Predictive Hints in Optimistic Online Learning for Better Optimizers

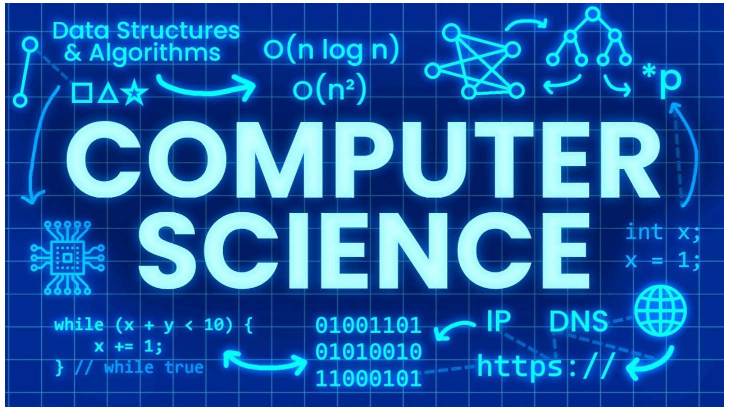
Luca Mezger

28.11.2024

Fachreferat, HSLU

Über mich

gym fms | THUN
Eine Institution des Kantons Bern



RSI
Research Science Institute

RSI am MIT

Ziel

- Schnellere/Genauere Optimierungsalgorithmen

→Bedeutung für maschinelles Lernen

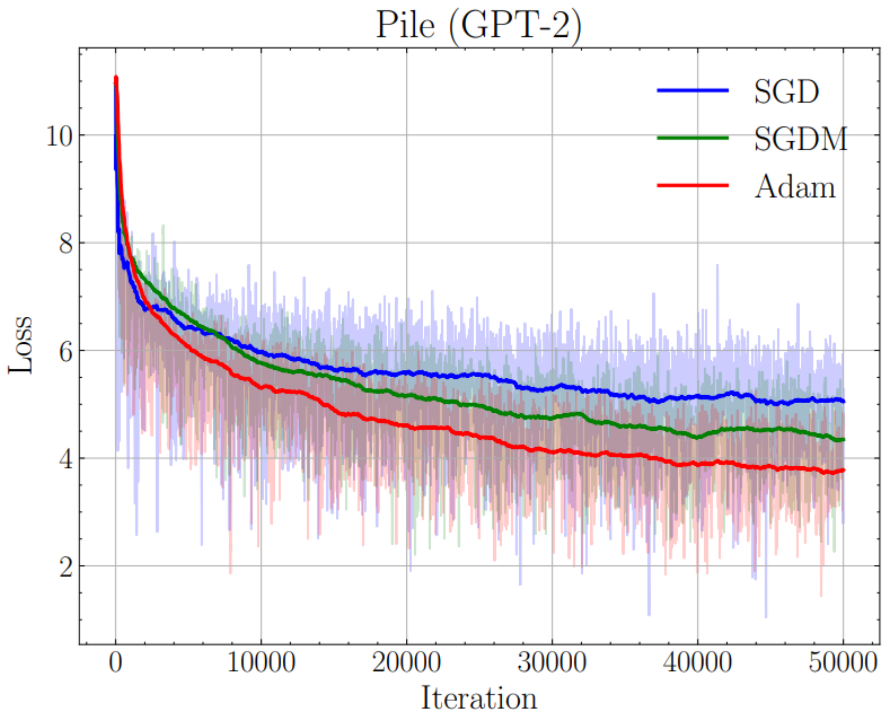
- Besseres Theoretische Verständnis von Optimierungsalgorithmen

SGD vs. Adam

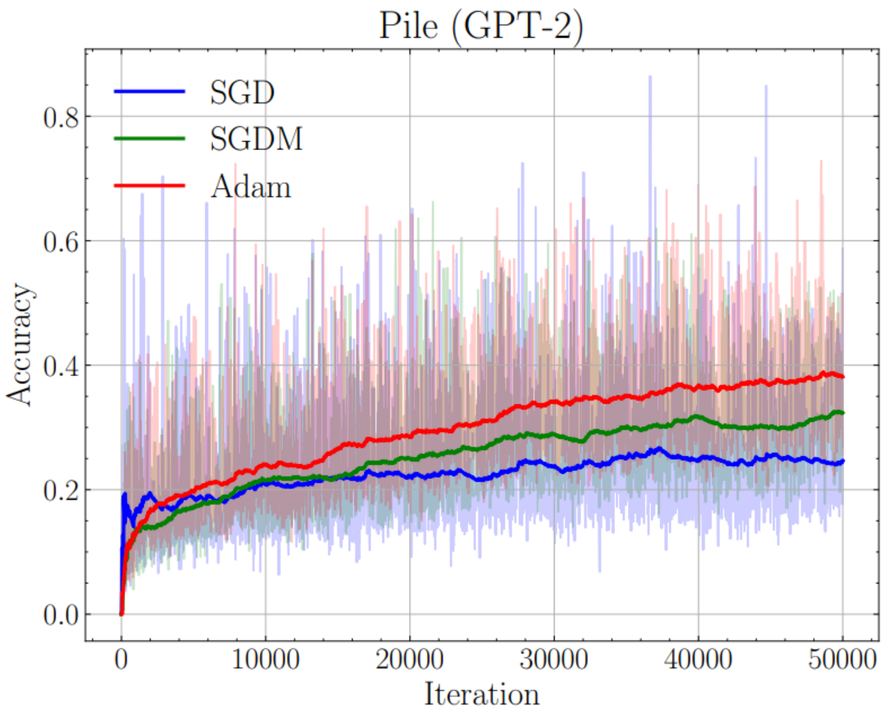
$$x_{t+1} = x_t - \eta_t \nabla \ell(x_t) \quad \textbf{vs.} \quad x_{t+1} = x_t - \eta \frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}$$

? ?

- Adam > SGD (in schlecht konditionierten Problemen)



?



?

