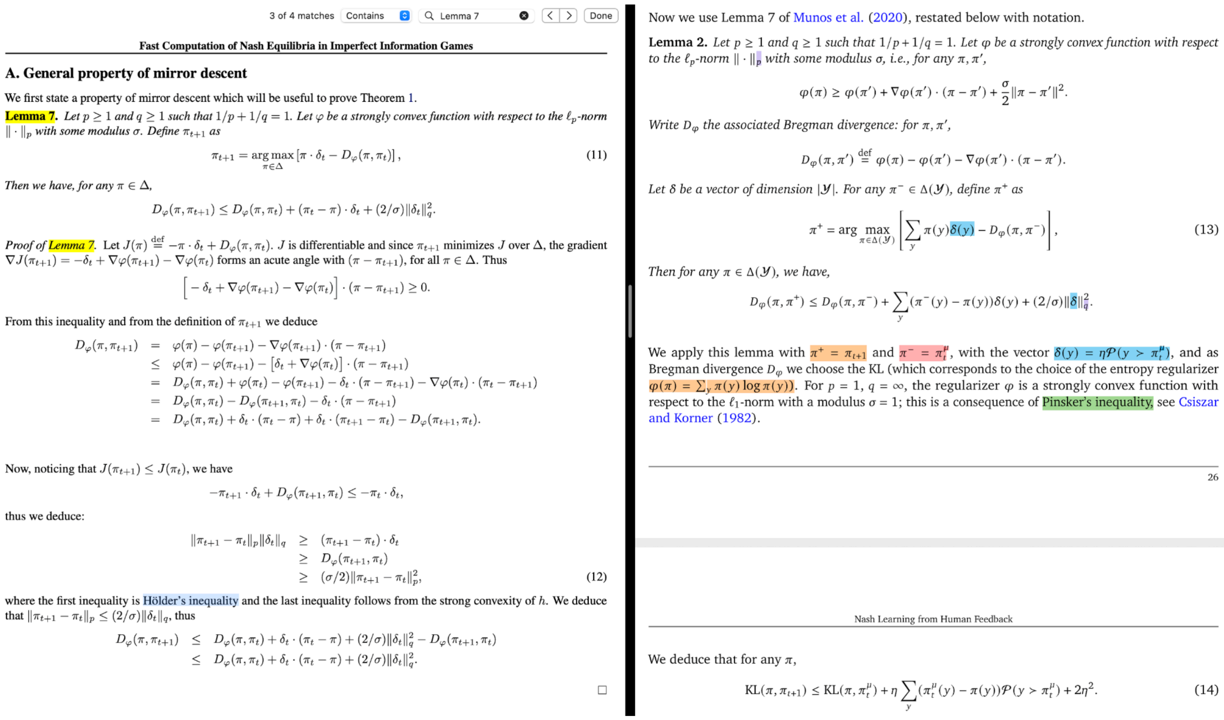
$$\nabla_\pi L(\pi^+, \lambda) = 0$$

Which results in:

$$\delta - (\nabla\Phi(\pi^+) - \nabla\Phi(\pi^-)) - \lambda\mathbf{1} = 0$$

This gives us the optimality condition for π^+:

$$\nabla\Phi(\pi^+) = \delta + \nabla\Phi(\pi^-) + \lambda\mathbf{1}$$



3 of 4 matches Contains Lemma 7 Done

Fast Computation of Nash Equilibria in Imperfect Information Games

A. General property of mirror descent

We first state a property of mirror descent which will be useful to prove Theorem 1.

Lemma 7. *Let $p \geq 1$ and $q \geq 1$ such that $1/p + 1/q = 1$. Let φ be a strongly convex function with respect to the ℓ_p-norm $\|\cdot\|_p$ with some modulus σ. Define π_{t+1} as*

$$\pi_{t+1} = \arg\max_{\pi \in \Delta} \left[\pi \cdot \delta_t - D_\varphi(\pi, \pi_t)\right], \tag{11}$$

Then we have, for any $\pi \in \Delta$,

$$D_\varphi(\pi, \pi_{t+1}) \leq D_\varphi(\pi, \pi_t) + (\pi_t - \pi) \cdot \delta_t + (2/\sigma)\|\delta_t\|_q^2.$$

