Predictive Hints in Optimistic Online Learning for Better Optimizers

Luca Mezger
Under the direction of Prof. Ashok Cutkosky and Qinzi Zhang
Department of Electrical & Computer Engineering, Boston University

Research Science Institute

August 1, 2024

Online Learning as a Game

- **Setup:**
  - At each round t, an adversary chooses $y_t \in \mathcal{Y}$

Online Learning as a Game

- **Setup:**
  - At each round t, an adversary chooses $y_t \in \mathcal{Y}$
  - Learner predicts $\hat{y}_t \in \mathcal{Y}$

Online Learning as a Game

- **Setup:**
  - At each round t, an adversary chooses $y_t \in \mathcal{Y}$
  - Learner predicts $\hat{y}_t \in \mathcal{Y}$
  - Adversary reveals y_t and learner incurs loss $\ell(\hat{y}_t, y_t)$

Online Learning as a Game

- **Setup:**
  - At each round t, an adversary chooses $y_t \in \mathcal{Y}$
  - Learner predicts $\hat{y}_t \in \mathcal{Y}$
  - Adversary reveals y_t and learner incurs loss $\ell(\hat{y}_t, y_t)$
- **Example: Number Guessing Game**
  - *Rounds:* $t = 1, 2, \ldots, T$

Online Learning as a Game

- **Setup:**
  - At each round t, an adversary chooses $y_t \in \mathcal{Y}$
  - Learner predicts $\hat{y}_t \in \mathcal{Y}$
  - Adversary reveals y_t and learner incurs loss $\ell(\hat{y}_t, y_t)$
- **Example: Number Guessing Game**
  - *Rounds:* $t = 1, 2, \ldots, T$
    - *Adversary's choice:* $y_t \in [0, 1]$

Online Learning as a Game

- **Setup:**
  - At each round t, an adversary chooses $y_t \in \mathcal{Y}$
  - Learner predicts $\hat{y}_t \in \mathcal{Y}$
  - Adversary reveals y_t and learner incurs loss $\ell(\hat{y}_t, y_t)$
- **Example: Number Guessing Game**
  - *Rounds:* $t = 1, 2, \ldots, T$
    - *Adversary's choice:* $y_t \in [0, 1]$
    - *Learner's prediction:* $\hat{y}_t \in [0, 1]$

Online Learning as a Game

- **Setup:**
  - At each round t, an adversary chooses $y_t \in \mathcal{Y}$
  - Learner predicts $\hat{y}_t \in \mathcal{Y}$
  - Adversary reveals y_t and learner incurs loss $\ell(\hat{y}_t, y_t)$
- **Example: Number Guessing Game**
  - *Rounds:* $t = 1, 2, \ldots, T$
    - *Adversary's choice:* $y_t \in [0, 1]$
    - *Learner's prediction:* $\hat{y}_t \in [0, 1]$
    - *Loss:* Squared error $\ell(\hat{y}_t, y_t) = (\hat{y}_t - y_t)^2$

Follow-the-Leader (FTL)

$$\hat{y}_t = \arg\min_{\hat{y}} \sum_{i=1}^{t-1} \ell_i(\hat{y})$$

Regret

$$R_T = \sum_{t=1}^{T} \ell_t(\hat{y}) - \min_x \sum_{t=1}^{T} \ell_t(x)$$

Regret

$$R_T = \sum_{t=1}^{T} \ell_t(\hat{y}) - \min_x \sum_{t=1}^{T} \ell_t(x)$$

- Goal: minimize regret

Online to Non-Convex Conversion (O2NC), Cutkosky et al. (2023)

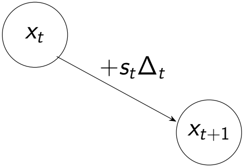
Key idea ($s_t \sim \exp(1)$):

$$\mathbb{E}\left[F(x_{t-1} + s_t \Delta_t) - F(x_{t-1})\right] = \mathbb{E}\left[\langle \nabla F(x_{t-1} + s_t \Delta_t), \Delta_t \rangle\right]$$

Online to Non-Convex Conversion (O2NC), Cutkosky et al. (2023)

Key idea ($s_t \sim \exp(1)$):