Online Learning

- Aufbau:
  - In jeder Runde t wählt ein Gegner $\boldsymbol{y_t}$
  - Der "Lerner" wählt $\widehat{\boldsymbol{y_t}}$
  - Der Gegner zeigt $\boldsymbol{y_t}$ und der Lerner bezahlt eine Verlustsfunktion $\ell(\widehat{\boldsymbol{y_t}}, \boldsymbol{y_t})$

- Beispiel: Zahlen erraten
  - Runden: $\boldsymbol{t = 1, 2, \dots, T}$
  - Gegner: $\boldsymbol{y_t} \in \{\boldsymbol{1, 2, \dots, 10}\}$
  - Lerner: $\widehat{\boldsymbol{y_t}} \in \{\boldsymbol{1, 2, \dots, 10}\}$
  - Verlust: $\ell(\widehat{\boldsymbol{y_t}}, \boldsymbol{yt}) = (\widehat{\boldsymbol{y_t}} - \boldsymbol{y_t})^2$

Follow-the-Leader (FTL)

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

Follow-the-Regularized-Leader (FTRL)

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

$$R(\hat{y}) = \frac{\lambda}{2} \|\hat{y}\|^2$$

Follow-the-Regularized-Leader (FTRL)

$$\hat{y}_t = \arg\min_{\hat{y}} \left(\sum_{i=1}^{t-1} \ell_i(\hat{y}) + \frac{\lambda}{2} \|\hat{y}\|^2 \right)$$

arg min lösen

$$\hat{y}_{t+1} = -\eta \sum_{i=1}^{t} \nabla \ell_i$$

$$\hat{y}_{t+1} = \hat{y}_t - \eta \nabla \ell_t$$

Regret

$$\text{Regret}_T = \sum_{t=1}^{T}(y_t - \hat{y}_t)^2 - \sum_{t=1}^{T}(y_t - u)^2$$

$$u = \arg\min_{v \in \mathcal{Y}} \sum_{t=1}^{T}(y_t - v)^2$$

Regret Bound FTRL

C.2 Key Lemma

Suppose F is α-strongly convex, f is convex, and let $x = \arg\min_z F(z)$ and $x' = \arg\min_z G(z)$, where $G(x) = F(x) + f(x)$. Then:

$$0 \leq f(x) - f(x') \leq \frac{1}{\alpha}\|\nabla f(x)\|^2$$

C.3 Proof of the Key Lemma

Since F is α-strongly convex, by definition, we have:

$$F(x') \geq F(x) + \langle \nabla F(x), x' - x \rangle + \frac{\alpha}{2}\|x - x'\|^2$$

Regret Bound FTRL

By the optimality condition of x:

$$\langle \nabla F(x), x' - x \rangle \geq 0$$

Therefore, we can simplify the strong convexity condition to:

$$F(x') \geq F(x) + \frac{\alpha}{2}\|x - x'\|^2 \tag{1}$$

Similarly, for G:

$$G(x') \geq G(x) + \langle \nabla G(x), x' - x \rangle + \frac{\alpha}{2}\|x - x'\|^2$$

Regret Bound FTRL

By the optimality condition of x':

$$\langle \nabla G(x'), x - x' \rangle \geq 0$$

Thus, we obtain:

$$G(x') \leq G(x)$$

Since $G(x) = F(x) + f(x)$, we have:

$$G(x') = F(x') + f(x')$$

$$F(x') + f(x') \leq F(x) + f(x)$$

Regret Bound FTRL

From (1), we have:

$$F(x') \geq F(x) + \frac{\alpha}{2}\|x - x'\|^2$$

Substituting this into (2):

$$f(x) - f(x') \geq \frac{\alpha}{2}\|x - x'\|^2 \tag{3}$$

Since f is convex, using the first-order condition:

$$f(x) \leq f(x') + \langle \nabla f(x), x - x' \rangle$$

Taking the absolute value:

$$f(x) - f(x') \leq |\langle \nabla f(x), x - x' \rangle|$$

Regret Bound FTRL

Applying the Cauchy-Schwarz inequality:

$$f(x) - f(x') \leq \|\nabla f(x)\| \cdot \|x - x'\|$$

From (3), we have:

$$\|x - x'\|^2 \leq \frac{2}{\alpha}(f(x) - f(x'))$$

Thus:

$$\|x - x'\| \leq \sqrt{\frac{2}{\alpha}(f(x) - f(x'))}$$

Substituting this back into the previous inequality:

$$f(x) - f(x') \leq \|\nabla f(x)\| \sqrt{\frac{2}{\alpha}(f(x) - f(x'))}$$

Regret Bound FTRL

Let $a = f(x) - f(x')$. Then:

$$a \leq \|\nabla f(x)\| \sqrt{\frac{2a}{\alpha}}$$

$$a \leq \frac{2}{\alpha} \|\nabla f(x)\|^2$$

So we get:

$$f(x) - f(x') \leq \frac{1}{\alpha} \|\nabla f(x)\|^2$$

The lemma is thus proven:

$$0 \leq f(x) - f(x') \leq \frac{1}{\alpha} \|\nabla f(x)\|^2$$

Regret Bound FTRL

C.4 Bounding the Regret

Using the key lemma:

$$\ell_t(x_t) - \ell_t(x_{t+1}) \leq \frac{1}{\alpha}\|\nabla \ell_t(x_t)\|^2$$

Summing over all t from 1 to T:

$$\sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(x_{t+1})) \leq \frac{1}{\alpha}\sum_{t=1}^{T}\|\nabla \ell_t(x_t)\|^2$$

We can write the regret as:

$$R = \sum_{t=1}^{T}\ell_t(x_t) - \min_{x\in\Omega}\sum_{t=1}^{T}\ell_t(x)$$

Notice that:

$$\sum_{t=1}^{T}\ell_t(x_t) - \min_{x\in\Omega}\sum_{t=1}^{T}\ell_t(x) \leq \sum_{t=1}^{T}(\ell_t(x_t) - \ell_t(x_{t+1}))$$

Regret Bound FTRL

Combining this with our previous bound, we get:

$$\sum_{t=1}^{T} \ell_t(x_t) - \min_{x \in \Omega} \sum_{t=1}^{T} \ell_t(x) \leq \frac{1}{\alpha} \sum_{t=1}^{T} \|\nabla \ell_t(x_t)\|^2$$

For the FTRL algorithm, $\alpha = \frac{1}{\eta}$ for the regularizer ϕ, thus:

$$R \leq \eta \sum_{t=1}^{T} \|\nabla \ell_t(x_t)\|^2$$

Regret Bound FTRL

C.5 Final Regret Bound

If $\|\nabla \ell_t(x_t)\| \leq G$ for all t, then:

$$R \leq \frac{D}{\eta} + \eta \sum_{t=1}^{T} \|\nabla \ell_t(x_t)\|^2$$

$$O(\sqrt{T})$$

Choosing $\eta = \sqrt{\frac{D}{TG^2}}$ gives:

$$R \leq 2G\sqrt{TD}$$

Thus, the regret of the FTRL algorithm is bounded by $R \leq 2G\sqrt{TD}$, where D is the range of the regularizer ϕ, and G is an upper bound on the gradient norms (Ma et al., 2018).