Proof of Lemma 7. Let $J(\pi) \stackrel{\text{def}}{=} -\pi \cdot \delta_t + D_\varphi(\pi, \pi_t)$. J is differentiable and since π_{t+1} minimizes J over Δ, the gradient $\nabla J(\pi_{t+1}) = -\delta_t + \nabla\varphi(\pi_{t+1}) - \nabla\varphi(\pi_t)$ forms an acute angle with $(\pi - \pi_{t+1})$, for all $\pi \in \Delta$. Thus

$$\Big[-\delta_t + \nabla\varphi(\pi_{t+1}) - \nabla\varphi(\pi_t)\Big] \cdot (\pi - \pi_{t+1}) \geq 0.$$

From this inequality and from the definition of π_{t+1} we deduce

$$\begin{aligned} D_\varphi(\pi, \pi_{t+1}) &= \varphi(\pi) - \varphi(\pi_{t+1}) - \nabla\varphi(\pi_{t+1}) \cdot (\pi - \pi_{t+1}) \\ &\leq \varphi(\pi) - \varphi(\pi_{t+1}) - \big[\delta_t + \nabla\varphi(\pi_t)\big] \cdot (\pi - \pi_{t+1}) \\ &= D_\varphi(\pi, \pi_t) + \varphi(\pi_t) - \varphi(\pi_{t+1}) - \delta_t \cdot (\pi - \pi_{t+1}) - \nabla\varphi(\pi_t) \cdot (\pi_t - \pi_{t+1}) \\ &= D_\varphi(\pi, \pi_t) - D_\varphi(\pi_{t+1}, \pi_t) - \delta_t \cdot (\pi - \pi_{t+1}) \\ &= D_\varphi(\pi, \pi_t) + \delta_t \cdot (\pi_t - \pi) + \delta_t \cdot (\pi_{t+1} - \pi_t) - D_\varphi(\pi_{t+1}, \pi_t). \end{aligned}$$

Now, noticing that $J(\pi_{t+1}) \leq J(\pi_t)$, we have

$$-\pi_{t+1} \cdot \delta_t + D_\varphi(\pi_{t+1}, \pi_t) \leq -\pi_t \cdot \delta_t,$$

thus we deduce:

$$\begin{aligned} \|\pi_{t+1} - \pi_t\|_p \|\delta_t\|_q &\geq (\pi_{t+1} - \pi_t) \cdot \delta_t \\ &\geq D_\varphi(\pi_{t+1}, \pi_t) \\ &\geq (\sigma/2)\|\pi_{t+1} - \pi_t\|_p^2, \end{aligned} \tag{12}$$

where the first inequality is Hölder's inequality and the last inequality follows from the strong convexity of h. We deduce that $\|\pi_{t+1} - \pi_t\|_p \leq (2/\sigma)\|\delta_t\|_q$, thus

$$\begin{aligned} D_\varphi(\pi, \pi_{t+1}) &\leq D_\varphi(\pi, \pi_t) + \delta_t \cdot (\pi_t - \pi) + (2/\sigma)\|\delta_t\|_q^2 - D_\varphi(\pi_{t+1}, \pi_t) \\ &\leq D_\varphi(\pi, \pi_t) + \delta_t \cdot (\pi_t - \pi) + (2/\sigma)\|\delta_t\|_q^2. \end{aligned}$$

□

Now we use Lemma 7 of Munos et al. (2020), restated below with notation.

Lemma 2. *Let $p \geq 1$ and $q \geq 1$ such that $1/p + 1/q = 1$. Let φ be a strongly convex function with respect to the ℓ_p-norm $\|\cdot\|_p$ with some modulus σ, i.e., for any π, π',*

$$\varphi(\pi) \geq \varphi(\pi') + \nabla\varphi(\pi') \cdot (\pi - \pi') + \frac{\sigma}{2}\|\pi - \pi'\|^2.$$

Write D_φ the associated Bregman divergence: for π, π',

$$D_\varphi(\pi, \pi') \stackrel{\text{def}}{=} \varphi(\pi) - \varphi(\pi') - \nabla\varphi(\pi') \cdot (\pi - \pi').$$

