A FAIRNESS DERIVATION FOR NUM PROBLEM

Now, based on this primal problem, we introduce the Lagrangian in the following Eq.

$$L(x,\lambda,\mu) = -\sum_{i=1}^{n} U_i(x_i) + \mu_0((\sum_{i=1}^{n} x_i) - C) + \sum_{i=1}^{n} \mu_i(-x_i), \quad \text{where} \quad \mu_i \geq 0. \tag{14}$$

We have defined the primal optimization problem and the associated Lagrangian. and We can explore the KKT conditions for this problem. These conditions will allow us to draw strong conclusions about the nature of the solution and ultimately solve the problem.

A.1 KKT Conditions:

Primal Feasibility:

$$x_i^* \geq 0, \quad \sum_{i=1}^{n} x_i^* \leq C. \tag{15}$$

Dual Feasibility:

$$\mu_0^* \geq 0, \quad \mu_i^* \geq 0. \tag{16}$$

Complementary Slackness:

$$\mu_0^*(\sum_{i=1}^{n} x_i^* - C) = 0, \quad \mu_i^* x_i^* = 0. \tag{17}$$

Lagrange Multiplier Theorem:

$$\begin{aligned} \text{general form:} \quad & \nabla_x f(x^*) + \sum_{i=1}^{m} \lambda_i^* \nabla_x h_i(x^*) + \sum_{j=1}^{l} \mu_j^* \nabla_x g_j(x^*) \\ \text{for this problem:} \quad & -U_i'(x_i^*) + \mu_0^* - \mu_i^* = 0 \quad \text{(Key Eq.)} \end{aligned} \tag{18}$$

Now that we have defined the KKT conditions for this problem, we can go through different possible solution cases to gain some insight into the nature of the optimal solution.

A.2 Case 1: if $x_i^* \geq 0$, then, $\mu_i^* = 0$

According to the Lagrange Multiplier Theorem:

$$U_i'(x_i^*) = \mu_0^* \tag{19}$$

Thus, for flows i, j where $x_i^* \geq 0$ and $x_j^* \geq 0$:

$$U_i'(x_i^*) = U_j'(x_j^*) \tag{20}$$

The interpretation of this result is that the "marginal utility" is the same at the optimal allocation for two flows that got assigned a non-zero rate.

A.3 Case 2: if $(\sum_{i=1}^{n} x_i^*) - C < 0$, then: $\mu_0^* = 0$

Also, by Case 1, if $x_i^* > 0$, then:

$$U_i'(x_i^*) = 0. \tag{21}$$

The interpretation of this result is that if we do not use the total capacity, then each flow with a positive rate should have zero marginal utility at the optimal allocation. Otherwise, we could use our excess capacity to increase utility.

A.4 General Non-Trivial Case: $\mu_0^* > 0, \quad \sum_{i=1}^{n} x_i^* = C$

According to the Key Eq., we can derive a general Eq. for the marginal utility:

$$U_i'(x_i^*) = \mu_0^* - \mu_i^*. \tag{22}$$

Thus, the Lagrange multipliers completely define the marginal utility for a given flow.

A.5 The log Utility Case: Proportional Fairness

Let us now investigate the nature of the optimal solution for the case where $U_i(x) = log(x_i)$.

$$U_i(x_i) = log(x_i), \text{ so,} \quad U_i'(x_i) = \frac{1}{x_i}. \tag{23}$$

Thus, by Case 1, for cases where $x_i > 0$ and $x_j > 0$,

$$\frac{1}{x_i^*} = \frac{1}{x_j^*}, \text{ so,} \quad x_i^* = x_j^*. \tag{24}$$

Thus, if we satisfy the capacity constraint with equality, then, for $x_i > 0$:

$$x_i^* = \frac{C}{k}, \tag{25}$$

where k is the number of flows with non-zero allocation

Now that we have defined the optimal solution for the log-utility case, what happens if we deviate slightly from the optimal solution? Recall that for convex function $f(x)$:

$$\nabla_x^\top f(x^*)(x - x^*) \geq 0. \tag{26}$$

Thus, applying this inequality to log-utility. Let

$$U(x) = \sum_{i=1}^{n} log(x_i) = -f(x) \tag{27}$$

since $f(x)$ is convex:

$$\begin{aligned} & -\nabla_x^\top U(x)(x - x^*) \geq 0, \\ \rightarrow \quad & \sum_{i=1}^{n} \frac{x_i - x_i^*}{x_i^*} \leq 0 \quad \text{(Proportional Fairness).} \end{aligned} \tag{28}$$

The interpretation of this result is: If the optimal solution satisfies the capacity constraint with equality, then if any flow adds additional traffic, the proportion that its traffic increases will be less than the sum of the proportions of traffic decreases along other flows. This property of the optimal solution of log-utility maximization is called "proportional fairness."