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

- Online Lerner → Modelparameter
- O2NC

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

- Zu optimierende Funktion: $F(x)$
- Iterationen: t
- Parameter: x_t
- Update: Δ_t

Standard Update Regel:

$$x_{t+1} = x_t + \Delta_t$$

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

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

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

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

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

Optimaler Fall: $\Delta_t \approx -\nabla F(x_t)$

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

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

Optimaler Fall: $\Delta_t \approx -\nabla F(x_t)$

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

Adam mit O2NC herstellen, Ahn et al. (2024)

$$\Delta_t = -\eta_t \sum_{i=1}^{t} g_i$$

$$\eta_t = \frac{\alpha}{\sqrt{\sum_{i=1}^{t} g_i^2}}$$

$$\Delta_t = -\alpha \frac{\sum_{i=1}^{t} \beta_1^{t-i} g_i}{\sqrt{\sum_{i=1}^{t} \beta_2^{t-i} g_i^2}}$$

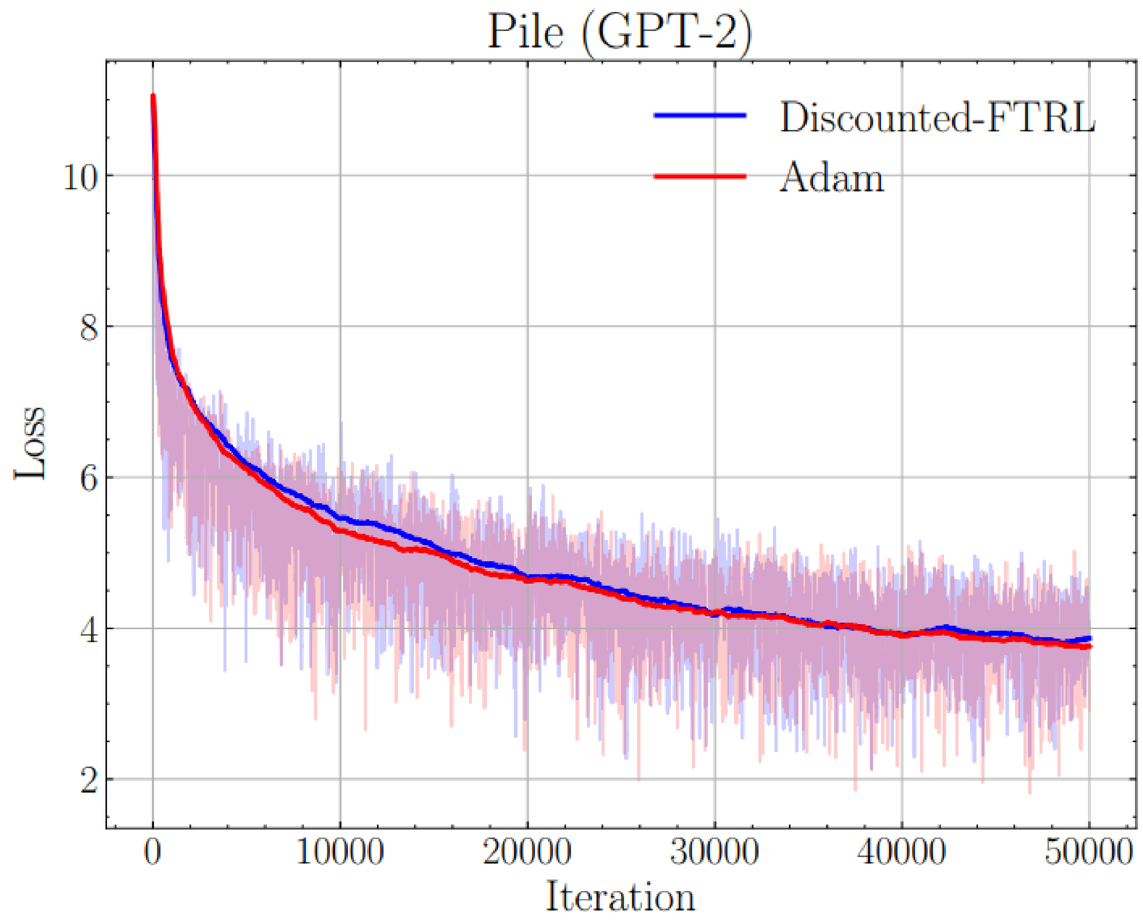


Figure 14: Loss function (smoothed with time-weighted EMA) with $\eta = 3 \times 10^{-4}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ for both Adam and FTRL.

