Stochastic Approximation with Block Coordinate Optimal Stepsizes

Tao Jiang Cornell University

Lin Xiao Meta FAIR

CRM Workshop:
Optimization and Learning: Theory and Applications

Centre de Recherches Mathématiques, Montréal
May 27-30, 2025

Outline

- background and motivation
- **BCOS: block coordinate optimal stepsizes**
- instantiations of BCOS
  - **search directions:** stochastic gradient, momentum, preconditioned, ...
  - **2nd moment estimators:** EMA, conditional estimator
- numerical experiments
- **convergence analysis**

Stochastic optimization

stochastic optimization problem

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \;\; F(x) := \mathbf{E}_\xi[f(x, \xi)]$$

classical stochastic approximation method [Robbins and Monro, 1951]

$$x_{t+1} = x_t - \gamma_t \, d_t$$

- $d_t \in \mathbf{R}^n$: stochastic search direction
  - stochastic gradient: $d_t = \nabla f(x_t, \xi_t)$
  - momentum (EMA): $d_t = \beta d_{t-1} + (1 - \beta) \nabla f(x_t, \xi_t)$

Stochastic optimization

stochastic optimization problem

$$\underset{x \in \mathbf{R}^n}{\text{minimize}} \;\; F(x) := \mathbf{E}_\xi[f(x, \xi)]$$

classical stochastic approximation method [Robbins and Monro, 1951]

$$x_{t+1} = x_t - \gamma_t \, d_t$$

- $d_t \in \mathbf{R}^n$: stochastic search direction
  - stochastic gradient: $d_t = \nabla f(x_t, \xi_t)$
  - momentum (EMA): $d_t = \beta d_{t-1} + (1 - \beta) \nabla f(x_t, \xi_t)$
- assumptions: **convexity and/or smoothness**
- convergence analysis: **guarantees in expectation, rate of convergence**

The myth about Adam and AdamW

Adam [Kingma and Ba, 2015] and AdamW [Loshchilov and Hutter, 2019]

$$x_{t+1} = x_t - \frac{\alpha_t}{\sqrt{v_t} + \epsilon} \odot m_t - \alpha_t \lambda x_t$$

where m_t and v_t are EMA of g_t and g_t^2:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

The myth about Adam and AdamW

Adam [Kingma and Ba, 2015] and AdamW [Loshchilov and Hutter, 2019]

$$x_{t+1} = x_t - \frac{\alpha_t}{\sqrt{v_t} + \epsilon} \odot m_t - \alpha_t \lambda x_t$$

where m_t and v_t are EMA of g_t and g_t^2:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

myths about Adam(W)

- role of 2nd moment: diagonal preconditioning or something else?
- choice of hyper-parameters (e.g., why $\beta_1 = 0.9$ and $\beta_2 = 0.99$?)
- why AdamW performs much better than Adam?
- what is a convincing convergence analysis (that can explain all of above)?

The myth about Adam and AdamW

Adam [Kingma and Ba, 2015] and AdamW [Loshchilov and Hutter, 2019]

$$x_{t+1} = x_t - \frac{\alpha_t}{\sqrt{v_t} + \epsilon} \odot m_t - \alpha_t \lambda x_t$$

where m_t and v_t are EMA of g_t and g_t^2:

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$
$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

myths about Adam(W)

- role of 2nd moment: diagonal preconditioning or something else?
- choice of hyper-parameters (e.g., why $\beta_1 = 0.9$ and $\beta_2 = 0.99$?)
- why AdamW performs much better than Adam?
- what is a convincing convergence analysis (that can explain all of above)?

motivation: demystify Adam(W) and derive **better/simpler** algorithms

Stochastic Approximation

- find x_* such that $G(x_*) = 0$ where $G : \mathbf{R}^n \to \mathbf{R}^n$ is defined as

$$G(x) := \mathbf{E}_\xi[g(x, \xi)]$$

- stochastic approximation [Robbins and Monro, 1951]

$$x_{t+1} = x_t - \alpha_t \, g(x_t, \xi_t)$$

Stochastic Approximation

- find x_* such that $G(x_*) = 0$ where $G : \mathbf{R}^n \to \mathbf{R}^n$ is defined as

$$G(x) := \mathbf{E}_\xi[g(x, \xi)]$$

- stochastic approximation [Robbins and Monro, 1951]

$$x_{t+1} = x_t - \alpha_t \, g(x_t, \xi_t)$$

- **rich literature on convergence analysis**
  - convergence in mean-square sense, almost sure convergence
  - rate of convergence (matching lower-bounds)

- **aiming condition** (weaker than convexity)

$$\big\langle x - x_*, \, \mathbf{E}_\xi[g(x, \xi)] \big\rangle > 0 \qquad \forall \, x \neq x_*$$

Stochastic Approximation

- find x_* such that $G(x_*) = 0$ where $G : \mathbf{R}^n \to \mathbf{R}^n$ is defined as

$$G(x) := \mathbf{E}_\xi[g(x, \xi)]$$

- stochastic approximation [Robbins and Monro, 1951]

$$x_{t+1} = x_t - \alpha_t \, g(x_t, \xi_t)$$

- **rich literature on convergence analysis**
  - convergence in mean-square sense, almost sure convergence
  - rate of convergence (matching lower-bounds)

- **aiming condition** (weaker than convexity)

$$\big\langle x - x_*, \, \mathbf{E}_\xi[g(x, \xi)] \big\rangle > 0 \qquad \forall \, x \neq x_*$$

- **BCOS: a general framework for algorithm design and analysis**
  - block-coordinate weighted aiming (neither convexity nor smoothness)
  - almost sure convergence (leveraging classic results in 1950's)