B FAIRNESS DERIVATIONS:

B.1 Derivation of the Fairness Condition for Univariate Objective:

1st KKT Condition: Primal constraints are feasible at the optimal point as in:

$$\sum_{i=1}^{N_r} \rho_i^* \leq P, \tag{29}$$

$$\rho_i^* \geq 0, \quad \forall i \in \{1, ..., N_r\}. \tag{30}$$

2nd KKT Condition: Dual constraints are feasible at the optimal point as in:

$$\mu_i^* \geq 0, \quad \forall i \in \{0, ..., N_r\}. \tag{31}$$

3rd KKT Condition: Complementary slackness holds at the optimal point as shown by:

$$\mu_0^*(\sum_{i}^{N_r} \rho_i^* - P) = 0, \tag{32}$$

$$\mu_i^* \rho_i^* = 0, \quad \forall i \in \{1, ..., N_r\}. \tag{33}$$

4th KKT Condition: The Lagrangian gradient is equal to zero at the optimal point.

The Lagrangian of the problem is formulated as:

$$L(\rho, \boldsymbol{\mu}; \rho^{\mathbf{0}}, \boldsymbol{\gamma}, P) = -\sum_{i}^{N_r} (\rho_i^0 + \rho_i)^{\gamma_i} + \mu_0(\sum_{i}^{N_r} \rho_i - P) + \sum_{i}^{N_r} (-\mu_i \rho_i). \tag{34}$$

The gradient of the Lagrangian is equal to zero at the optimal point which implies:

$$\nabla_\rho L(\rho^*, \boldsymbol{\mu}^*; \rho^{\mathbf{0}}, \boldsymbol{\gamma}, P) = 0 \Rightarrow \mu_0^* - \mu_i^* = \frac{\gamma_i}{(\rho_i^0 + \rho_i^*)^{1-\gamma_i}}, \forall i \in \{1, ..., N_r\}. \tag{35}$$

Case 1: If $\rho_i^* > 0$, then $\mu_i^* = 0$ by the complementary slackness condition. Then, the Eq. 35 reduces to:

$$U_i'(\rho_i^*; \rho_i^0, \gamma_i) = \frac{\gamma_i}{(\rho_i^0 + \rho_i^*)^{1-\gamma_i}} = \mu_0^*, \quad \forall i \in \{1, ..., N_r\}. \tag{36}$$

Thus, for two robots r_i and r_j where $\rho_i^* > 0$ and $\rho_j^* > 0$ satisfies:

$$U_i'(\rho_i^*; \rho_i^0, \gamma_i) = U_j'(\rho_j^*; \rho_j^0, \gamma_j) = \mu_0^*. \tag{37}$$

The interpretation of this result is that the *marginal utility* is the same at the optimal allocation for two robots. In other words, all accuracies would increase by the same amount with additional cloud resources.

Case 2: If $(\sum_{i=1}^{N_r} \rho_i^*) - P < 0$, then, $\mu_0^* = 0$ by complementary slackness. By Case 1, if $\rho_i^* > 0$, then, $\mu_i^* = 0$ and Eq. 35 becomes:

$$U_i'(\rho_i^*; \rho_i^0, \gamma_i) = 0. \tag{38}$$

The interpretation of this result is that if we do not use the total resources in the cloud, then each robot with a positive early exit allocation should have zero marginal utility at the optimal point. Otherwise, robots could use the excess resources in the cloud to increase their utility.

General Non-Trivial Case: If $\mu_0^* > 0$ and $\sum_{i=1}^{N_r} \rho_i^* = P$, we can derive a general Eq. for the marginal utility as:

$$U_i'(\rho_i^*; \rho_i^0, \gamma_i) = \mu_0^* - \mu_i^*. \tag{39}$$

Thus, the Lagrange multipliers completely define the marginal utility for a given resource allocation.

B.2 Derivation of The Fairness Condition of Multi-dimensional Objective Problem:

We employ a similar methodology for deriving fairness in multi-robot early exit inference resource allocation, akin to the approach described in §B.1. However, in this scenario, we have two decision variables that must be taken into account when examining the KKT conditions.