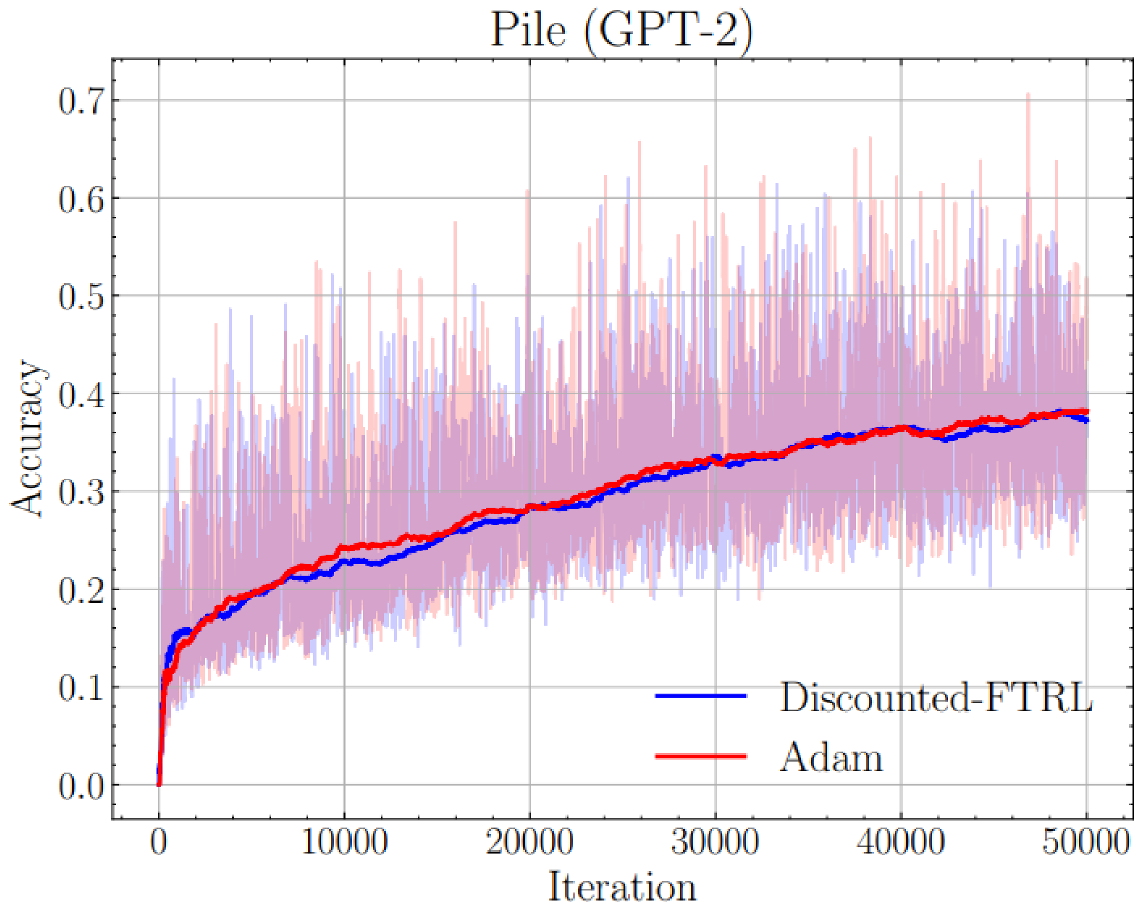


Figure 15: Accuracy (smoothed with time-weighted EMA) with $\eta = 3 \times 10^{-4}$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ for both Adam and FTRL.

Recap

- Online Learning
  - FTL, FTRL
  - Regret
  - Regret Bounds
- O2NC

Optimismus im Online Lernen (FTRL)

$$x_{t+1} = x_0 - \eta_t \left(\sum_{i=0}^{t} g_i + h_t \right)$$

$$x_{t+1} = x_t - \eta_t g_t + \eta_{t-1} h_{t-1} - \eta_t h_t$$

Optimismus im Online Lernen (OMD)

$$x_{t+1} = x_t - \eta_t(g_t + h_t - h_{t-1})$$

Ziel: $h_t \approx g_{t+1}$

Theoretische Analyse der Optimistischen Lerner

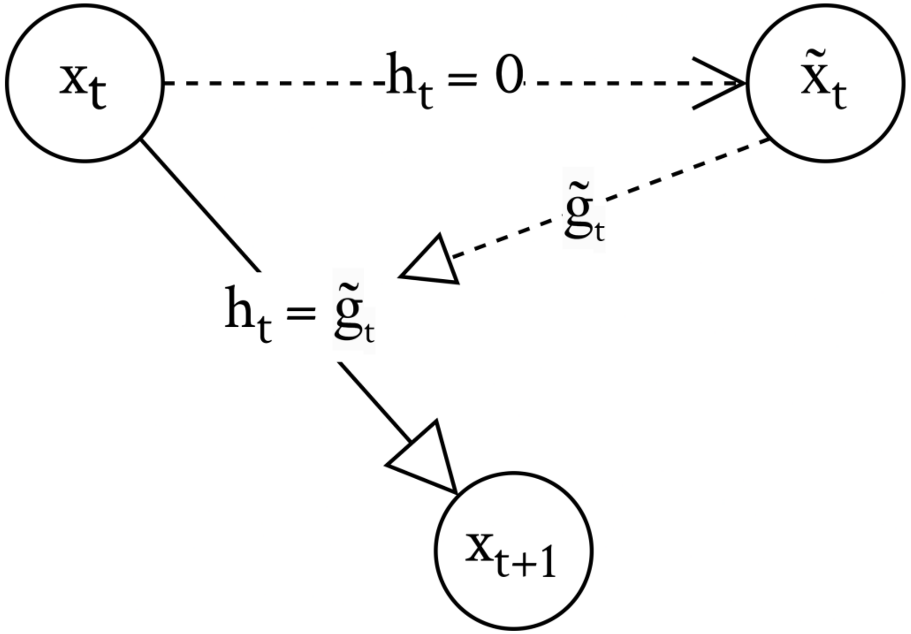
$$R_T \leq \frac{1}{2\eta}\|x_1 - u\|^2 + \frac{\eta}{2}\sum_{t=1}^{T}\|\nabla \ell_{t+1}(x_{t+1}) - h_t\|^2$$