Let δ be a vector of dimension $|\mathcal{Y}|$. For any $\pi^- \in \Delta(\mathcal{Y})$, define π^+ as

$$\pi^+ = \arg\max_{\pi \in \Delta(\mathcal{Y})} \left[\sum_y \pi(y)\delta(y) - D_\varphi(\pi, \pi^-)\right], \tag{13}$$

Then for any $\pi \in \Delta(\mathcal{Y})$, we have,

$$D_\varphi(\pi, \pi^+) \leq D_\varphi(\pi, \pi^-) + \sum_y (\pi^-(y) - \pi(y))\delta(y) + (2/\sigma)\|\delta\|_q^2.$$

We apply this lemma with $\pi^+ = \pi_{t+1}$ and $\pi^- = \pi_t^\mu$, with the vector $\delta(y) = \eta\mathcal{P}(y > \pi_t^\mu)$, and as Bregman divergence D_φ we choose the KL (which corresponds to the choice of the entropy regularizer $\varphi(\pi) = \sum_y \pi(y)\log\pi(y)$). For $p = 1$, $q = \infty$, the regularizer φ is a strongly convex function with respect to the ℓ_1-norm with a modulus $\sigma = 1$; this is a consequence of Pinsker's inequality, see Csiszar and Korner (1982).

26

Nash Learning from Human Feedback

We deduce that for any π,

$$\mathrm{KL}(\pi, \pi_{t+1}) \leq \mathrm{KL}(\pi, \pi_t^\mu) + \eta\sum_y (\pi_t^\mu(y) - \pi(y))\mathcal{P}(y > \pi_t^\mu) + 2\eta^2. \tag{14}$$

In the screenshot, both left and right sides are using different equations for π+ if we compare their second term notation a bit more carefully. 👀 but they are still using the same logic, it is just that the right side is replacing π_t with π_μ_t given that the right side is a newer paper introducing the alternative mixture policy.

Equation (11) π+ = arg maxπ∈Δ(Y) {∑y $\pi(y)\delta(y)$ − $D\varphi(\pi$,π_t)}

versus

Equation (13) π+ = arg maxπ∈Δ(Y) {∑y $\pi(y)\delta(y)$ − $D\varphi(\pi$,π_μ_t)}

For Lemma 7 proof, the first inequality line of equation (12) is just a direct application of Hölder's inequality, while the last inequality line of equation (12) about the strong convexity is explained by GPT4 below:

Note: For those interested in understanding more on Hölder's inequality, see Young's inequality for products

ChatGPT

The inequality $D_\Phi(\pi_{t+1}, \pi_t) \geq \frac{\alpha}{2}\|\pi_{t+1} - \pi_t\|_p^2$ arises from the strong convexity of the function Φ with respect to the ℓ_p-norm. A function Φ is strongly convex with modulus α if, for any two points x and y, the following condition holds:

$$\Phi(x) \geq \Phi(y) + \nabla\Phi(y) \cdot (x - y) + \frac{\alpha}{2}\|x - y\|_p^2$$

In the context of the Bregman divergence, $D_\Phi(x, y)$ is defined as:

$$D_\Phi(x, y) = \Phi(x) - \Phi(y) - \nabla\Phi(y) \cdot (x - y)$$

If we compare the definition of strong convexity with the definition of Bregman divergence, we can see that the strong convexity condition essentially provides a lower bound on the Bregman divergence. Specifically, the term $\frac{\alpha}{2}\|x - y\|_p^2$ in the strong convexity condition becomes a lower bound for the Bregman divergence:

$$D_\Phi(x, y) \geq \frac{\alpha}{2}\|x - y\|_p^2$$

In equation (12) from your document, π_{t+1} and π_t play the roles of x and y, respectively, and thus we have:

$$D_\Phi(\pi_{t+1}, \pi_t) \geq \frac{\alpha}{2}\|\pi_{t+1} - \pi_t\|_p^2$$

This means that the Bregman divergence between π_{t+1} and π_t is at least as large as half the modulus of strong convexity α times the ℓ_p-norm of the difference between π_{t+1} and π_t, squared. This lower bound is used in the proof of Lemma 7 to relate the change in Φ due to a change from π_t to π_{t+1} to the actual change in the policies as measured by the ℓ_p-norm.