1st KKT Condition: Primal constraints are feasible at the optimal point as in:

$$\sum_{i}^{N_r} \rho_i^* \leq P, \tag{40}$$

$$\sum_{i}^{N_r} \alpha_i^* \leq T, \tag{41}$$

$$\alpha_i^* \geq 0, \quad \forall i = 1, ..., N, \tag{42}$$

$$\rho_i^* \geq 0, \quad \forall i = 1, ..., N. \tag{43}$$

2nd KKT Condition: Dual constraints are feasible at the optimal point as in:

$$0 \leq \mu_i^*; \quad \forall i = 0, 1, ..., N, \tag{44}$$

$$0 \leq \lambda_i^*; \quad \forall i = 0, 1, ..., N. \tag{45}$$

3rd KKT Condition: Complementary slackness holds at the optimal point as shown by:

$$\mu_0^*(\sum_{i}^{N_r} \rho_i^* - P) = 0, \tag{46}$$

$$\lambda_0^*(\sum_{i}^{N_r} \alpha_i^* - T) = 0, \tag{47}$$

$$\mu_i^* \rho_i^* = 0; \quad \forall i = 1, ..., N, \tag{48}$$

$$\lambda_i^* \alpha_i^* = 0; \quad \forall i = 1, ..., N. \tag{49}$$

4th KKT Condition: The Lagrangian gradient is equal to zero at the optimal point.

The Lagrangian of the problem, including vector variables denoted in bold, is formulated as:

$$L(\boldsymbol{\rho}, \boldsymbol{\alpha}, \boldsymbol{\lambda}, \boldsymbol{\mu}; \boldsymbol{\rho}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}, T, P) = -\sum_{i=1}^{N_r} U_i(\rho_i, \alpha_i; \rho_i^0, \alpha_i^0) + \mu_0(\sum_{i=1}^{N_r} \rho_i - P) + \sum_{i=1}^{N_r} \mu_i * (-\rho_i) + \lambda_0(\sum_{i=1}^{N_r} \alpha_i - T) + \sum_{i}^{N_r} \lambda_i * (-\alpha_i).$$

The gradient of the Lagrangian is equal to zero at the optimal point, which implies:

$$\nabla_{\boldsymbol{\rho},\boldsymbol{\alpha}} L(\boldsymbol{\rho}^*, \boldsymbol{\alpha}^*, \boldsymbol{\lambda}^*, \boldsymbol{\mu}^*; \boldsymbol{\rho}^{\mathbf{0}}, \boldsymbol{\alpha}^{\mathbf{0}}, T, P) = 0 = \nabla_{\boldsymbol{\rho},\boldsymbol{\alpha}}(-\sum_{i=1}^{N_r} U_i(\rho_i^*, \alpha_i^*; \rho_i^0, \alpha_i^0) + \mu_0^*(\sum_{i=1}^{N_r} \rho_i^* - P) + \sum_{i=1}^{N_r} \mu_i^*(-\rho_i^*) + \lambda_0^*(\sum_{i=1}^{N_r} \alpha_i^* - T) + \sum_{i}^{N_r} \lambda_i^*(-\alpha_i^*)). \tag{50}$$

Taking the gradient with respect to each decision variable, we obtain the vectorial form depicted as:

$$-\begin{bmatrix} U_1'(\rho_1^*) \\ U_2'(\rho_2^*) \\ \dots \\ U_n'(\rho_n^*) \\ U_1'(\alpha_1^*) \\ U_2'(\alpha_2^*) \\ \dots \\ U_n'(\alpha_n^*) \end{bmatrix} + \begin{bmatrix} \mu_0^* \\ \mu_0^* \\ \dots \\ \mu_0^* \\ \lambda_0^* \\ \lambda_0^* \\ \dots \\ \lambda_0^* \end{bmatrix} - \begin{bmatrix} \mu_1^* \\ \mu_2^* \\ \dots \\ \mu_n^* \\ \lambda_1^* \\ \lambda_2^* \\ \dots \\ \lambda_n^* \end{bmatrix} = 0, \tag{51}$$

which implies:

$$-U_i'(\rho_i^*) + \mu_0^* - \mu_i = 0, \quad \forall i \in \{1, ..., N_r\}, \tag{52}$$

$$-U_i'(\alpha_i^*) + \lambda_0^* - \lambda_i = 0, \quad \forall i \in \{1, ..., N_r\}. \tag{53}$$