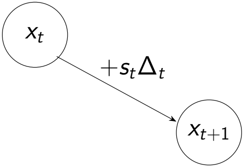
$$\mathbb{E}\left[F(x_{t-1} + s_t\Delta_t) - F(x_{t-1})\right] = \mathbb{E}\left[\langle \nabla F(x_{t-1} + s_t\Delta_t), \Delta_t\rangle\right]$$



Online to Non-Convex Conversion (O2NC), Cutkosky et al. (2023)

Key idea ($s_t \sim \exp(1)$):

$$\mathbb{E}\left[F(x_{t-1} + s_t\Delta_t) - F(x_{t-1})\right] = \mathbb{E}\left[\langle \nabla F(x_{t-1} + s_t\Delta_t), \Delta_t \rangle\right]$$



- Simplifies to:

$$\mathbb{E}\left[F(x_t) - F(x_{t-1})\right] = \mathbb{E}\left[\langle \nabla F(x_t), \Delta_t \rangle\right]$$

Minimizing the Gap

$$\mathbb{E}\left[F(x_t) - F(x_{t-1})\right] = \mathbb{E}\left[\langle \nabla F(x_t), \Delta_t \rangle\right]$$

- **Objective:** Minimize $\langle \nabla F(x_t), \Delta_t \rangle$

Minimizing the Gap

$$\mathbb{E}\left[F(x_t) - F(x_{t-1})\right] = \mathbb{E}\left[\langle \nabla F(x_t), \Delta_t \rangle\right]$$

- **Objective:** Minimize $\langle \nabla F(x_t), \Delta_t \rangle$
- **Desired step:** $\Delta_t \approx -\nabla F(x_t)$

Minimizing the Gap

$$\mathbb{E}\left[F(x_t) - F(x_{t-1})\right] = \mathbb{E}\left[\langle \nabla F(x_t), \Delta_t \rangle\right]$$

- **Objective:** Minimize $\langle \nabla F(x_t), \Delta_t \rangle$
- **Desired step:** $\Delta_t \approx -\nabla F(x_t)$
- **Caveat:** $\nabla F(x_t)$ is unknown.

Minimizing the Gap

$$\mathbb{E}\left[F(x_t) - F(x_{t-1})\right] = \mathbb{E}\left[\langle \nabla F(x_t), \Delta_t \rangle\right]$$

- **Objective:** Minimize $\langle \nabla F(x_t), \Delta_t \rangle$
- **Desired step:** $\Delta_t \approx -\nabla F(x_t)$
- **Caveat:** $\nabla F(x_t)$ is unknown.
- $\ell_t(\Delta) = \langle g_t, \Delta \rangle$:

$$\mathrm{Regret}_T(u) := \sum_{t=1}^{T} \langle g_t, \Delta_t - u \rangle$$

Optimistic Online Gradient Descent

- Update Rule (O2NC: $x = \Delta$):

$$x_{t+1} = x_t - \eta(\nabla F(x_t) + h_t - h_{t-1})$$

Optimistic Online Gradient Descent

- Update Rule (O2NC: $x = \Delta$):

$$x_{t+1} = x_t - \eta(\nabla F(x_t) + h_t - h_{t-1})$$

- Unrolling the recursion:

$$x_{t+1} = x_0 - \eta \sum_{i=0}^{t} (\nabla F(x_i) + h_i - h_{i-1})$$

Optimistic Online Gradient Descent

- Update Rule (O2NC: $x = \Delta$):

$$x_{t+1} = x_t - \eta(\nabla F(x_t) + h_t - h_{t-1})$$

- Unrolling the recursion:

$$x_{t+1} = x_0 - \eta \sum_{i=0}^{t} (\nabla F(x_i) + h_i - h_{i-1})$$

- Goal: $h_t \approx g_{t+1}$

$$x_{t+1} \approx x_0 - \eta \sum_{i=0}^{t} (\nabla F(x_i)) - \eta h_t$$

Regret Bound for Optimistic Online Gradient Descent

$$R_T \leq \frac{1}{2\eta}\|\theta_1 - u\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t - h_{t-1}\|^2$$

Regret Bound for Optimistic Online Gradient Descent

$$R_T \leq \frac{1}{2\eta}\|\theta_1 - u\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|g_t - h_{t-1}\|^2$$