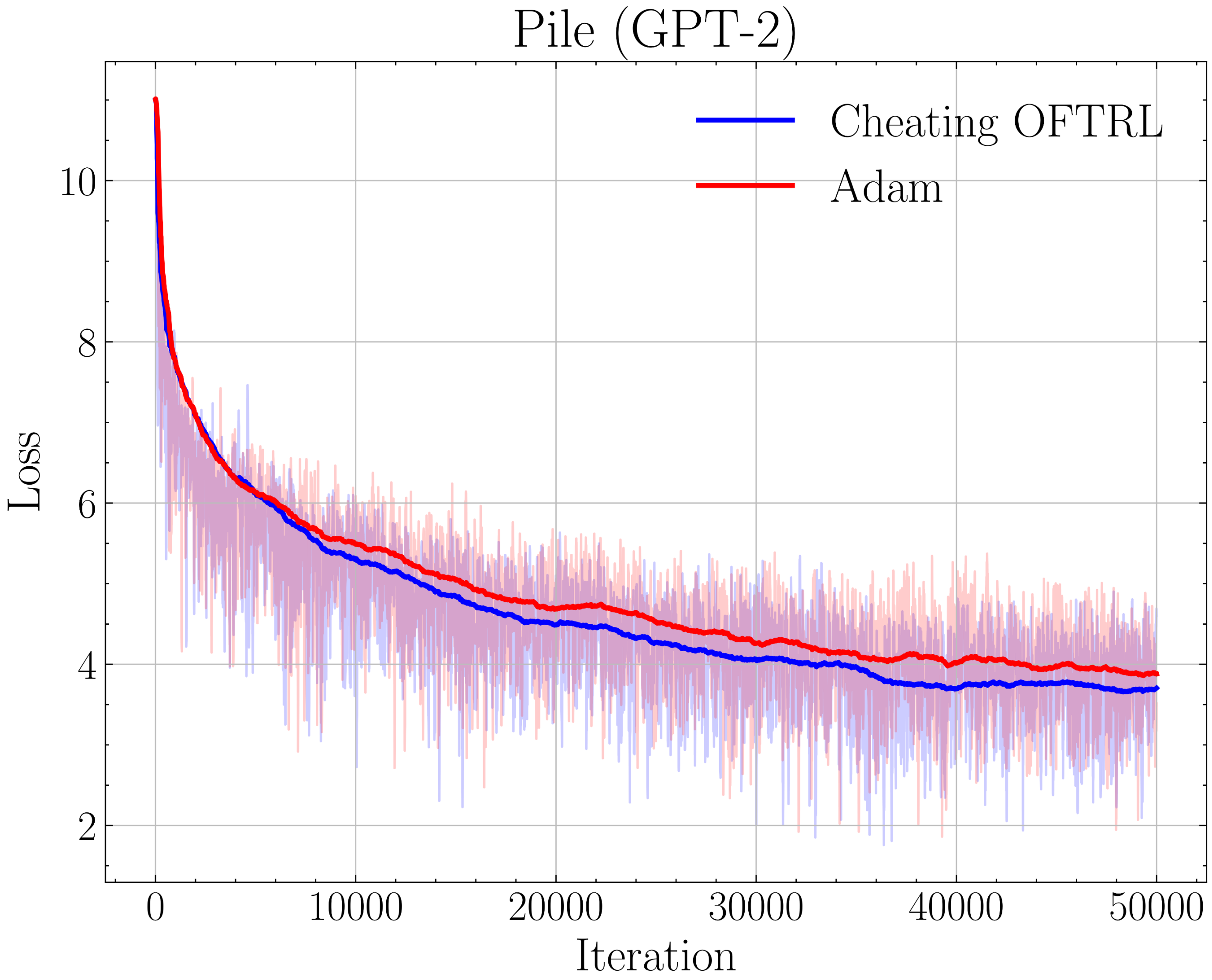
Theorem 7 von Cutkosky et al. (2023)

$$\mathbb{E}[F(x_M)] = F(x_0) + \mathbb{E}\left[\sum_{n=1}^{M}\langle g_n, \Delta_n - u_n\rangle\right] + \mathbb{E}\left[\sum_{n=1}^{M}\langle g_n, u_n\rangle\right]$$

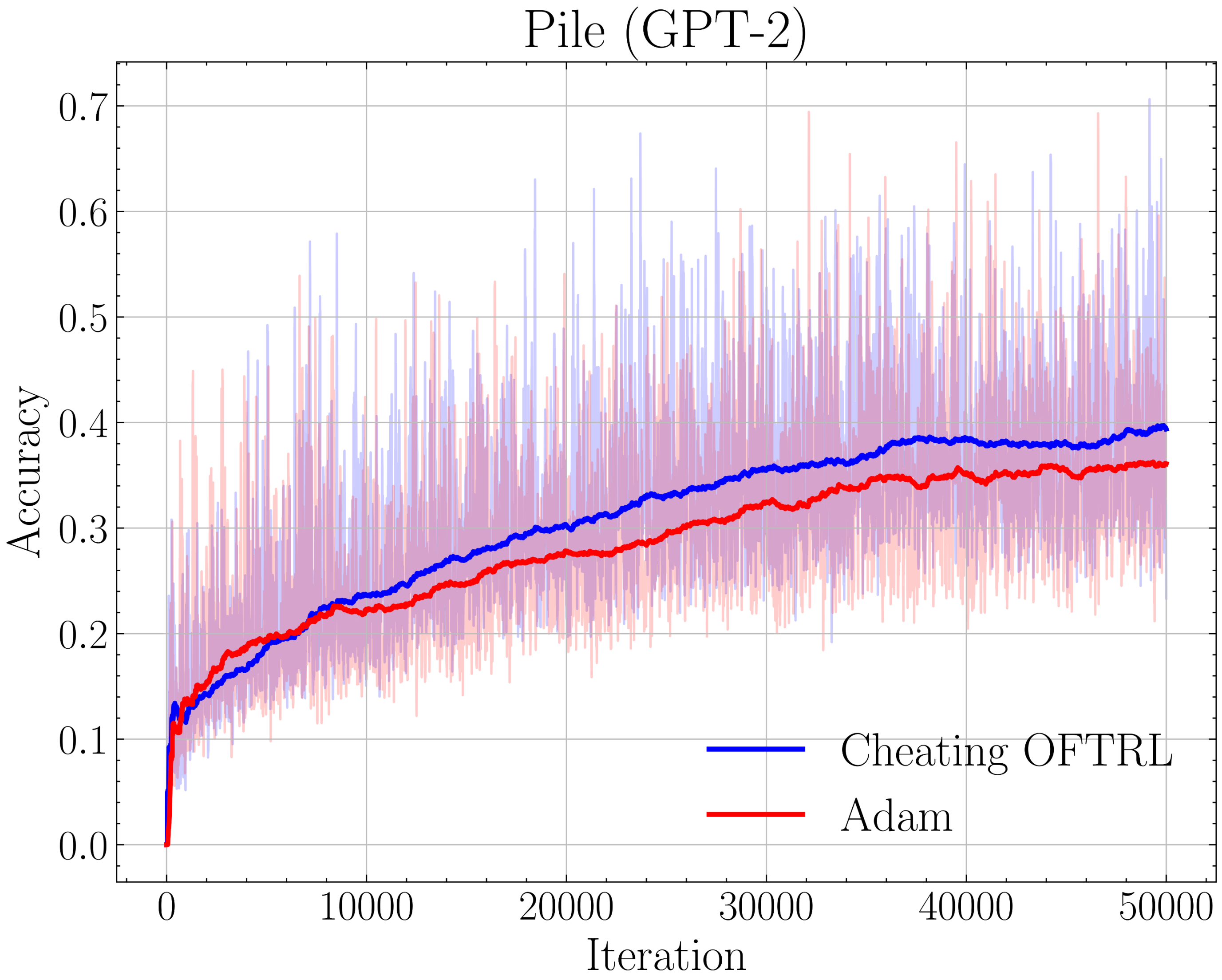
Wie können wir jetzt **h** bestimmen???