Case 1: If $\alpha_i^* > 0$ and $\rho_i^* > 0$, it implies that $\mu_i^* = 0$ and $\lambda_i^* = 0$ due to the complementary slackness conditions outlined in Eq.s 46 and 47. By removing these zero variables in Eq.s 51-53, we obtain $U_i'(\alpha_i^*) = \lambda_0^*$ and $U_i'(\rho_i^*) = \mu_0^*$ for robot i. Note that λ_0^* and μ_0^* are independent of i and are the same for all robots. Therefore, for two robots i and j satisfying $\alpha_i^* > 0$, $\rho_i^* > 0$, $\alpha_j^* > 0$, and $\rho_j^* > 0$, the fairness condition holds as:

$$U_i'(\alpha_i^*) = U_j'(\alpha_j^*) = \lambda_0^*, \tag{54}$$

$$U_i'(\rho_i^*)\mu_0^* = U_j'(\rho_j^*) = \mu_0^*. \tag{55}$$

Therefore, it is also true that:

$$\frac{U_i'(\alpha_i^*)}{\lambda_0^*} = \frac{U_j'(\alpha_j^*)}{\lambda_0^*} = \frac{U_i'(\rho_i^*)}{\mu_0^*} = \frac{U_j'(\rho_j^*)}{\mu_0^*}. \tag{56}$$

This implies that when resources are fully utilized and there is a slight deviation from the optimal allocation, it leads to unfairness in resource distribution. If any agent is allocated more resources than the optimal point, the others experience a greater loss compared to the benefit gained by the agent with excess resources. As a result, overall satisfaction and aggregate accuracy of the agents decrease, even if some individual model accuracies show slight improvements.

Case 2: If $\sum_{i=1}^{N_r} \rho_i \leq P$ and $\sum_{i=1}^{N_r} \alpha_i \leq T$, by the complementary slackness, it implies that $\lambda_0^* = 0$ and $\mu_0^* = 0$. Then the Eq.s 51-53 reduce to:

$$\lambda_0^* = 0 \Rightarrow U_i'(\alpha_i^*) = 0, \tag{57}$$

$$\mu_0^* = 0 \Rightarrow U_i'(\rho_i^*) = 0. \tag{58}$$

The explanation for this outcome suggests that when we don't fully utilize the cloud resources, robots assigned a positive resource allocation should ideally have no additional benefit at the optimal point. Otherwise, robots might exploit the surplus cloud resources to enhance their utility unfairly. In other words, this case illustrates the inflation of resource allocation to the point where none of the robots have a positive shadow price for the resource. Note that, since there are two independent decision variables and corresponding upper limits, either Eq. 54 or 57 can hold together with either Eq. 55 or 58, meaning one of the resources might inflate while the other is fully consumed.

General Non-Trivial Case: If $\lambda_0 \geq 0$ and $\mu_0 \geq 0$, then $\sum_{i=1}^{N_r} \rho_i = P$ and $\sum_{i=1}^{N_r} \alpha_i = T$. Based on this and Eq.s 51-53, we can derive the general Eq. for the marginal utility as:

$$U_i'(\rho_i^*) = \mu_0^* - \mu_i^*, \tag{59}$$

$$U_i'(\alpha_i^*) = \lambda_0^* - \lambda_i^*. \tag{60}$$

By gathering all results we obtained and taking the derivative, we obtained the fairness condition formulated as:

$$(\gamma_i^\alpha)\frac{(\rho_i^0 + \rho_i)^{\gamma_i^\rho}}{(\alpha_i^0 + \alpha_i)^{1-\gamma_i^\alpha}} = (\gamma_j^\alpha)\frac{(\rho_j^0 + \rho_j)^{\gamma_j^\rho}}{(\alpha_j^0 + \alpha_j)^{1-\gamma_j^\alpha}}. \tag{61}$$

C BICONVEX PROBLEMS:

The term "biconvex" refers to functions that exhibit convexity in each set of variables when the other set remains fixed. This unique property makes biconvex optimization invaluable for modeling complex real-world challenges, including the intricacies of resource allocation in ML applications.

A biconvex optimization problem is formally defined as follows.

$$\begin{aligned} \text{minimize} \quad & f(x, y) \\ \text{subject to} \quad & g_1(x, y) \leq 0 \\ & g_2(x, y) \leq 0 \\ & h(x, y) = 0 \\ & x \in \mathcal{X}, \quad y \in \mathcal{Y} \end{aligned} \tag{62}$$

Here, x and y represent sets of variables partitioned into $\mathcal{X}$ and $\mathcal{Y}$ respectively.

The biconvex property is unique because it ensures that the objective function $f(x, y)$ remains convex in both sets of variables (x and y) individually, even when the other set is kept constant. This dual convexity property sets biconvex optimization apart from other optimization paradigms. Unlike non-biconvex problems, biconvex optimization guarantees the existence of a single, globally optimal solution. In practical terms, this means that our allocation mechanism, `FairSynergy`, can make precise, optimal decisions in complex, multi-variable scenarios.