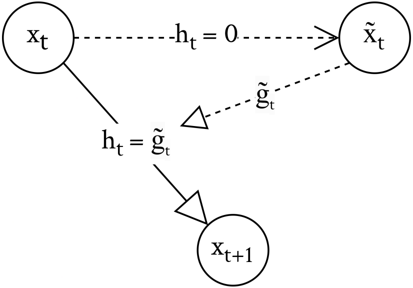
- $h_t \approx g_{t+1}$ minimizes regret bound R_T

Theorem 7 from Cutkosky et al. (2023)

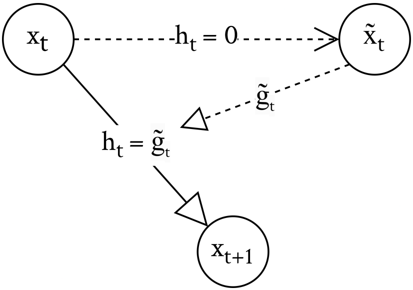
$$\mathbb{E}[F(\theta_M)] = F(\theta_0) + \mathbb{E}\left[\sum_{n=1}^{M}\langle g_n, \Delta_n - u_n\rangle\right] + \mathbb{E}\left[\sum_{n=1}^{M}\langle g_n, u_n\rangle\right]$$

- Loss bound for O2NC

Proof of Concept

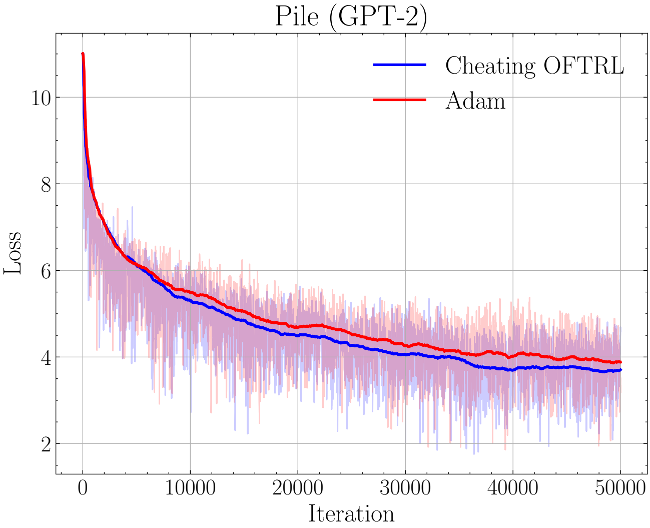


Proof of Concept

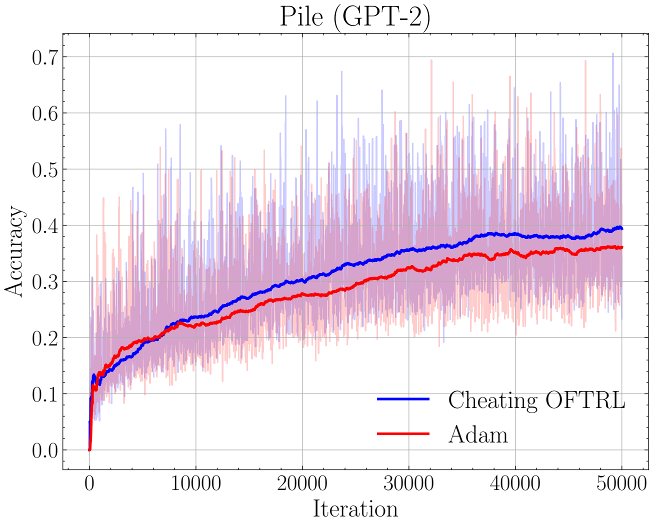


- Evaluation:
  - GPT-2
  - Pile Dataset
  - Train loss

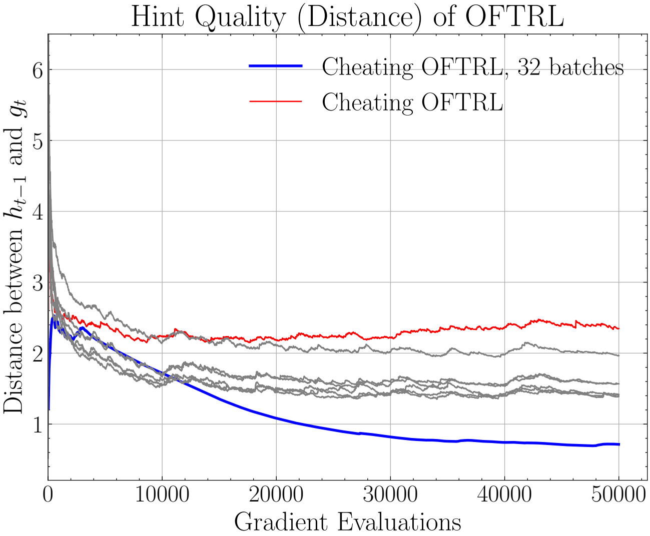
Proof of Concept



Proof of Concept



Results



Hint Quality (Distance) of OFTRL
Cheating OFTRL, 32 batches
Cheating OFTRL
Distance between h_{t-1} and g_t
1
2
3
4
5
6
0
10000
20000
30000
40000
50000
Gradient Evaluations

Results

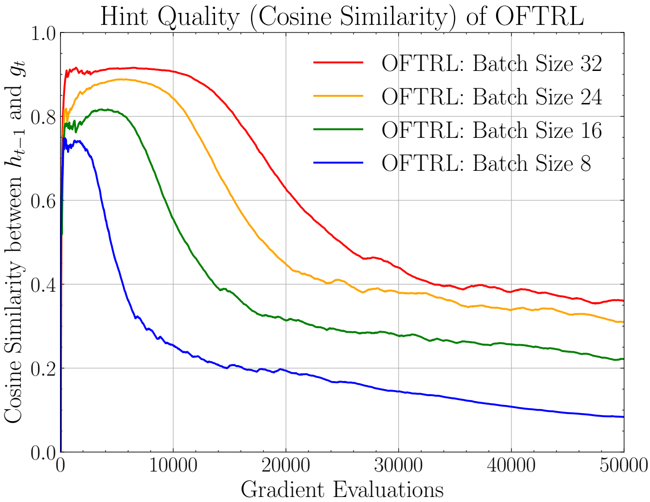


Figure: Cosine Similarity $\frac{h_{t-1} \cdot g_t}{\|h_{t-1}\| \|g_t\|}$ of OFTRL, with $h_{t+1} = \beta \Delta_t + (1 - \beta) g_t$

Hint Calculations

Formula	Hyperparameter	EMA Loss
$h_{t+1} = 0$ (Adam)	$\eta = 0.0003$	3.89
$h_{t+1} = g_t$	$\eta = 0.0003$	3.90