Using the first **inequality** line of equation (12) as well as the last line of **equality** expression of *DΦ(π_t+1, π_t)* , we could obtain the final two **inequality** of *DΦ(π_t+1, π_t)* for Lemma 7.

Now that we have proven Lemma 7 as described above, we are halfway done between equation (13) and equation (14) within Lemma 2 proof after we replace *π*_t with *π_μ*_t within the final inequality expression of *DΦ(π_t+1, π_t)* from Lemma 7.

As for deriving equation (14) in Lemma 2, see the next two screenshots from Claude3 and GPT4 replies:

Claude-3-Opus-200k Poe

I apologize for the confusion. Let me provide a more detailed explanation of how we obtain that inequality.

We start with the result from Lemma 2:

$D\varphi(\pi, \pi+) \leq D\varphi(\pi, \pi-) + \sum y(\pi-(y) - \pi(y))\delta(y) + (2/\sigma)\|\delta\|2q.$

To apply this lemma to our context, we make the following substitutions:

1. $\pi+ = \pi t+1$: This is the policy we are trying to bound in terms of KL divergence.
2. $\pi- = \pi\mu t$: This is the regularized policy defined in the paper.
3. $\delta(y) = \eta\mathrm{P}(y \succ \pi\mu t)$: This is the specific choice of δ used in the paper.
4. $D\varphi(\cdot,\cdot) = \mathrm{KL}(\cdot,\cdot)$: We choose the KL divergence as the Bregman divergence.

With these substitutions, the left-hand side of the inequality becomes:

$D\varphi(\pi, \pi+) = \mathrm{KL}(\pi, \pi t+1).$

The right-hand side of the inequality becomes:

$D\varphi(\pi, \pi-) + \sum y(\pi-(y) - \pi(y))\delta(y) + (2/\sigma)\|\delta\|2q$
$= \mathrm{KL}(\pi, \pi\mu t) + \sum y(\pi\mu t(y) - \pi(y))(\eta\mathrm{P}(y \succ \pi\mu t)) + (2/\sigma)\|\eta\mathrm{P}(\cdot \succ \pi\mu t)\|2q$
$= \mathrm{KL}(\pi, \pi\mu t) + \eta \sum y(\pi\mu t(y) - \pi(y))\mathrm{P}(y \succ \pi\mu t) + (2/\sigma)\|\eta\mathrm{P}(\cdot \succ \pi\mu t)\|2q.$

Therefore, applying Lemma 2 with these substitutions directly gives us:

$\mathrm{KL}(\pi, \pi t+1) \leq \mathrm{KL}(\pi, \pi\mu t) + \eta \sum y(\pi\mu t(y) - \pi(y))\mathrm{P}(y \succ \pi\mu t) + (2/\sigma)\|\eta\mathrm{P}(\cdot \succ \pi\mu t)\|2q,$

which is the inequality we wanted to derive.

I hope this clarifies the steps involved in obtaining this inequality from Lemma 2. Let me know if you have any further questions!

Let's go through each of the steps briefly to understand how the last term $2\eta^2$ is derived in equation (14):

1. **Lemma 2 Application**:
   Lemma 2 provides an upper bound on the Bregman divergence (in this case, the KL divergence) when the policy is updated. This lemma is applied with $\pi^+ = \pi_{t+1}$ and $\pi^- = \pi_{\mu t}$, with $\delta(y)$ representing the scaled immediate reward vector $\eta P(y \succ \pi_{\mu t})$.
2. **Hölder's Inequality**:
   Hölder's inequality is then applied to $\|\eta P(\cdot \succ \pi_{\mu t})\|_q$ to obtain an upper bound for the norm. Since probability values are bounded by 1, the term simplifies to $\eta |Y|^{1/q}$, where $|Y|$ is the size of the action space and q is the conjugate of the p-norm used in the context (where $1/p + 1/q = 1$).
3. **Simplifying the Bound**:
   With $\sigma = 1$ and $q = \infty$, the expression $|Y|^{2/q}$ equals 1, which simplifies the bound to $(2/\sigma)\eta^2|Y|^{2/q}$, and hence simplifies further to $2\eta^2$ since σ is 1 and $|Y|^{2/q}$ is 1 by the choice of q.
4. **Final Substitution**:
   Substituting this result back into the inequality obtained from Lemma 2, you get the desired form of equation (14):

$$KL(\pi, \pi_{t+1}) \leq KL(\pi, \pi_{\mu t}) + \eta \sum_{y \in Y} (\pi_{\mu t}(y) - \pi(y)) P(y \succ \pi_{\mu t}) + 2\eta^2$$