- Proof of Concept:





Pile (GPT-2)
Cheating OFTRL
Adam
Loss
Iteration
2
4
6
8
10
0
10000
20000
30000
40000
50000



Pile (GPT-2)
Accuracy
Iteration
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0
10000
20000
30000
40000
50000
Cheating OFTRL
Adam

Jetzt ohne Cheating!

Formula	Hyperparameter	EMA Loss
$h_{t+1} = 0$ (Adam)	$\eta = 0.0003$	3.89
$h_{t+1} = g_t$	$\eta = 0.0003$	3.90
$h_{t+1} = \beta h_t + (1-\beta) g_t$	$\eta = 0.0003,\ \beta = 0.5$	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 1: Hint update methods, their formulas, hyperparameters, and time-weighted EMA of the train loss at the 50,000th iteration (same computational budget).

Hint Qualität

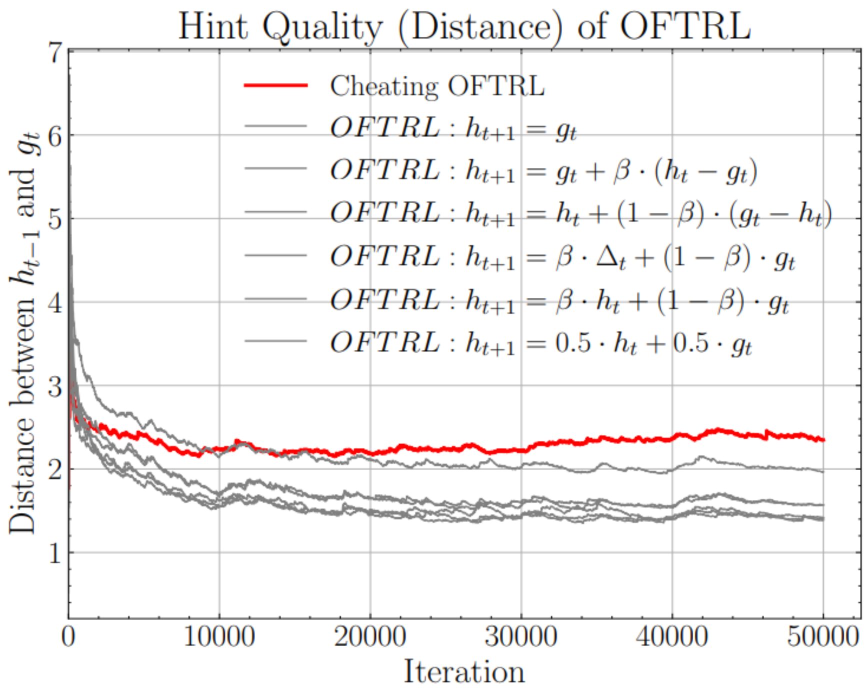


Figure 6: Distance $\|h_{t-1} - g_t\|$ of OFTRL according to table 1 (smoothed with time-weighted EMA).

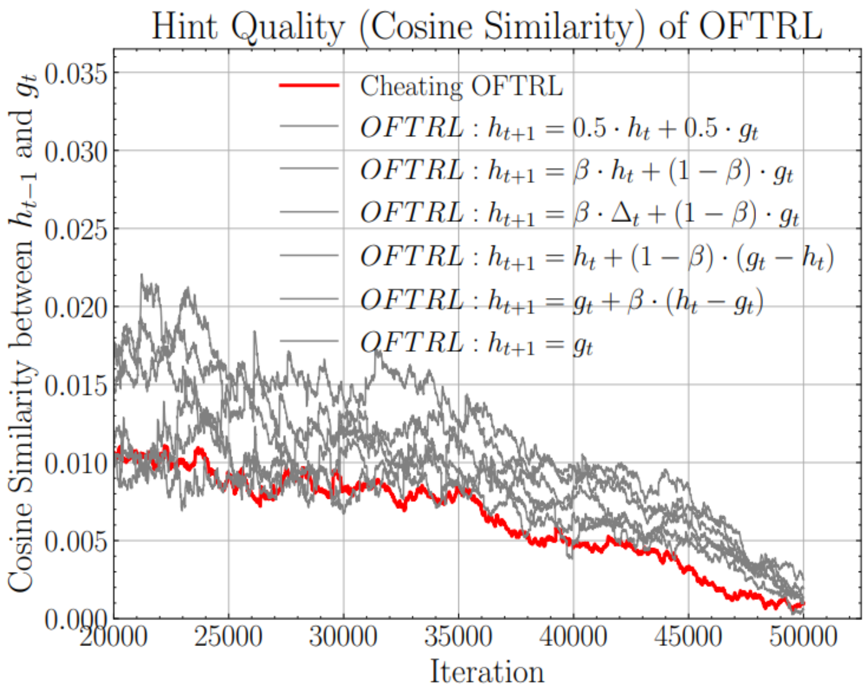


Figure 7: Cosine Similarity $\frac{h_{t-1} \cdot g_t}{\|h_{t-1}\| \|g_t\|}$ of OFTRL according to table 1 (smoothed with time-weighted EMA).

