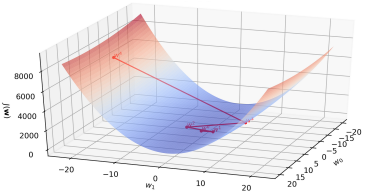
`Adam` Optimization with Adaptive Batch Selection

Gyu Yeol Kim and **Min-hwan Oh**

Seoul National University

Optimization Method

- We want to minimize the function $f(\theta)$: $\underset{\theta}{\text{minimize}}\, f(\theta)$

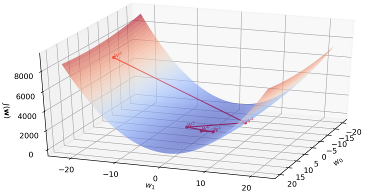


- `ADAM` optimizer (Kingma and Ba, 2015)
  - Uses past gradient (momentum)
  - Adapt to individual parameters (adaptive learning rate)

$\rightarrow$ One of the most widely used optimizers for deep learning!

Optimization Method

- We want to minimize the function $f(\theta)$: $\underset{\theta}{\text{minimize}}\, f(\theta)$



- `ADAM` optimizer (Kingma and Ba, 2015)
  - Uses past gradient (momentum)
  - Adapt to individual parameters (adaptive learning rate)

$\rightarrow$ One of the most widely used optimizers for deep learning!

Limitations of ADAM and its variants

Original ADAM uses uniform sampling over dataset

- Treats each training sample equally.
- However, different samples can influence model updates differently.
- Full dataset sweeps ⇒ possible inefficient convergence
- Same issues exist in the follow-up works (Reddi et al., 2018; Huang et al., 2019; Chen et al., 2023)

Limitations of ADAM and its variants

Original ADAM uses uniform sampling over dataset

- Treats each training sample equally.
- However, different samples can influence model updates differently.
- Full dataset sweeps ⇒ possible inefficient convergence
- Same issues exist in the follow-up works (Reddi et al., 2018; Huang et al., 2019; Chen et al., 2023)

Adaptive approach: Adam with Bandit Sampling (Liu et al., 2020)

- Learns importance of samples dynamically during training
- Adaptive batch selection using **multi-armed bandit** (MAB) algorithm
  - Treats each training sample as an arm in MAB
  - Partial feedbacks : per-sample gradients for selected batch

Limitations of ADAM and its variants

Original ADAM uses uniform sampling over dataset

- Treats each training sample equally.
- However, different samples can influence model updates differently.
- Full dataset sweeps $\Rightarrow$ possible inefficient convergence
- Same issues exist in the follow-up works (Reddi et al., 2018; Huang et al., 2019; Chen et al., 2023)

Adaptive approach: Adam with Bandit Sampling (Liu et al., 2020)

- Learns importance of samples dynamically during training
- Adaptive batch selection using **multi-armed bandit** (MAB) algorithm
  - Treats each training sample as an arm in MAB
  - Partial feedbacks : per-sample gradients for selected batch
- However, proposed under limited settings: features assumed to follow a doubly heavy-tailed distribution
- Incorrect convergence analysis
- Poor numerical performances

Research Motivations

Research Questions:

- Can we design a **provably correct** and **practical** `Adam` optimization algorithm with **convergence guarantees**?
- Can we show that our new `Adam` optimization algorithm with even **faster convergence**?

Performance Measure: Regret

Online Optimization as **Regret Minimization** framework

Performance Measure: Regret

Online Optimization as **Regret Minimization** framework

- **Cumulative Regret**

  Consider an online optimization algorithm π that generates a sequence of model parameters $\theta_1, \theta_2, \dots, \theta_T$ over T iterations. The cumulative regret after T iterations:

$$\mathcal{R}^{\pi}(T) := \mathbb{E}\left[\sum_{t=1}^{T} f(\theta_t; \mathcal{D}) - T \cdot \min_{\theta \in \mathbb{R}^d} f(\theta; \mathcal{D})\right]$$

- Regret defined under the whole dataset $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^{n}$ where $f(\theta_t; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; x_i, y_i)$

Performance Measure: Regret

Online Optimization as **Regret Minimization** framework

- **Cumulative Regret**

  Consider an online optimization algorithm π that generates a sequence of model parameters $\theta_1, \theta_2, \ldots, \theta_T$ over T iterations. The cumulative regret after T iterations:

$$\mathcal{R}^{\pi}(T) := \mathbb{E}\left[\sum_{t=1}^{T} f(\theta_t; \mathcal{D}) - T \cdot \min_{\theta \in \mathbb{R}^d} f(\theta; \mathcal{D})\right]$$

- Regret defined under the whole dataset $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^{n}$ where $f(\theta_t; \mathcal{D}) := \frac{1}{n}\sum_{i=1}^{n} \ell(\theta; x_i, y_i)$

Want: Design an algorithm that gives sublinear regret $\mathcal{R}^{\pi}(T) = o(T)$.