Outline

- background and motivation
- **BCOS: block coordinate optimal stepsizes**
- instantiations of BCOS
  - **search directions:** stochastic gradient, momentum, preconditioned, ...
  - **2nd moment estimators:** EMA, conditional estimator
- numerical experiments
- **convergence analysis**

A general form of stochastic approximation (SA)

SA with block-coordinate stepsizes

$$x_{t+1,k} = x_{t,k} - \gamma_{t,k} d_{t,k}, \qquad k = 1, \ldots, m$$

- block partition: $x_{t,k}, d_{t,k} \in \mathbf{R}^{n_k}$ for $k = 1, \ldots, m$, with $\sum_{i=1}^{m} n_k = n$
- each block share common stepsize $\gamma_{t,k} > 0$ for each $k = 1, \ldots, m$
- full-vector notation

$$x_{t+1} = x_t - s_t \odot d_t$$

where $s_t = [s_{t,1}, \ldots, s_{t,m}] \in \mathbf{R}^n$ with $s_{t,k} = \gamma_{t,k} \mathbf{1}_{n_k} \in \mathbf{R}^{n_k}$

A general form of stochastic approximation (SA)

SA with block-coordinate stepsizes

$$x_{t+1,k} = x_{t,k} - \gamma_{t,k} d_{t,k}, \qquad k = 1, \ldots, m$$

- block partition: $x_{t,k}, d_{t,k} \in \mathbf{R}^{n_k}$ for $k = 1, \ldots, m$, with $\sum_{i=1}^{m} n_k = n$
- each block share common stepsize $\gamma_{t,k} > 0$ for each $k = 1, \ldots, m$
- full-vector notation

$$x_{t+1} = x_t - s_t \odot d_t$$

where $s_t = [s_{t,1}, \ldots, s_{t,m}] \in \mathbf{R}^n$ with $s_{t,k} = \gamma_{t,k} \mathbf{1}_{n_k} \in \mathbf{R}^{n_k}$

examples of search direction $d_t \in \mathbf{R}^n$

- stochastic gradient: $d_t = \nabla f(x_t, \xi_t)$
- stochastic momentum: $d_t = \beta d_{t-1} + (1-\beta)\nabla f(x_t, \xi_t)$
- can also work with preconditioned directions and Muon

Expected distance to an optimal point

distance of x_{t+1} to an optimal point x_*

$$\begin{aligned}\|x_{t+1} - x_*\|^2 &= \|x_t - s_t \odot d_t - x_*\|^2 \\ &= \|x_t - x_*\|^2 - 2\langle x_t - x_*, s_t \odot d_t\rangle + \|s_t \odot d_t\|^2 \\ &= \|x_t - x_*\|^2 + \sum_{k=1}^{m} \left(-2\gamma_{t,k}\langle x_{t,k} - x_{*,k},\ d_{t,k}\rangle + \gamma_{t,k}^2 \|d_{t,k}\|^2\right)\end{aligned}$$

Expected distance to an optimal point

distance of x_{t+1} to an optimal point x_*

$$\begin{aligned}\|x_{t+1} - x_*\|^2 &= \|x_t - s_t \odot d_t - x_*\|^2 \\ &= \|x_t - x_*\|^2 - 2\langle x_t - x_*, s_t \odot d_t\rangle + \|s_t \odot d_t\|^2 \\ &= \|x_t - x_*\|^2 + \sum_{k=1}^{m} \left(-2\gamma_{t,k}\langle x_{t,k} - x_{*,k},\, d_{t,k}\rangle + \gamma_{t,k}^2 \|d_{t,k}\|^2\right)\end{aligned}$$

conditional expectation

$$\mathbf{E}_t[\cdot] := \mathbf{E}[\cdot | x_0, d_0, x_1, d_1, \ldots, d_{t-1}, x_t]$$

expected distance

$$\mathbf{E}_t\left[\|x_{t+1} - x_*\|^2\right] = \|x_t - x_*\|^2 + \sum_{k=1}^{m} \left(-2\gamma_{t,k}\left\langle x_{t,k} - x_{*,k},\, \mathbf{E}_t[d_{t,k}]\right\rangle + \gamma_{t,k}^2 \mathbf{E}_t\left[\|d_{t,k}\|^2\right]\right)$$

Inspiration from Polyak stepsize

- (sub)gradient method for **convex optimization**: $x_{t+1} = x_t - \gamma_t \nabla f(x_t)$
- distance of x_{t+1} to an optimal point x_*

$$\begin{aligned} \|x_{t+1} - x_*\|^2 &= \|x_t - \gamma_t \nabla f(x_t) - x_*\|^2 \\ &= \|x_t - x_*\|^2 - 2\gamma_t \nabla f(x_t)^T (x_t - x_*) + \gamma_t^2 \|\nabla f(x_t)\|^2 \\ \textbf{(by convexity)} &\leq \|x_t - x_*\|^2 - 2\gamma_k \big(f(x_t) - f^\star\big) + \gamma_t^2 \|\nabla f(x_t)\|^2 \end{aligned}$$

Inspiration from Polyak stepsize

- (sub)gradient method for **convex optimization**: $x_{t+1} = x_t - \gamma_t \nabla f(x_t)$
- distance of x_{t+1} to an optimal point x_*

$$\begin{aligned} \|x_{t+1} - x_*\|^2 &= \|x_t - \gamma_t \nabla f(x_t) - x_*\|^2 \\ &= \|x_t - x_*\|^2 - 2\gamma_t \nabla f(x_t)^T (x_t - x_*) + \gamma_t^2 \|\nabla f(x_t)\|^2 \\ \textbf{(by convexity)} &\leq \|x_t - x_*\|^2 - 2\gamma_k \big(f(x_t) - f^\star\big) + \gamma_t^2 \|\nabla f(x_t)\|^2 \end{aligned}$$