This term $2\eta^2$ essentially represents a bound on the variance of the updates to the policy, accounting for the maximum possible deviation due to the stochastic nature of the update process, and hence completes the derivation of equation (14) from the NLHF paper.

Lemma 2 needs Pinsker's inequality for the following text snippet within Lemma 2.

> For p = 1, q = ∞, the regularizer φ is a strongly convex function with respect to the ℓ1-norm with a modulus σ = 1; this is a consequence of Pinsker's inequality, see Csiszar and Korner (1982).

$$\|\pi - \pi'\|_{TV} \leq \sqrt{\tfrac{1}{2}KL(\pi||\pi')}$$

Here $\|\pi - \pi'\|_{TV}$ denotes the total variation distance, which for probability distributions is defined as $\frac{1}{2}\|\pi - \pi'\|_1$, because total variation distance is half the ℓ_1-norm of their difference. Therefore, Pinsker's inequality can be expressed as:

$$\tfrac{1}{2}\|\pi - \pi'\|_1 \leq \sqrt{\tfrac{1}{2}KL(\pi||\pi')}$$

Squaring both sides gives us:

$$\tfrac{1}{4}\|\pi - \pi'\|_1^2 \leq \tfrac{1}{2}KL(\pi||\pi')$$

Multiplying by 2 to get rid of the 1/4, we obtain:

$$\|\pi - \pi'\|_1^2 \leq 2KL(\pi||\pi')$$

Using Pinsker's inequality to provide a lower bound for the KL divergence gives us:

$$KL(\pi||\pi') \geq \tfrac{1}{2}\|\pi - \pi'\|_1^2$$

So using this above Pinsker's inequality or lower bound for the KL divergence, if we refer below to the first two expressions just before equation (13) inside Lemma 2, and then as instructed by Lemma 2 text, we choose KL divergence as Bregman divergence $D\varphi$, we could actually see and prove the strong convexity, by just rearranging the terms for the second expression.

Lemma 2. *Let $p \geq 1$ and $q \geq 1$ such that $1/p + 1/q = 1$. Let φ be a strongly convex function with respect to the ℓ_p-norm $\|\cdot\|_p$ with some modulus σ, i.e., for any π, π',*

$$\varphi(\pi) \geq \varphi(\pi') + \nabla\varphi(\pi') \cdot (\pi - \pi') + \frac{\sigma}{2}\|\pi - \pi'\|^2.$$

Write D_φ the associated Bregman divergence: for π, π',

$$D_\varphi(\pi, \pi') \overset{\text{def}}{=} \varphi(\pi) - \varphi(\pi') - \nabla\varphi(\pi') \cdot (\pi - \pi').$$

We deduce that for any π,

$$\mathrm{KL}(\pi, \pi_{t+1}) \leq \mathrm{KL}(\pi, \pi_t^\mu) + \eta \sum_y (\pi_t^\mu(y) - \pi(y)) \mathcal{P}(y \succ \pi_t^\mu) + 2\eta^2. \tag{14}$$

