Lion Secretly Solves a Constrained Optimization: As Lyapunov Predict

Lizhang Chen*, Bo Liu*, Kaizhao Liang*, Qiang Liu,

The University of Texas at Austin

1 Theoretical Exploration of Lion Optimizer

We unveil the theoretical foundation of the Lion optimizer, which, despite its empirical success, lacked a clear theoretical basis. Our work establishes that Lion is not just an empirically effective optimizer but is theoretically grounded in solving a specific bound-constrained optimization problem.

given $\beta_1, \beta_2, \lambda, \eta, f$
initialize $\theta_0, m_0 \leftarrow 0$
while θ_t not converged **do**
  $g_t \leftarrow \nabla_\theta f(\theta_{t-1})$
  update model parameters
  $c_t \leftarrow \beta_1 m_{t-1} + (1-\beta_1) g_t$
  $\theta_t \leftarrow \theta_{t-1} - \eta_t(\text{sign}(c_t) + \lambda\theta_{t-1})$
  update EMA of g_t
  $m_t \leftarrow \beta_2 m_{t-1} + (1-\beta_2) g_t$
end while
return θ_t

2 Lyapunov: Weight decay is a constraint

Our study introduces a new Lyapunov function for Lion updates. This development is instrumental in demonstrating that Lion minimizes a general loss function *f(x)* while enforcing a bound constraint $||x||_\infty \le 1/\lambda$ with λ as the weight decay coefficient.

- Main result: Lion-$\mathcal{K}$ solves

$$\min_x F(x) := \alpha f(x) + \frac{\gamma}{\lambda}\mathcal{K}^*(\lambda x),$$

$\mathcal{K}^*$ is the convex conjugate of $\mathcal{K}$: $\mathcal{K}^*(x) = \sup_y \left(x^\top y - \mathcal{K}(y)\right)$.

- When $\mathcal{K}^*(x)$ can take infinite values, it is a constrained optimization:

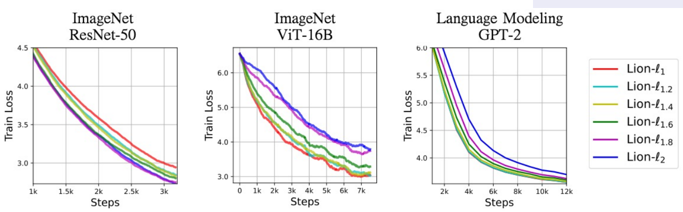
$$\min_x F(x), \qquad s.t. \quad x \in \text{dom}\mathcal{K}^*,$$

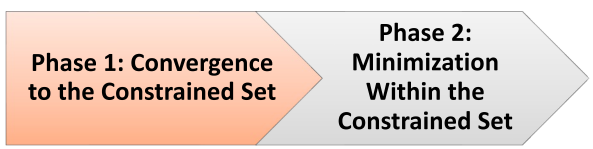
where $\text{dom}\mathcal{K}^* = \{x : \mathcal{K}^*(x) < +\infty\}$.

Example: $\mathcal{K}(x) = \|x\|_1$, then $\mathcal{K}^*(x) = \begin{cases} 0 & \text{if } \|x\|_\infty \le 1 \\ +\infty & \text{if } \|x\|_\infty > 1. \end{cases}$

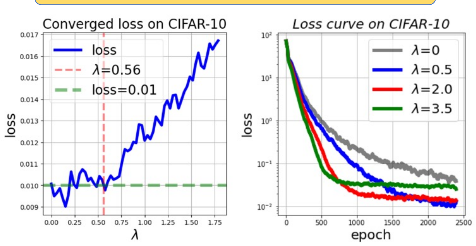
3 Extension to Lion-$\mathcal{K}$ Algorithms

We extend the Lion optimizer to a broader family of Lion- $\mathcal{K}$ algorithms.





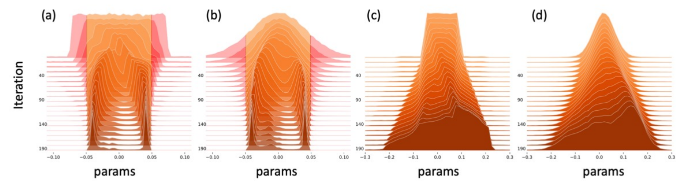
Weight Decay λ MATTERS!



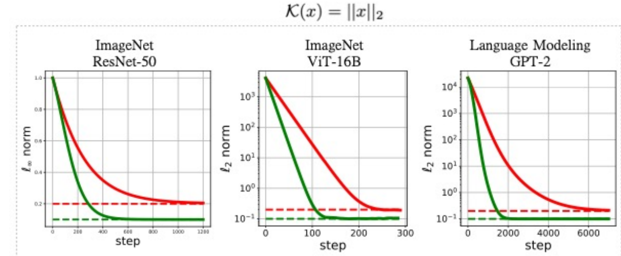
Line ID	$\mathcal{K}(x)$	$\nabla\mathcal{K}(x)$	$\min_x f(x) + \mathcal{K}^*(x)$
①	$\|x\|_1$	$\text{sign}(x)$	$\min f(x) \ \ s.t. \ \ \|x\|_\infty \le 1$
②	$\|x\|_p$	$\frac{\text{sign}(x)\|x\|^{p-1}}{\|x\|_p^{p-1}}$	$\min f(x) \ \ s.t. \ \ \|x\|_q \le 1$
③	$\sum_i \max(\|x_i\| - e, 0)$	$\text{sign}(x)\mathbb{I}(\|x\| > e)$	$\min f(x) + e\|x\|_1 \ \ s.t. \ \ \|x\|_\infty \le 1$
④	$\sum_{i \le i^{cut}} \|x_{(i)}\|$	$\text{sign}(x)\mathbb{I}(\|x\| > \|x_{(i^{cut})}\|)$	$\min f(x) \ \ s.t. \ \ \|x\|_1 \le i^{cut}, \ \ \|x\|_\infty \le 1$
⑤	$\sum_i \text{huber}_e(x_i)$	$\text{clip}(x, -e, e)/e$	$\min f(x) + \frac{e}{2}\|x\|_2^2 \ \ s.t. \ \ \|x\|_\infty < 1$

4 Empirical Validation

Through extensive experiments, including on a toy example and large-scale tasks like ImageNet classification and language modeling, we empirically validate our theoretical findings. We observe that different choices of ϕ in Lion-ϕ influence the optimization behavior and performance, confirming our theoretical predictions.



$\mathcal{K}(x) = \|x\|_2$



$\mathcal{K}(x) = \|x\|_1$