Grössere Batches

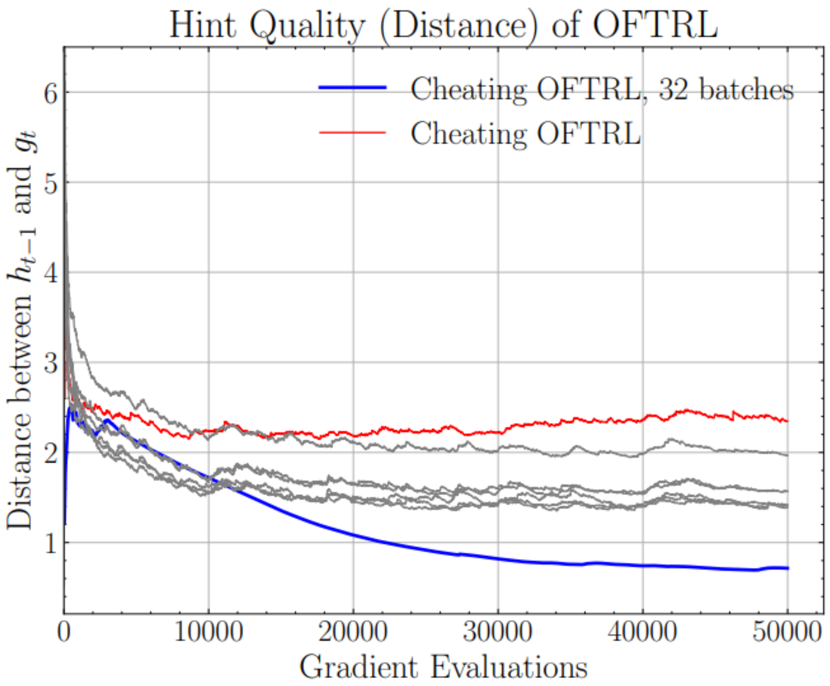
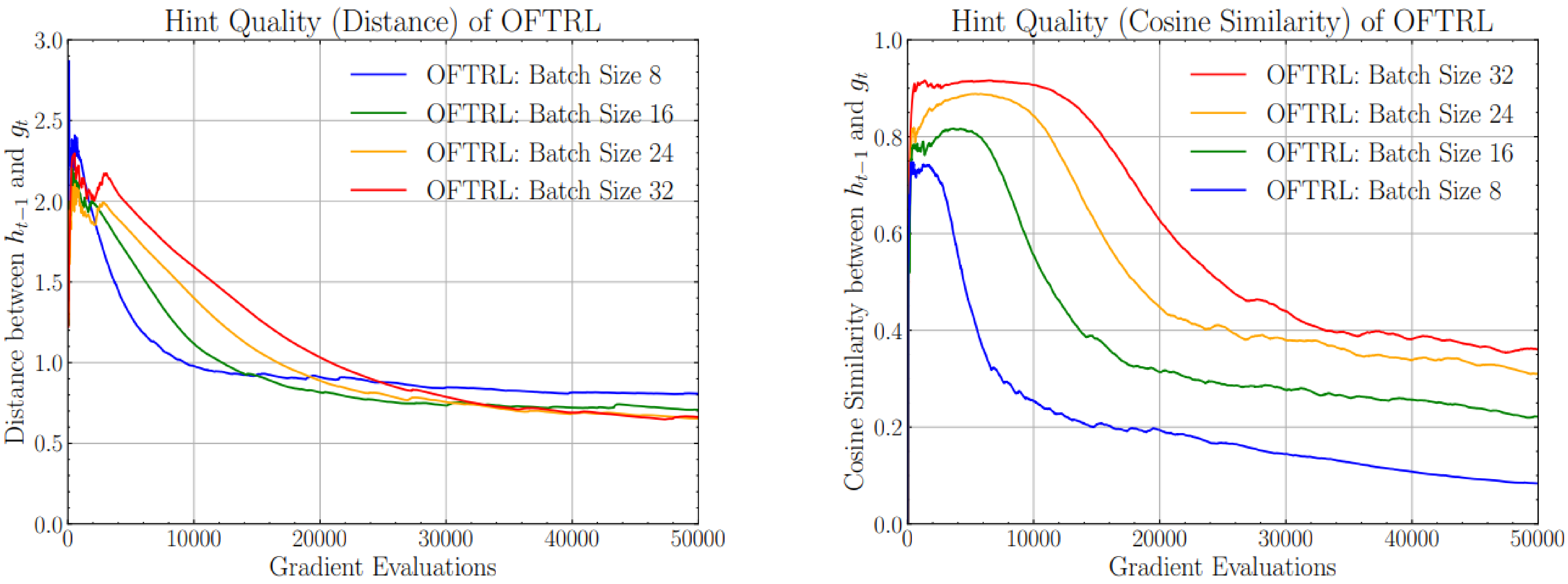


Figure 8: Distance $\|h_{t-1} - g_t\|$ of OFTRL. Each line represents a different hint calculation method, according to Table 1, using a single batch, except for the method that uses 32 batches (smoothed with time-weighted EMA).


Hint Quality (Distance) of OFTRL
Distance between h_{t-1} and g_t
Gradient Evaluations
OFTRL: Batch Size 8
OFTRL: Batch Size 16
OFTRL: Batch Size 24
OFTRL: Batch Size 32
Hint Quality (Cosine Similarity) of OFTRL
Cosine Similarity between h_{t-1} and g_t
Gradient Evaluations
OFTRL: Batch Size 32
OFTRL: Batch Size 24
OFTRL: Batch Size 16
OFTRL: Batch Size 8

Fazit

- O2NC erlaubt neue Analyse von Optimizern
- Erster Prototyp
  - Grössere Batches
  - Bessere Hint Generierungsmethoden

Vielen Dank!

Noch Fragen?

Luca Mezger | lucamezger.com/ma
28.11.2024

Fachreferat, HSLU

Probleme

- Nicht-differenzierbare Punkte
- Noisy loss function
- Lokale Minima und Sattelpunkte
- Verschwindende / Explodierende Gradienten
- Hohe Dimensionalität Hyperparameter Tuning
- Grosse datasets
- Sparse data

FTL Worst case Scenario

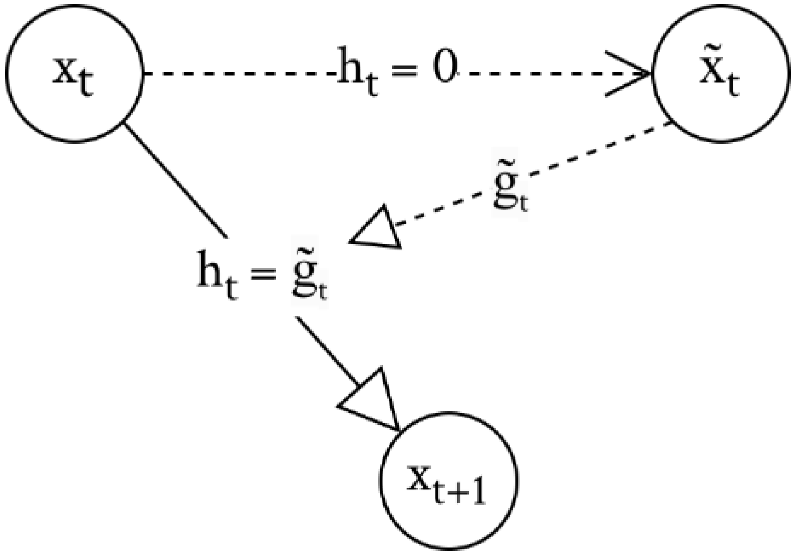
Imagine a similar game to the one described in section 2.2. Let the feasible set $V = [-1, 1]$. Define $\ell_1(x) = \frac{1}{2}x$ and let ℓ_t for $t = 2, \ldots, T$ alternate between $-x$ and x. Thus,

$$\sum_{t=1}^{T} \ell_t(x) = \begin{cases} \frac{1}{2}x & \text{if } t \text{ is odd,} \\ -\frac{1}{2}x & \text{if } t \text{ is even.} \end{cases}$$