- **choose γ_t to minimize upper bound on $\|x_{t+1} - x_*\|^2$**

$$\widehat{\gamma}_t = \frac{f(x_t) - f^\star}{\|\nabla f(x_t)\|^2}$$

Inspiration from Polyak stepsize

- (sub)gradient method for **convex optimization**: $x_{t+1} = x_t - \gamma_t \nabla f(x_t)$
- distance of x_{t+1} to an optimal point x_*

$$\begin{aligned} \|x_{t+1} - x_*\|^2 &= \|x_t - \gamma_t \nabla f(x_t) - x_*\|^2 \\ &= \|x_t - x_*\|^2 - 2\gamma_t \nabla f(x_t)^T (x_t - x_*) + \gamma_t^2 \|\nabla f(x_t)\|^2 \\ \textbf{(by convexity)} &\leq \|x_t - x_*\|^2 - 2\gamma_k \big(f(x_t) - f^\star\big) + \gamma_t^2 \|\nabla f(x_t)\|^2 \end{aligned}$$

- **choose γ_t to minimize upper bound on $\|x_{t+1} - x_*\|^2$**

$$\widehat{\gamma}_t = \frac{f(x_t) - f^\star}{\|\nabla f(x_t)\|^2}$$

- **limitations**
  - requires convexity (but can derive similar stepsizes using smoothness)
  - do not have access of $f^\star$ in general (and $f(x_t)$ in stochastic setting)

Block coordinate optimal stepsizes (BCOS)

choose $\gamma_{t,k}$ minimize expected distance of x_{t+1} from x_*

$$\mathbf{E}_t\left[\|x_{t+1}-x_*\|^2\right] = \|x_t-x_*\|^2+\sum_{k=1}^{m}\left(-2\gamma_{t,k}\left\langle x_{t,k}-x_{*,k},\, \mathbf{E}_t[d_{t,k}]\right\rangle+\gamma_{t,k}^2\mathbf{E}_t\left[\|d_{t,k}\|^2\right]\right)$$

block-coordinate optimal stepsizes

$$\widehat{\gamma}_{t,k} = \frac{\left\langle x_{t,k} - x_{*,k},\, \mathbf{E}_t[d_{t,k}]\right\rangle}{\mathbf{E}_t[\|d_{t,k}\|^2]}, \qquad k = 1, \ldots, m$$

Block coordinate optimal stepsizes (BCOS)

choose $\gamma_{t,k}$ minimize expected distance of x_{t+1} from x_*

$$\mathbf{E}_t\left[\|x_{t+1}-x_*\|^2\right] = \|x_t-x_*\|^2+\sum_{k=1}^{m}\left(-2\gamma_{t,k}\left\langle x_{t,k}-x_{*,k},\, \mathbf{E}_t[d_{t,k}]\right\rangle+\gamma_{t,k}^2\mathbf{E}_t\left[\|d_{t,k}\|^2\right]\right)$$

block-coordinate optimal stepsizes

$$\widehat{\gamma}_{t,k} = \frac{\left\langle x_{t,k} - x_{*,k},\, \mathbf{E}_t[d_{t,k}]\right\rangle}{\mathbf{E}_t[\|d_{t,k}\|^2]}, \qquad k = 1, \ldots, m$$

obviously, does not work in practice

- do not have access to x_*
- cannot compute $\mathbf{E}_t[\cdot]$ precisely

Simplify optimal stepsize rule

$$\widehat{\gamma}_{t,k} = \frac{\langle x_{t,k} - x_{*,k},\, \mathbf{E}_t[d_{t,k}] \rangle}{\mathbf{E}_t[\|d_{t,k}\|^2]}$$

- absorb quantities related to x_* into tunable parameters $\alpha_{t,k}$

$$\begin{aligned} \left\langle x_{t,k} - x_{*,k},\, \mathbf{E}_t[d_{t,k}] \right\rangle &= \|x_{t,k} - x_{*,k}\| \|\mathbf{E}_t[d_{t,k}]\| \cos\theta_{t,k} \\ &\approx \alpha_{t,k} \|\mathbf{E}_t[d_{t,k}]\| \end{aligned}$$

restrict $\alpha_{t,k} > 0$: being "**optimistic**" that $\langle x_{t,k} - x_{*,k}, \mathbf{E}_t[d_{t,k}] \rangle > 0$

Simplify optimal stepsize rule

$$\widehat{\gamma}_{t,k} = \frac{\langle x_{t,k} - x_{*,k},\, \mathbf{E}_t[d_{t,k}] \rangle}{\mathbf{E}_t[\|d_{t,k}\|^2]}$$

- absorb quantities related to x_* into tunable parameters $\alpha_{t,k}$

$$\begin{aligned} \left\langle x_{t,k} - x_{*,k},\, \mathbf{E}_t[d_{t,k}] \right\rangle &= \|x_{t,k} - x_{*,k}\| \|\mathbf{E}_t[d_{t,k}]\| \cos\theta_{t,k} \\ &\approx \alpha_{t,k} \|\mathbf{E}_t[d_{t,k}]\| \end{aligned}$$

  restrict $\alpha_{t,k} > 0$: being "**optimistic**" that $\langle x_{t,k} - x_{*,k}, \mathbf{E}_t[d_{t,k}] \rangle > 0$

- use a single $\alpha_t > 0$

$$\widetilde{\gamma}_{t,k} = \alpha_t \frac{\|\mathbf{E}_t[d_{t,k}]\|}{\mathbf{E}_t[\|d_{t,k}\|^2]}$$

  - assuming $\alpha_{t,k}$ similar (better to exploit block-wise structure if possible)
  - **stepsize schedule**: α_t should decrease as $\mathbf{E}[\|x_t - x_*\|]$ becomes small