This distinct characteristic makes biconvex optimization particularly powerful for applications where multiple independent variables significantly impact the optimization process, as is the case in resource allocation for ML tasks. Using these properties, our approach ensures that the resource allocation adapts to varying numbers of robots, parameters, training dataset sizes, and other pertinent variables. The formal expression of a biconvex optimization problem underlines the elegance and effectiveness of this approach in solving intricate resource allocation challenges, marking a significant leap in the field of ML resource management.

C.1 Alternate Convex Search (ACS) Solvers for Biconvex Optimization

Achieving iterative convergence in biconvex optimization is pivotal for practical applications. The Alternate Convex Search (ACS) method provides a rigorous solution to this challenge. ACS iteratively optimizes the variables x and y of a biconvex function by alternating between fixing one set and optimizing the other.

Algorithm 1: Alternate Convex Search Algorithm for Biconvex Optimization

Data: Initial values for $x^{(0)}$ and $y^{(0)}$
Result: Optimized values for x and y
while *not converged* **do**
  Fix $y^{(k-1)}$, optimize $x^{(k)}$: $x^{(k)} = \text{argmin}_x f(x, y^{(k-1)})$;
  Fix $x^{(k)}$, optimize $y^{(k)}$: $y^{(k)} = \text{argmin}_y f(x^{(k)}, y)$;
  if $|f(x^{(k)}, y^{(k)}) - f(x^{(k-1)}, y^{(k-1)})| < \epsilon$ *or maximum iterations reached* **then**
    break;
  end
end

The convergence of ACS is established through the dual minimization perspective, ensuring convergence to a critical point of the biconvex objective function under specific conditions. This critical point can be a local minimum, a saddle point, or a global minimum, depending on the problem's nature.

In the context of our research, ACS plays a vital role in ensuring the iterative convergence of the biconvex optimization problem embedded in our resource allocation approach, `FairSynergy`. Using ACS, our method systematically explores the solution space, refining resource allocation based on intricate interdependencies between variables. This iterative optimization process guarantees algorithm convergence while enhancing the precision and efficiency of resource allocation. This adaptability is essential for dynamic ML environments with diverse and evolving parameters.

D SLATER'S CONDITION

Slater's condition is a crucial constraint qualification in convex optimization theory. It ensures the existence of a feasible point strictly within the feasible region of a convex optimization problem. Consider a convex optimization problem in standard form:

$$\begin{aligned} &\text{Minimize} && f(x) \\ &\text{Subject to} && h_i(x) \leq 0, \quad i = 1, 2, \ldots, m, \\ &&& Ax = b. \end{aligned}$$

Slater's condition holds if there exists a point x inside the feasible region such that all the inequality constraints $h_i(x) < 0$ and the equality constraints $Ax = b$ are satisfied.

Slater's Condition: There exists an x such that $h_i(x) < 0$ for all $i = 1, 2, \ldots, m$ and $Ax = b$.

When Slater's condition is satisfied in a convex optimization problem, it guarantees strong duality. This implies that the optimal values of the primal and dual problems are equal, and there exist primal and dual optimal solutions achieving these optimal values.

In summary, Slater's condition ensures the existence of a strictly feasible point within the feasible region, which is essential for strong duality in convex optimization problems. It establishes a vital connection between primal feasibility, dual feasibility, and optimality conditions (KKT conditions), providing a solid theoretical foundation for understanding the behavior of convex optimization problems and their dual counterparts.

Slater's condition is closely related to the Karush-Kuhn-Tucker (KKT) conditions, which are necessary conditions for optimality in convex optimization. When Slater's condition is met, it ensures strong duality, meaning that the optimal value of the primal problem equals the optimal value of the dual problem. The KKT conditions for a convex optimization problem with Slater's condition are as follows:

- **Primal Feasibility:** $h_i(x) \leq 0$ for all $i = 1, 2, \ldots, m$
- **Dual Feasibility:** $\lambda_i \geq 0$ for all $i = 1, 2, \ldots, m$ (where λ_i are Lagrange multipliers corresponding to the inequality constraints)
- **Complementary Slackness:** $\lambda_i \cdot h_i(x) = 0$ for all $i = 1, 2, \ldots, m$
- **Gradient of the Lagrangian:** $\nabla f(x) + \sum_{i=1}^{m} \lambda_i \nabla h_i(x) + A^T \nu = 0$ (where ν represents Lagrange multipliers corresponding to the equality constraints $Ax = b$)