The FTL strategy will constantly switch between $x_t = -1$ and $x_t = 1$, making the incorrect decision at every iteration t. This demonstrates that the seemingly intuitive FTL approach fails in this scenario due to its instability (Hazan et al., 2016). We can improve this algorithm if we introduce a regularizer to reduce the instability. This algorithm is called Follow-The-Regularized-Leader (FTRL).

POC OFTRL

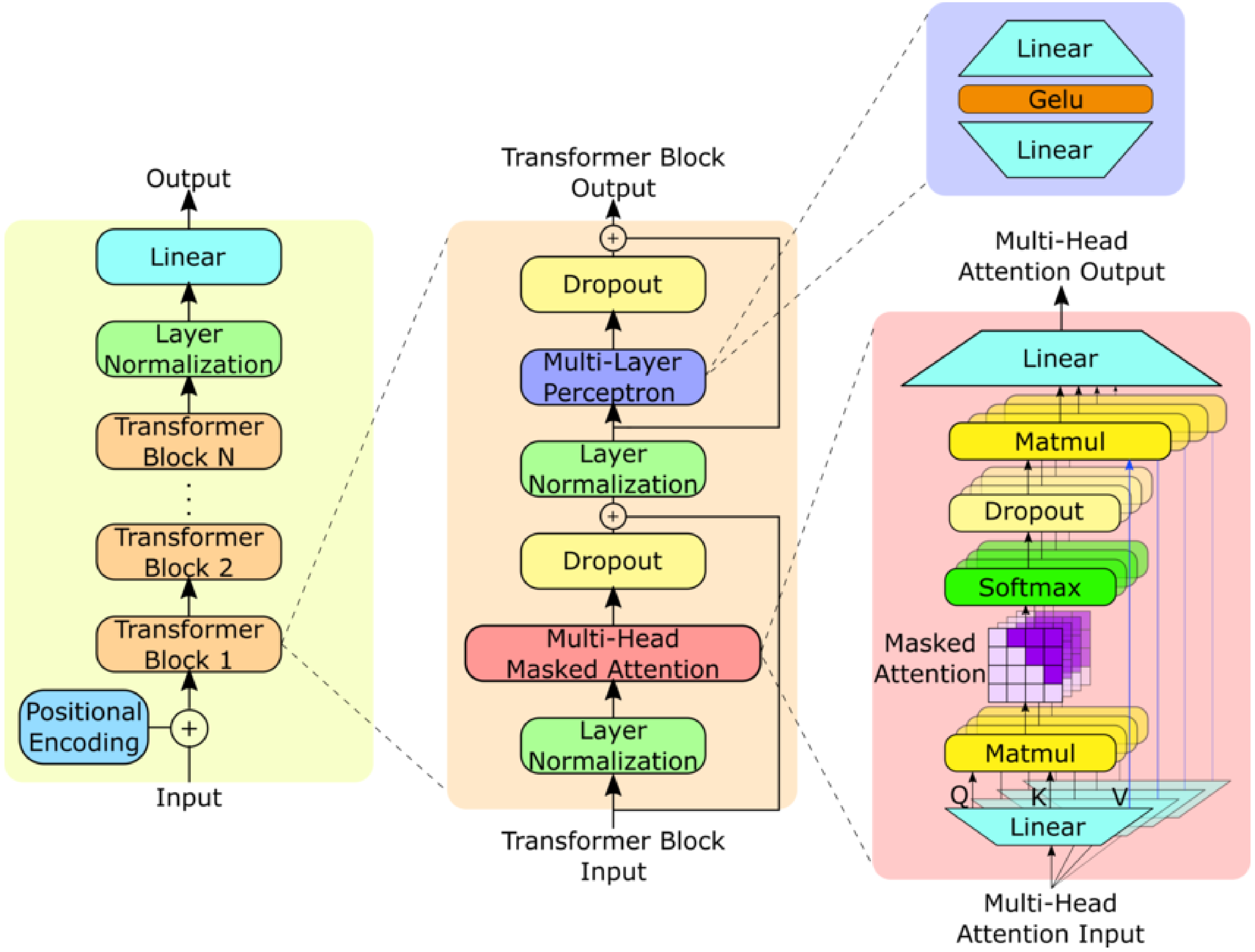
bias correction: $$\Delta_t = -\alpha \frac{\frac{\beta_1 \sum_{i=1}^{t} \beta_1^{t-i} g_i + (1-\beta_1) h_t}{1-\beta_1^t}}{\sqrt{\frac{\beta_2 \sum_{i=1}^{t} \beta_2^{t-i} g_i^2 + (1-\beta_2) h_t^2}{1-\beta_2^t}}}$$



$$\Delta_t = -\alpha \frac{\sum_{i=1}^{t} \beta_1^{t-i} g_i}{\sqrt{\sum_{i=1}^{t} \beta_2^{t-i} g_i^2}}$$

$$\Delta_t = -\alpha \frac{\beta_1 \sum_{i=1}^{t} \beta_1^{t-i} g_i + (1-\beta_1) h_t}{\sqrt{\beta_2 \sum_{i=1}^{t} \beta_2^{t-i} g_i^2 + (1-\beta_2) h_t^2}}$$

GPT-2 Model



Output
Linear
Layer Normalization
Transformer Block N
Transformer Block 2
Transformer Block 1
Positional Encoding
Input
Transformer Block Output
Dropout
Multi-Layer Perceptron
Layer Normalization
Dropout
Multi-Head Masked Attention
Layer Normalization
Transformer Block Input
Linear
Gelu
Linear
Multi-Head Attention Output
Linear
Matmul
Dropout
Softmax
Masked Attention
Matmul
Q
K
V
Linear
Multi-Head Attention Input