Performance Measure: Regret

Online Optimization as **Regret Minimization** framework

- **Cumulative Regret**

  Consider an online optimization algorithm π that generates a sequence of model parameters $\theta_1, \theta_2, \ldots, \theta_T$ over T iterations. The cumulative regret after T iterations:

$$\mathcal{R}^\pi(T) := \mathbb{E}\left[\sum_{t=1}^{T} f(\theta_t; \mathcal{D}) - T \cdot \min_{\theta \in \mathbb{R}^d} f(\theta; \mathcal{D})\right]$$

- Regret defined under the whole dataset $\mathcal{D} := \{(x_i, y_i)\}_{i=1}^n$ where $f(\theta_t; \mathcal{D}) := \frac{1}{n} \sum_{i=1}^{n} \ell(\theta; x_i, y_i)$

Want: Design an algorithm that gives sublinear regret $\mathcal{R}^\pi(T) = o(T)$.

- e.g., $\mathcal{R}^\pi(T) = \mathcal{O}(\sqrt{T}) \implies$ Average regret $\frac{\mathcal{R}^\pi(T)}{T} = \mathcal{O}(\frac{1}{\sqrt{T}})$

Contributions

- New optimization algorithm
  - Propose Adam-based optimization with adaptive sample selection using combinatorial bandit method
  - Adam with Combinatorial Bandit Sampling (AdamCB)

Contributions

- New optimization algorithm
  - Propose Adam-based optimization with adaptive sample selection using combinatorial bandit method
  - Adam with Combinatorial Bandit Sampling (AdamCB)
- Convergence guarantee with improved regret performance
  - Theoretically show a sub-linear regret bound for AdamCB that achieves sharper regret than existing AdamBS and Adam

Contributions

- New optimization algorithm
  - Propose Adam-based optimization with adaptive sample selection using combinatorial bandit method
  - Adam with Combinatorial Bandit Sampling (AdamCB)
- Convergence guarantee with improved regret performance
  - Theoretically show a sub-linear regret bound for AdamCB that achieves sharper regret than existing AdamBS and Adam
- Practically efficient
  - Numerical experiments show that AdamCB outperforms the existing Adam-based methods

Contributions

- New optimization algorithm
  - Propose Adam-based optimization with adaptive sample selection using combinatorial bandit method
  - Adam with Combinatorial Bandit Sampling (AdamCB)
- Convergence guarantee with improved regret performance
  - Theoretically show a sub-linear regret bound for AdamCB that achieves sharper regret than existing AdamBS and Adam
- Practically efficient
  - Numerical experiments show that AdamCB outperforms the existing Adam-based methods

Theoretical guaranteed (provably efficient) and practically superior Adam optimizer

Adam with Combinatorial Bandit Sampling (AdamCB)

Algorithm 1: Adam with Combinatorial Bandit Sampling (`AdamCB`)

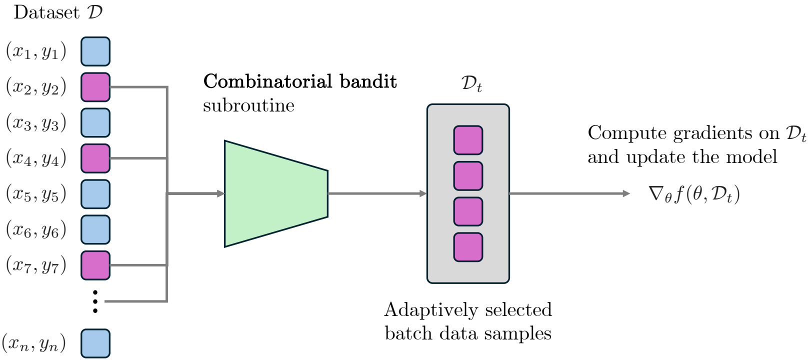
Input: learning rate $\{\alpha_t\}_{t=1}^T$, decay rates $\{\beta_{1,t}\}_{t=1}^T$, β_2, batch size K, exploration parameter $\gamma \in [0, 1)$

Initialize: model parameters θ_0, first moment estimate $m_0 \leftarrow 0$, second moment estimate $v_0 \leftarrow 0, \hat{v}_0 \leftarrow 0$, sample weights $w_{i,0} \leftarrow 1$ for all $i \in [n]$

for $t = 1$ **to** T **do**

- $J_t, p_t, S_{\text{null},t} \leftarrow$ `Batch-Selection`(w_{t-1}, K, γ) (Algorithm 2)
- Compute unbiased gradient estimate g_t with respect to J_t using Eq.(8)
- $m_t \leftarrow \beta_{1,t} m_{t-1} + (1 - \beta_{1,t}) g_t$
- $v_t \leftarrow \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$
- $\hat{v}_1 \leftarrow v_1$, $\hat{v}_t \leftarrow \max\left\{ \frac{(1-\beta_{1,t})^2}{(1-\beta_{1,t-1})^2} \hat{v}_{t-1}, v_t \right\}$ if $t \geq 2$
- $\theta_{t+1} \leftarrow \theta_t - \alpha_t \frac{m_t}{\sqrt{\hat{v}_t} + \epsilon}$
- $w_t \leftarrow$ `Weight-Update`$(w_{t-1}, p_t, J_t, \{g_{j,t}\}_{j \in J_t}, S_{\text{null},t}, \gamma)$ (Algorithm 3)