Approximate the expectations

$$\widetilde{\gamma}_{t,k} = \alpha_t \frac{\|\mathbf{E}_t[d_{t,k}]\|}{\mathbf{E}_t[\|d_{t,k}\|^2]}$$

- approximate $\mathbf{E}_t[\cdot]$ using exponential moving average (EMA)

$$u_{t,k} = \beta u_{t-1,k} + (1-\beta) d_{t,k}$$
$$v_{t,k} = \beta v_{t-1,k} + (1-\beta)\|d_{t,k}\|^2$$

- practical block-coordinate stepsize

$$\gamma_{t,k} = \alpha_t \frac{\|u_{t,k}\|}{v_{t,k} + \epsilon}$$

  - $u_{t,k} \in \mathbf{R}^{n_k}$ and $v_{t,k} \in \mathbf{R}_+$: mean and 2nd-moment estimators respectively
  - $\epsilon > 0$ on denominator: avoid numerical instability when $v_{t,k}$ too small

Approximate the expectations

$$\widetilde{\gamma}_{t,k} = \alpha_t \frac{\|\mathbf{E}_t[d_{t,k}]\|}{\mathbf{E}_t[\|d_{t,k}\|^2]}$$

- approximate $\mathbf{E}_t[\cdot]$ using exponential moving average (EMA)

$$u_{t,k} = \beta u_{t-1,k} + (1-\beta) d_{t,k}$$
$$v_{t,k} = \beta v_{t-1,k} + (1-\beta)\|d_{t,k}\|^2$$

- practical block-coordinate stepsize

$$\gamma_{t,k} = \alpha_t \frac{\|u_{t,k}\|}{v_{t,k} + \epsilon}$$

  - $u_{t,k} \in \mathbf{R}^{n_k}$ and $v_{t,k} \in \mathbf{R}_+$: mean and 2nd-moment estimators respectively
  - $\epsilon > 0$ on denominator: avoid numerical instability when $v_{t,k}$ too small

problem: ratio of two EMA estimators can be volatile

Further simplfication

- define **signal fraction** (SiF)

$$\rho_{t,k} = \frac{\|\mathbf{E}_t[d_{t,k}]\|^2}{\mathbf{E}_t[\|d_{t,k}\|^2]} = \frac{\|\mathbf{E}_t[d_{t,k}]\|^2}{\|\mathbf{E}_t[d_{t,k}]\|^2 + \mathsf{Var}(d_{t,k})} \in [0,1]$$

- decomposition of stepsize rule (keeping scaling-invariance)

$$\widetilde{\gamma}_{t,k} = \alpha_t \sqrt{\frac{\|\mathbf{E}_t[d_{t,k}]\|^2}{\mathbf{E}_t[\|d_{t,k}\|^2]}} \frac{1}{\sqrt{\mathbf{E}_t[\|d_{t,k}\|^2]}} = \alpha_t \sqrt{\rho_{t,k}} \frac{1}{\sqrt{\mathbf{E}_t[\|d_{t,k}\|^2]}}$$

Further simplfication

- define **signal fraction** (SiF)

$$\rho_{t,k} = \frac{\|\mathbf{E}_t[d_{t,k}]\|^2}{\mathbf{E}_t[\|d_{t,k}\|^2]} = \frac{\|\mathbf{E}_t[d_{t,k}]\|^2}{\|\mathbf{E}_t[d_{t,k}]\|^2 + \mathrm{Var}(d_{t,k})} \in [0,1]$$

- decomposition of stepsize rule (keeping scaling-invariance)

$$\widetilde{\gamma}_{t,k} = \alpha_t \sqrt{\frac{\|\mathbf{E}_t[d_{t,k}]\|^2}{\mathbf{E}_t[\|d_{t,k}\|^2]}} \frac{1}{\sqrt{\mathbf{E}_t[\|d_{t,k}\|^2]}} = \alpha_t \sqrt{\rho_{t,k}} \frac{1}{\sqrt{\mathbf{E}_t[\|d_{t,k}\|^2]}}$$

- approximate 2nd-moment with EMA

$$\widetilde{\gamma}_{t,k} = \alpha'_t \frac{1}{\sqrt{\mathbf{E}_t[d_{t,k}^2]}} \qquad \Longrightarrow \qquad \gamma_{t,k} = \alpha'_t \frac{1}{\sqrt{v_{t,k}} + \epsilon}$$

  - assimilate effect of $\rho_{t,k}$ into α'_t together with $\|x_{t,k} - x_{*,k}\| \cos(\measuredangle)$
  - only one EMA estimator on denominator (**under square-root**)

Outline

- background and motivation
- **PBCOS: block coordinate optimal stepsizes**
- instantiations of BCOS
  - **search directions:** stochastic gradient, momentum, preconditioned, ...
  - **2nd moment estimators:** EMA, **conditional estimator**
- numerical experiments
- **convergence analysis**

Instantiations of BCOS

focus on **single-coordinate blocks** (element-wise arithmetic)

$$v_t = \beta v_{t-1} + (1-\beta) d_t^2$$

$$x_{t+1} = x_t - \alpha_t \frac{d_t}{\sqrt{v_t} + \epsilon}$$

Instantiations of BCOS

focus on **single-coordinate blocks** (element-wise arithmetic)

$$v_t = \beta v_{t-1} + (1-\beta) d_t^2$$
$$x_{t+1} = x_t - \alpha_t \frac{d_t}{\sqrt{v_t} + \epsilon}$$

- search direction d_t
  - stochastic gradient: $g_t = \nabla f(x_t, \xi_t)$
  - stochastic momentum: $m_t = \beta d_{t-1} + (1-\beta) g_t$
- 2nd-moment estimator v_t
  - EMA estimator: $v_t = \beta' v_{t-1} + (1-\beta') d_t^2$
  - **conditional estimator** when $d_t = m_t$
- BCOS with decoupled weight decay