For the choice $\pi = \pi_\tau^*$ and using the previous lemma, we have

$$\begin{aligned}
\mathrm{KL}(\pi_\tau^*, \pi_{t+1}) &\leq \mathrm{KL}(\pi_\tau^*, \pi_t^\mu) + \eta \sum_y (\pi_t^\mu(y) - \pi_\tau^*(y)) \mathcal{P}(y \succ \pi_t^\mu) + 2\eta^2 \\
&\leq (1-\eta\tau)\mathrm{KL}(\pi_\tau^*, \pi_t) + \eta\tau \left(\mathrm{KL}(\pi_\tau^*, \mu) - \mathrm{KL}(\pi_t^\mu, \mu)\right) \\
&\quad + \eta \left(\mathcal{P}(\pi_t^\mu \succ \pi_t^\mu) - \mathcal{P}(\pi_\tau^* \succ \pi_t^\mu)\right) + 2\eta^2 \\
&= (1-\eta\tau)\mathrm{KL}(\pi_\tau^*, \pi_t) + \eta \left[1/2 - \mathcal{P}(\pi_\tau^* \succ \pi_t^\mu) + \tau \mathrm{KL}(\pi_\tau^*, \mu) - \tau \mathrm{KL}(\pi_t^\mu, \mu)\right] + 2\eta^2 \\
&= (1-\eta\tau)\mathrm{KL}(\pi_\tau^*, \pi_t) + \eta \left[1/2 - \mathcal{P}_\tau(\pi_\tau^* \succ \pi_t^\mu)\right] + 2\eta^2 \\
&\leq (1-\eta\tau)\mathrm{KL}(\pi_\tau^*, \pi_t) + 2\eta^2,
\end{aligned}$$

where the last inequality comes from the fact that π_τ^* is the Nash of the regularized game $\mathcal{P}_\tau$: $\mathcal{P}_\tau(\pi_\tau^* \succ \pi_t^\mu) \geq \mathcal{P}_\tau(\pi_\tau^* \succ \pi_\tau^*) = 1/2$ and the last equality comes from the definition of the regularized preference.

By iterating this inequality we deduce

$$\begin{aligned}
\mathrm{KL}(\pi_\tau^*, \pi_T) &\leq 2\eta^2 \sum_{t=0}^{T-1} (1-\eta\tau)^i + (1-\eta\tau)^T \mathrm{KL}(\pi_\tau^*, \pi_0) \\
&\leq 2\frac{\eta}{\tau} + e^{-\eta\tau T} \mathrm{KL}(\pi_\tau^*, \pi_0).
\end{aligned}$$

We deduce that for the choice $\eta = \log T / (\tau T)$ we have

$$\mathrm{KL}(\pi_\tau^*, \pi_T) \leq \left(\mathrm{KL}(\pi_\tau^*, \pi_0) + \frac{2 \log T}{\tau^2}\right) \frac{1}{T}.$$

Just right after equation (14) , regarding the second green inequality line: KL($\pi*\tau$, π_μt) ≤ *(1 − $\eta\tau$)KL($\pi*\tau$, πt) + $\eta\tau$(KL($\pi*\tau$, μ) − KL(π_μt, μ)*), it comes from Lemma 1 in the paper.

If we substitute π = $\pi*\tau$ in Lemma 1, we get: KL($\pi*\tau$, π_μt) ≤ *$\eta\tau$KL($\pi*\tau$, μ) + (1 − $\eta\tau$)KL($\pi*\tau$, πt) − $\eta\tau$KL(π_μt, μ), which can be rearranged to: KL($\pi*\tau$, π_μt) ≤ (1 − $\eta\tau$)KL($\pi*\tau$, πt) + $\eta\tau$(KL($\pi*\tau$, μ) − KL(π_μt, μ*)). This is the inequality used in the proof of Theorem 1.

As for the last yellow line of the inequality $KL(\pi*\tau, \pi t+1)$, $1/2 - P\tau(\pi*\tau > \pi\_\mu t)$ becomes negative where we use the following logic:

Claude-3-Opus-200k Poe

The statement "where the last inequality comes from the fact that $\pi*\tau$ is the Nash of the regularized game $P\tau$: $P\tau(\pi*\tau \succ \pi\mu t) \geq P\tau(\pi*\tau \succ \pi*\tau) = 1/2$" is based on the properties of the Nash equilibrium in the regularized preference model.

