Mathematical Formulation of the Onyttig Pipeline

Kaustubh Sharma — EET 110 Term Paper Update

Objective: To formally define the evolutionary mechanisms of *Onyttig* that synthesize a large-scale, mathematically valid dataset of Linear Programming (LP) instances from a small seed set, mapping them to Natural Language (L) and OptBNF structural formats ($\mathcal{F}$) for supervised fine-tuning.

---

1. Canonical Formulation of the Seed Space

Let the seed space $\mathcal{S}_{seed}$ consist of canonical LP instances. A single instance $M \in \mathcal{S}_{seed}$ is defined by the tuple $M = (\mathbf{c}, \mathbf{A}, \mathbf{b}, \mathcal{B})$, representing the objective vector $\mathbf{c} \in \mathbb{R}^n$, constraint matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$, right-hand side $\mathbf{b} \in \mathbb{R}^m$, and variable bounds $\mathcal{B}$. The standard mathematical evaluation computes the optimal solution $\mathbf{x}^*$:

$$\min_{\mathbf{x} \in \mathbb{R}^n} \quad \mathbf{c}^T \mathbf{x} \quad \text{s.t.} \quad \mathbf{A}\mathbf{x} \leq \mathbf{b}, \quad \mathbf{x} \geq 0 \tag{1}$$

2. Evolutionary Synthesis Operators

To generate a massive synthetic dataset $\mathcal{S}_{large}$, we apply stochastic mutation operators $\mathcal{M}_{mut} : M \to M'$. A critical mathematical requirement is that M' must remain strictly feasible and bounded. Given a parent instance M with known optimal solution $\mathbf{x}^*$, we define three primary operators:

2.1. Valid Constraint Addition (Hyperplane Injection)

To increase problem complexity without inducing infeasibility, a new constraint row $\mathbf{a}_{new} \in \mathbb{R}^n$ and scalar b_{new} are appended. Feasibility is preserved iff the existing optimal solution $\mathbf{x}^*$ resides within the new closed half-space:

$$\mathbf{a}_{new}^T \mathbf{x}^* \leq b_{new} \quad \implies \quad \mathbf{A}' = \begin{bmatrix} \mathbf{A} \\ \mathbf{a}_{new}^T \end{bmatrix}, \quad \mathbf{b}' = \begin{bmatrix} \mathbf{b} \\ b_{new} \end{bmatrix} \tag{2}$$

2.2. RHS Perturbation (Capacity Scaling)

We perturb the capacity vector $\mathbf{b}$ by a noise vector $\boldsymbol{\epsilon} \in \mathbb{R}^m$. To ensure the original $\mathbf{x}^*$ remains in the feasible polytope $\mathcal{P}' = \{\mathbf{x} \mid \mathbf{A}\mathbf{x} \leq \mathbf{b} + \boldsymbol{\epsilon}\}$, the perturbation is lower-bounded by the slack $\mathbf{s}^*$:

$$\boldsymbol{\epsilon} \geq \mathbf{A}\mathbf{x}^* - \mathbf{b} \quad \implies \quad \boldsymbol{\epsilon} \geq -\mathbf{s}^* \tag{3}$$

where $\mathbf{s}^* \geq 0$.

2.3. Objective Rotation

To alter the optimal solution mapping, the objective vector is rotated by a random orthogonal matrix $\mathbf{R}$ or scaled via scalar $\alpha \in \mathbb{R}^+$: $\mathbf{c}' = \alpha\mathbf{c} + \boldsymbol{\eta}$, ensuring the problem remains bounded.

3. API Annotation and Optimization Grammar Mapping

Once $\mathcal{S}_{large} = \{M_1, M_2, \ldots, M_N\}$ is generated, instances are serialized into the OptBNF JSON grammar, denoted as $\mathcal{F}(M_i)$. An LLM API generates the natural language description $L_i = g_{api}(\mathcal{F}(M_i))$. This yields the final aligned dataset: $\mathcal{D} = \{(L_i, \mathcal{F}(M_i))\}_{i=1}^N$.

4. Supervised Fine-Tuning Objective

The base model (DeepSeek-Math-7B) with parameters θ models the conditional probability $P_\theta(\mathcal{F} \mid L)$. The model is fine-tuned to minimize the standard auto-regressive Cross-Entropy loss over the structured OptBNF tokens sequence $y_{1:T}$:

$$\mathcal{L}_{SFT}(\theta) = -\mathbb{E}_{(L,\mathcal{F})\sim\mathcal{D}} \left[\sum_{t=1}^{T} \log P_\theta \left(y_t \mid y_{<t}, L \right) \right] \tag{4}$$

Evaluation metrics are defined as the discrete indicator functions: 1. $\mathbb{I}_{syntax}(\hat{\mathcal{F}})$ (Valid OptBNF JSON parsability) 2. $\mathbb{I}_{solver}(\hat{\mathcal{F}})$ (Feasible and bounded under `linprog`).