Adam with Combinatorial Bandit Sampling (AdamCB) Illustration



Regret Analysis

Theorem (Regret Bound of AdamCB)

Cumulative regret of AdamCB over T iterations with mini-batch size K is upper-bounded by:

$$\mathcal{R}^{\texttt{AdamCB}}(T) \leq \mathcal{O}\left(d\sqrt{T} + \frac{\sqrt{d}}{n^{3/4}} \left(\frac{T}{K} \ln \frac{n}{K} \right)^{1/4} \right)$$

Regret Analysis

Theorem (Regret Bound of AdamCB)

Cumulative regret of AdamCB over T iterations with mini-batch size K is upper-bounded by:

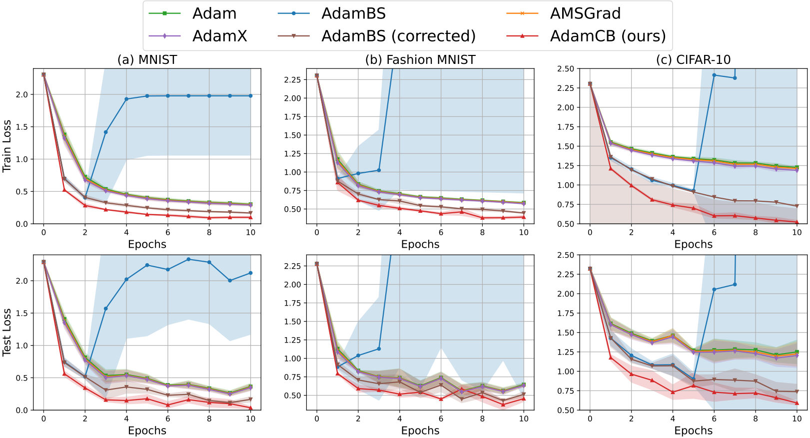
$$\mathcal{R}^{\textit{AdamCB}}(T) \leq \mathcal{O}\left(d\sqrt{T} + \frac{\sqrt{d}}{n^{3/4}}\left(\frac{T}{K}\ln\frac{n}{K}\right)^{1/4}\right)$$

Comparison:

Optimizer	Convergence Rate
AdamX (Tran et al., 2019) (variant of Adam)	$\mathcal{O}\left(d\sqrt{T} + \frac{\sqrt{d}}{n^{1/2}}\sqrt{T}\right)$
AdamBS (Liu et al., 2020) (corrected)	$\mathcal{O}\left(d\sqrt{T} + \frac{\sqrt{d}}{n^{3/4}}\left(T\ln n\right)^{1/4}\right)$
AdamCB (**Ours**)	$\mathcal{O}\left(d\sqrt{T} + \frac{\sqrt{d}}{n^{3/4}}\left(\frac{T}{K}\ln\frac{n}{K}\right)^{1/4}\right)$

- To our best knowldge, fastest regret convergence with AdamCB
- Sub-linear in T $\Rightarrow$ Convergence Guarantee
- Benefits from mini-batch size K

Numerical Experiments



Conclusion

- **AdamCB**
  - Novel method that adpats **combinatorial bandit sampling** to Adam optimization.
  - Introduces a batch selection strategy for sampling without replacement.

Conclusion

- **AdamCB**
  - Novel method that adpats **combinatorial bandit sampling** to Adam optimization.
  - Introduces a batch selection strategy for sampling without replacement.
- **Impact on Model Convergence**
  - Significantly accelerates model convergence
  - Provides rigorous theoretical analysis on regret bound
- **Better numerical performances compared to existing methods**

Achieves both theoretical and practical efficiency!

References I

Chen, Y., Li, Z., Zhang, L., Du, B., and Zhao, H. (2023). Bidirectional looking with a novel double exponential moving average to adaptive and non-adaptive momentum optimizers. In International Conference on Machine Learning, pages 4764–4803. PMLR.

Huang, H., Wang, C., and Dong, B. (2019). Nostalgic adam: weighting more of the past gradients when designing the adaptive learning rate. In Proceedings of the 28th International Joint Conference on Artificial Intelligence, pages 2556–2562.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings.

Liu, R., Wu, T., and Mozafari, B. (2020). Adam with bandit sampling for deep learning. Advances in Neural Information Processing Systems, 33:5393–5404.

Reddi, S. J., Kale, S., and Kumar, S. (2018). On the convergence of adam and beyond. In International Conference on Learning Representations.

Tran, P. T. et al. (2019). On the convergence proof of amsgrad and a new version. IEEE Access, 7:61706–61716.