Let's break it down:

1. $\pi*\tau$ is defined as the Nash equilibrium of the regularized preference model $P\tau$. This means that $\pi*\tau$ is the policy that maximizes its preference against any other policy, i.e., for any policy π,

   $P\tau(\pi*\tau \succ \pi) \geq P\tau(\pi \succ \pi*\tau)$.
2. In particular, if we consider $\pi = \pi\mu t$, we have:

   $P\tau(\pi*\tau \succ \pi\mu t) \geq P\tau(\pi\mu t \succ \pi*\tau)$.
3. Since the regularized preference model is symmetric, i.e., $P\tau(\pi \succ \pi') = 1 - P\tau(\pi' \succ \pi)$, we can rewrite the above inequality as:

   $P\tau(\pi*\tau \succ \pi\mu t) \geq 1 - P\tau(\pi*\tau \succ \pi\mu t)$.
4. This implies that $P\tau(\pi*\tau \succ \pi\mu t) \geq 1/2$.
5. Now, consider the case where $\pi = \pi*\tau$. By the definition of the regularized preference model (Equation 2 in the paper),

   $P\tau(\pi*\tau \succ \pi*\tau) = P(\pi*\tau \succ \pi*\tau) - \tau KL(\pi*\tau, \mu) + \tau KL(\pi*\tau, \mu) = P(\pi*\tau \succ \pi*\tau)$.

   Since $P(\pi*\tau \succ \pi*\tau) = 1/2$ (a policy is equally preferred to itself), we have $P\tau(\pi*\tau \succ \pi*\tau) = 1/2$.

Combining points 4 and 5, we get:

$P\tau(\pi*\tau \succ \pi\mu t) \geq P\tau(\pi*\tau \succ \pi*\tau) = 1/2$,

which is the inequality used in the proof of Theorem 1.

Regarding the second last inequality line of $KL(\pi*\tau, \pi\_T)$, see the following GPT4 derivation:

1. Start with $KL(\pi_\tau^*, \pi_T)$ and apply the inequality recursively:
   $KL(\pi_\tau^*, \pi_T) \leq (1-\eta\tau)KL(\pi_\tau^*, \pi_{T-1}) + 2\eta^2$
   $\leq (1-\eta\tau)^2 KL(\pi_\tau^*, \pi_{T-2}) + 2\eta^2(1+(1-\eta\tau))$
   $\leq \ldots$
   $\leq (1-\eta\tau)^T KL(\pi_\tau^*, \pi_0) + 2\eta^2 \sum_{t=0}^{T-1}(1-\eta\tau)^t$
2. Recognize the geometric series in the last term:
   $\sum_{t=0}^{T-1}(1-\eta\tau)^t$

   The sum of a geometric series with T terms, a first term a, and a ratio r is given by:
   $\sum_{t=0}^{T-1} ar^t = a\frac{1-r^T}{1-r}$

   For our case, with $a = 1$ and $r = 1 - \eta\tau$, this becomes:
   $\sum_{t=0}^{T-1}(1-\eta\tau)^t = \frac{1-(1-\eta\tau)^T}{\eta\tau}$

   Using the approximation $(1-\eta\tau)^T \approx e^{-\eta\tau T}$ for small $\eta\tau$, we have:
   $\sum_{t=0}^{T-1}(1-\eta\tau)^t = \frac{1-e^{-\eta\tau T}}{\eta\tau}$

1. Substitute this back into the inequality:
   $KL(\pi_\tau^*, \pi_T) \leq (1-\eta\tau)^T KL(\pi_\tau^*, \pi_0) + 2\eta^2 \frac{1-e^{-\eta\tau T}}{\eta\tau}$
   $= e^{-\eta\tau T} KL(\pi_\tau^*, \pi_0) + 2\eta\frac{1-e^{-\eta\tau T}}{\tau}$
   $\leq e^{-\eta\tau T} KL(\pi_\tau^*, \pi_0) + 2\eta/\tau$

   This derives the term $2\eta/\tau + e^{-\eta\tau T} KL(\pi_\tau^*, \pi_0)$, accounting for both the accumulated effect of the updates over time and the exponential decay of the initial divergence due to the learning rate.

The NLHF paper's choice of $\eta = \log T/(\tau T)$ does not apply for all situations especially given more and more new papers such as [Direct Nash Optimization: Teaching Language Models to Self-Improve with General Preferences] as well as [A Theoretical Analysis of Nash Learning from Human Feedback under General KL-Regularized Preference]