BCOS with stochastic gradient as search direction

Algorithm BCOS-g

input: x_0, $\{\alpha_t\}_{t\geq 0}$, $\beta \in [0,1)$, $\epsilon > 0$

$v_{-1} = g_0^2$

for $t = 0, 1, 2, \ldots$ **do**

$g_t = \nabla f(x_t, \xi_t)$

$v_t = \beta v_{t-1} + (1-\beta) g_t^2$

$$x_{t+1} = x_t - \alpha_t \frac{g_t}{\sqrt{v_t} + \epsilon}$$

- same as RMSprop [Tieleman and Hinton, 2012]
- special case: $\beta = 0$ and $\epsilon = 0$

$$x_{t+1} = x_t - \alpha_t \text{sign}(g_t)$$

(sign-SGD method)

BCOS with momentum as search direction

Algorithm BCOS-m

input: x_0, $\{\alpha_t\}$, $\beta_1, \beta_2 \in [0, 1)$, $\epsilon > 0$

$m_{-1} = g_0, \quad v_{-1} = g_0^2$

for $t = 0, 1, 2, \ldots$ **do**

$$g_t = \nabla f(x_t, \xi_t)$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) m_t^2$$

$$x_{t+1} = x_t - \alpha_t \frac{m_t}{\sqrt{v_t} + \epsilon}$$

- two smoothing factors β_1, β_2
  - can choose independently
  - $\beta_1 = \beta_2$ works well in practice

BCOS with momentum as search direction

Algorithm BCOS-m

input: x_0, $\{\alpha_t\}$, $\beta_1, \beta_2 \in [0, 1)$, $\epsilon > 0$

$m_{-1} = g_0$, $v_{-1} = g_0^2$

for $t = 0, 1, 2, \ldots$ **do**

$$g_t = \nabla f(x_t, \xi_t)$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) m_t^2$$

$$x_{t+1} = x_t - \alpha_t \frac{m_t}{\sqrt{v_t} + \epsilon}$$

- two smoothing factors β_1, β_2
  - can choose independently
  - $\beta_1 = \beta_2$ works well in practice
- **Adam: replace m_t^2 by g_t^2**
  - **mismatch** between direction and 2nd moment estimator
  - **need to compensate** with larger β_2 (e.g., $\beta_1 = 0.9$, $\beta_2 = 0.99$)

BCOS with conditional estimator

- notice **conditional expectation** in $\widetilde{\gamma}_{t,k} = \dfrac{\alpha_t}{\sqrt{\mathbf{E}_t[d_t^2]}}$

BCOS with conditional estimator

- notice **conditional expectation** in $\widetilde{\gamma}_{t,k} = \dfrac{\alpha_t}{\sqrt{\mathbf{E}_t[d_t^2]}}$
- exploit structure when $d_t = m_t$:

$$\begin{aligned}
\mathbf{E}_t\left[m_t^2\right] &= \mathbf{E}_t\left[(\beta m_{t-1} + (1-\beta) g_t)^2\right] \\
&= \beta^2 \mathbf{E}_t[m_{t-1}^2] + 2\beta(1-\beta)\mathbf{E}_t[m_{t-1} \odot g_t] + (1-\beta)^2 \mathbf{E}_t[g_t^2] \\
&= \beta^2 m_{t-1}^2 + 2\beta(1-\beta) m_{t-1} \odot \mathbf{E}_t[g_t] + (1-\beta)^2 \mathbf{E}_t[g_t^2]
\end{aligned}$$

BCOS with conditional estimator

- notice **conditional expectation** in $\widetilde{\gamma}_{t,k} = \dfrac{\alpha_t}{\sqrt{\mathbf{E}_t[d_t^2]}}$
- exploit structure when $d_t = m_t$:

$$\begin{aligned}\mathbf{E}_t\left[m_t^2\right] &= \mathbf{E}_t\left[(\beta m_{t-1} + (1-\beta) g_t)^2\right] \\ &= \beta^2 \mathbf{E}_t[m_{t-1}^2] + 2\beta(1-\beta)\mathbf{E}_t[m_{t-1} \odot g_t] + (1-\beta)^2 \mathbf{E}_t[g_t^2] \\ &= \beta^2 m_{t-1}^2 + 2\beta(1-\beta) m_{t-1} \odot \mathbf{E}_t[g_t] + (1-\beta)^2 \mathbf{E}_t[g_t^2]\end{aligned}$$

  - approximate $\mathbf{E}_t[g_t]$ with m_t
  - approximate $\mathbf{E}_t[g_t^2]$ with g_t^2 (coefficient $(1-\beta)^2$ very small in practice)

BCOS with conditional estimator

- notice **conditional expectation** in $\widetilde{\gamma}_{t,k} = \dfrac{\alpha_t}{\sqrt{\mathbf{E}_t[d_t^2]}}$
- exploit structure when $d_t = m_t$:

$$\begin{aligned}\mathbf{E}_t\left[m_t^2\right] &= \mathbf{E}_t\left[(\beta m_{t-1} + (1-\beta) g_t)^2\right] \\ &= \beta^2 \mathbf{E}_t[m_{t-1}^2] + 2\beta(1-\beta)\mathbf{E}_t[m_{t-1} \odot g_t] + (1-\beta)^2 \mathbf{E}_t[g_t^2] \\ &= \beta^2 m_{t-1}^2 + 2\beta(1-\beta) m_{t-1} \odot \mathbf{E}_t[g_t] + (1-\beta)^2 \mathbf{E}_t[g_t^2]\end{aligned}$$

  - approximate $\mathbf{E}_t[g_t]$ with m_t
  - approximate $\mathbf{E}_t[g_t^2]$ with g_t^2 (coefficient $(1-\beta)^2$ very small in practice)