$h_{t+1} = \beta h_t + (1-\beta) g_t$	$\eta = 0.0003,\ \beta = 0.9$	3.96
$h_{t+1} = h_t + (1-\beta)(g_t - h_t)$	$\eta = 0.0003,\ \beta = 0.8$	3.94
$h_{t+1} = g_t + \beta(h_t - g_t)$	$\eta = 0.0003,\ \beta = 0.8$	3.98
$h_{t+1} = \beta h_t + \beta g_t$	$\eta = 0.0003,\ \beta = 0.5$	3.93
$h_{t+1} = \frac{t}{t+1} h_t + \frac{1}{t+1} g_t$	$\eta = 0.0003$	4.08
$h_{t+1} = \frac{\sqrt{h_t^2 + g_t^2}}{\sqrt{2}}$	$\eta = 0.0001,\ \beta = 0.9$	4.72
$h_{t+1} = \beta \Delta_t + (1-\beta) g_t$	$\eta = 0.0003,\ \beta = 0.8$	3.88

Table: time-weighted EMA of the train loss at the 50,000th iteration (same computational budget) for each hint.

Conclusion

- Potential (shown by POC)

Conclusion

- Potential (shown by POC)
- Future Work:
  - bigger batch sizes

Conclusion

- Potential (shown by POC)
- Future Work:
  - bigger batch sizes
  - more effective hint generations

Acknowledgments

- Prof. Ashok Cutkosky
- Qinzi Zhang
- Shuvom Sadhuka
- Victor Kolev, River Grace, Jenny Sendova, and Canaan He
- Rickoids
- Research Science Institute (RSI)
- MIT & Boston University
- FBK Bern, René & Susanne Braginsky Stiftung, and Fritz-Gerber-Stiftung