- resulting 2nd-moment estimator

$$v_t = \beta^2 m_{t-1}^2 + 2\beta(1-\beta) m_{t-1} \odot m_t + (1-\beta)^2 g_t^2$$

BCOS with conditional estimator

Algorithm BCOS-c (ugly version)

input: x_0, $\{\alpha_t\}$, $\beta \in [0, 1)$, $\epsilon > 0$

$m_{-1} = g_0, \quad v_{-1} = g_0^2$

for $t = 0, 1, 2, \ldots$ **do**

$g_t = \nabla f(x_t, \xi_t)$

$m_t = \beta m_{t-1} + (1 - \beta) g_t$

$v_t = \beta^2 m_{t-1}^2 + 2\beta(1-\beta) m_{t-1} \odot m_t + (1-\beta)^2 g_t^2$

$x_{t+1} = x_t - \alpha_t \dfrac{m_t}{\sqrt{v_t} + \epsilon}$

- works well in practice!
- fewer optimizer state (do not store v_{t-1})
- fewer hyper-parameters (one smoothing factor)

BCOS with conditional estimator

Algorithm BCOS-c (ugly version)

input: x_0, $\{\alpha_t\}$, $\beta \in [0, 1)$, $\epsilon > 0$

$m_{-1} = g_0, \quad v_{-1} = g_0^2$

for $t = 0, 1, 2, \ldots$ **do**

$g_t = \nabla f(x_t, \xi_t)$

$m_t = \beta m_{t-1} + (1 - \beta) g_t$

$v_t = \beta^2 m_{t-1}^2 + 2\beta(1-\beta) m_{t-1} \odot m_t + (1-\beta)^2 g_t^2$

$x_{t+1} = x_t - \alpha_t \dfrac{m_t}{\sqrt{v_t} + \epsilon}$

- works well in practice!
- fewer optimizer state (do not store v_{t-1})
- fewer hyper-parameters (one smoothing factor)
- **but it looks ugly!**

BCOS with conditional estimator

- notice **conditional expectation** in $\widetilde{\gamma}_{t,k} = \dfrac{\alpha_t}{\sqrt{\mathbf{E}_t[d_t^2]}}$
- exploit structure when $d_t = m_t$:

$$\begin{aligned}\mathbf{E}_t\left[m_t^2\right] &= \mathbf{E}_t\left[(\beta m_{t-1} + (1-\beta)g_t)^2\right] \\ &= \beta^2 \mathbf{E}_t[m_{t-1}^2] + 2\beta(1-\beta)\mathbf{E}_t[m_{t-1} \odot g_t] + (1-\beta)^2\mathbf{E}_t[g_t^2] \\ &= \beta^2 m_{t-1}^2 + 2\beta(1-\beta)m_{t-1} \odot \mathbf{E}_t[g_t] + (1-\beta)^2\mathbf{E}_t[g_t^2]\end{aligned}$$

  - approximate $\mathbf{E}_t[g_t]$ with ~~m_t~~ $\rightarrow m_{t-1}$
  - approximate $\mathbf{E}_t[g_t^2]$ with g_t^2 (coefficient $(1-\beta)^2$ very small in practice)

BCOS with conditional estimator

- notice **conditional expectation** in $\widetilde{\gamma}_{t,k} = \dfrac{\alpha_t}{\sqrt{\mathbf{E}_t[d_t^2]}}$
- exploit structure when $d_t = m_t$:

$$\begin{aligned}\mathbf{E}_t\left[m_t^2\right] &= \mathbf{E}_t\left[(\beta m_{t-1} + (1-\beta) g_t)^2\right] \\ &= \beta^2 \mathbf{E}_t[m_{t-1}^2] + 2\beta(1-\beta)\mathbf{E}_t[m_{t-1} \odot g_t] + (1-\beta)^2 \mathbf{E}_t[g_t^2] \\ &= \beta^2 m_{t-1}^2 + 2\beta(1-\beta) m_{t-1} \odot \mathbf{E}_t[g_t] + (1-\beta)^2 \mathbf{E}_t[g_t^2]\end{aligned}$$

  - approximate $\mathbf{E}_t[g_t]$ with ~~m_t~~ $\rightarrow m_{t-1}$
  - approximate $\mathbf{E}_t[g_t^2]$ with g_t^2 (coefficient $(1-\beta)^2$ very small in practice)

- resulting 2nd-moment estimator

$$v_t = \left(1 - (1-\beta)^2\right) m_{t-1}^2 + (1-\beta)^2 g_t^2$$

BCOS with conditional estimator

- notice **conditional expectation** in $\widetilde{\gamma}_{t,k} = \dfrac{\alpha_t}{\sqrt{\mathbf{E}_t[d_t^2]}}$
- exploit structure when $d_t = m_t$:

$$\begin{aligned}\mathbf{E}_t\left[m_t^2\right] &= \mathbf{E}_t\left[(\beta m_{t-1} + (1-\beta) g_t)^2\right] \\ &= \beta^2 \mathbf{E}_t[m_{t-1}^2] + 2\beta(1-\beta)\mathbf{E}_t[m_{t-1} \odot g_t] + (1-\beta)^2 \mathbf{E}_t[g_t^2] \\ &= \beta^2 m_{t-1}^2 + 2\beta(1-\beta) m_{t-1} \odot \mathbf{E}_t[g_t] + (1-\beta)^2 \mathbf{E}_t[g_t^2]\end{aligned}$$

  - approximate $\mathbf{E}_t[g_t]$ with ~~m_t~~ $\rightarrow m_{t-1}$
  - approximate $\mathbf{E}_t[g_t^2]$ with g_t^2 (coefficient $(1-\beta)^2$ very small in practice)

- resulting 2nd-moment estimator

$$v_t = \left(1 - (1-\beta)^2\right) m_{t-1}^2 + (1-\beta)^2 g_t^2$$

memoryless "EMA" form: $v_t = \beta' m_{t-1}^2 + (1-\beta') g_t^2$ with $\beta' = 1 - (1-\beta)^2$

BCOS with conditional estimator

Algorithm BCOS-c

input: x_0, $\{\alpha_t\}$, $\beta \in [0, 1)$, $\epsilon > 0$

$m_{-1} = g_0, \quad v_{-1} = g_0^2$

for $t = 0, 1, 2, \ldots$ **do**

$$g_t = \nabla f(x_t, \xi_t)$$

$$m_t = \beta m_{t-1} + (1 - \beta) g_t$$

$$v_t = \left(1 - (1 - \beta)^2\right) m_{t-1}^2 + (1 - \beta)^2 g_t^2$$

$$x_{t+1} = x_t - \alpha_t \frac{m_t}{\sqrt{v_t} + \epsilon}$$

- fewer optimizer state (do not store v_{t-1})
- fewer hyper-parameters (one smoothing factor)
- and it looks **beautiful**

BCOS with conditional estimator

Algorithm BCOS-c

input: x_0, $\{\alpha_t\}$, $\beta \in [0, 1)$, $\epsilon > 0$

$m_{-1} = g_0, \quad v_{-1} = g_0^2$

for $t = 0, 1, 2, \ldots$ **do**

$g_t = \nabla f(x_t, \xi_t)$

$m_t = \beta m_{t-1} + (1 - \beta) g_t$

$v_t = \big(1 - (1 - \beta)^2\big) m_{t-1}^2 + (1 - \beta)^2 g_t^2$

$x_{t+1} = x_t - \alpha_t \dfrac{m_t}{\sqrt{v_t} + \epsilon}$

- fewer optimizer state (do not store v_{t-1})
- fewer hyper-parameters (one smoothing factor)
- and it looks **beautiful**

compare with Adam

- replace v_{t-1} with m_{t-1}^2
- explains $\beta_2 \approx 1 - (1 - \beta_1)^2$

BCOS with decoupled weight decay

- minimize regularized $\mathbf{E}_\xi[f(x,\xi)] + \frac{\lambda}{2}\|x\|^2$
- can apply BCOS with $g_t := \nabla f(x_t, \xi_t) + \lambda x_t$, but does not work well

BCOS with decoupled weight decay

- minimize regularized $\mathbf{E}_\xi[f(x,\xi)] + \frac{\lambda}{2}\|x\|^2$
- can apply BCOS with $g_t := \nabla f(x_t, \xi_t) + \lambda x_t$, but does not work well
- AdamW: Adam with decoupled weight decay [Loshchilov and Hutter, 2019]

$$\begin{aligned} x_{t+1} &= x_t - \alpha_t \frac{m_t}{\sqrt{v_t} + \epsilon} - \alpha_t \lambda x_t \\ &= (1 - \alpha_t \lambda) x_t - \alpha_t \frac{m_t}{\sqrt{v_t} + \epsilon} \end{aligned}$$

where m_t is momentum of $g_t = \nabla f(x_t, \xi_t)$, not including λx_t

BCOS with decoupled weight decay

- minimize regularized $\mathbf{E}_\xi[f(x,\xi)] + \frac{\lambda}{2}\|x\|^2$
- can apply BCOS with $g_t := \nabla f(x_t, \xi_t) + \lambda x_t$, but does not work well
- AdamW: Adam with decoupled weight decay [Loshchilov and Hutter, 2019]

$$\begin{aligned} x_{t+1} &= x_t - \alpha_t \frac{m_t}{\sqrt{v_t} + \epsilon} - \alpha_t \lambda x_t \\ &= (1 - \alpha_t \lambda) x_t - \alpha_t \frac{m_t}{\sqrt{v_t} + \epsilon} \end{aligned}$$

  where m_t is momentum of $g_t = \nabla f(x_t, \xi_t)$, not including λx_t

- **BCOSW: apply the same change to BCOS**

BCOSW with conditional estimator

Algorithm BCOSW-c

input: x_0, $\{\alpha_t\}$, $\beta_1, \beta_2 \in [0, 1)$, $\epsilon > 0$

$m_{-1} = g_0, \quad v_{-1} = g_0^2$

for $t = 0, 1, 2, \ldots$ **do**

$$g_t = \nabla f(x_t, \xi_t)$$

$$m_t = \beta_1 m_{t-1} + (1 - \beta_1) g_t$$

$$v_t = \left(1 - (1 - \beta)^2\right) m_{t-1}^2 + (1 - \beta)^2 g_t^2$$

$$x_{t+1} = (1 - \alpha_t \lambda) x_t - \alpha_t \frac{m_t}{\sqrt{v_t} + \epsilon}$$

- decoupled weight decay
- apply similar change for BCOSW-g, BCOSW-m
- BCOSW-c perform best in experiments

Outline

- background and motivation
- **BCOS: block coordinate optimal stepsizes**
- instantiations of BCOS
  - **search directions:** stochastic gradient, momentum, preconditioned, ...
  - **2nd moment estimators:** EMA, **conditional estimator**
- **numerical experiments**
- **convergence analysis**

Experiments: varying smoothing factors

training GPT2 (126M) on OpenWebText dataset

- $\alpha = 0.002$, warm up 2k iterations, then cosine decay ($\times 0.01$)
- weight decay $\lambda = 0.1$

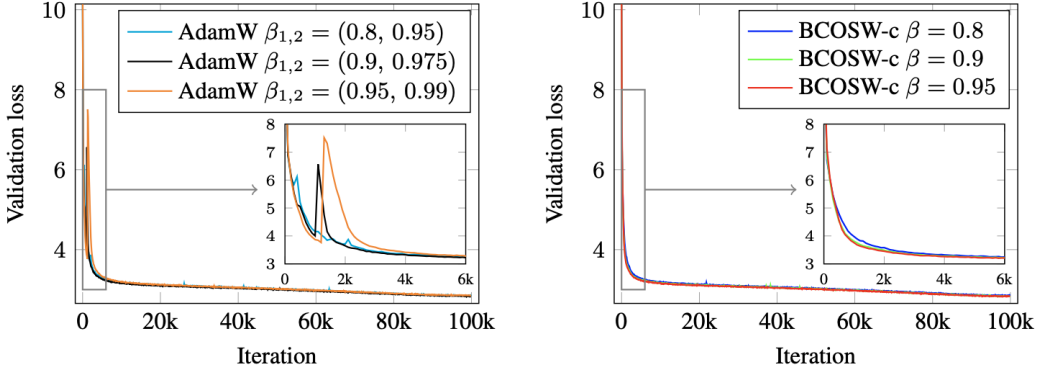


Figure 1: Comparing AdamW and BCOSW-c with different momentum parameters.

Experiments: comparing -g, -m, -c variants

training GPT2 (126M) on OpenWebText dataset

- $\alpha = 0.002$, warm up 2k iterations, then cosine decay ($\times 0.01$)
- weight decay $\lambda = 0.1$

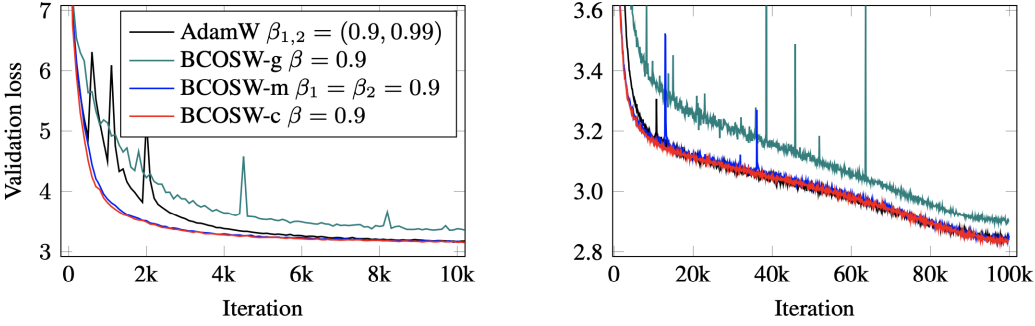
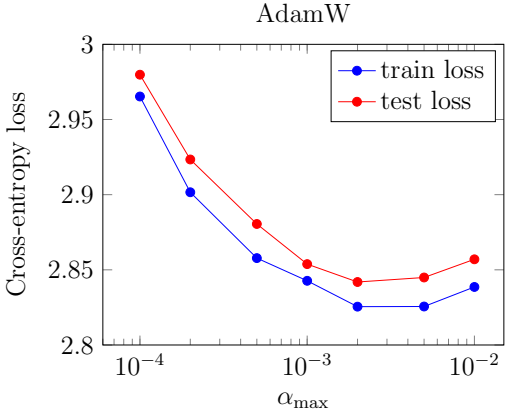


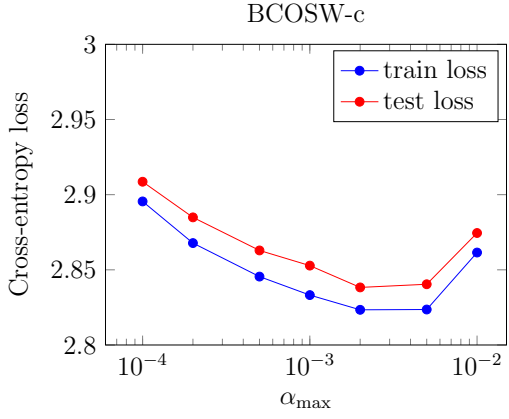
Figure 2: Comparing AdamW and BCOSW. Left: first 10k iterations; Right: all 100k iterations.

Experiments: loss vs lr

training GPT2 (126M) on OpenWebText dataset

- warm up 2k iterations, then cosine decay ($\times 0.01$)
- weight decay $\lambda = 0.1$





Experiments: decoupled weight decay

training GPT2 (126M) on OpenWebText dataset

- $\alpha = 0.002$, warm up 2k iterations, then cosine decay ($\times 0.01$)
- weight decay $\lambda = 0.1$

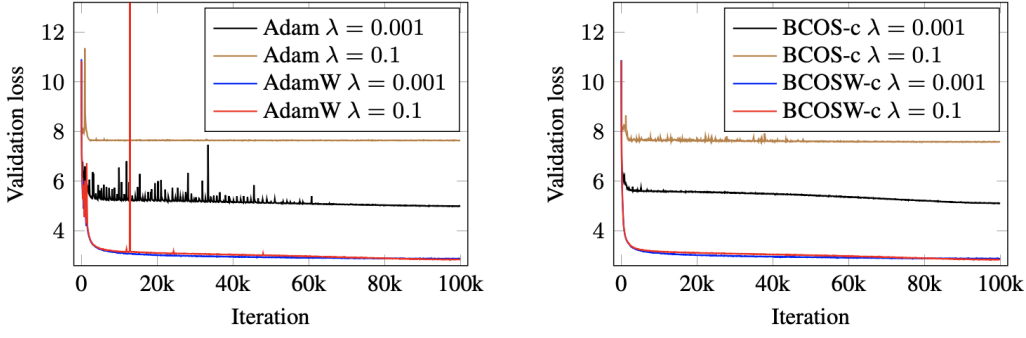


Figure 3: Left: Adam/ AdamW with $\beta_{1,2} = (0.9, 0.99)$. Right: BCOS/BCOSW with $\beta = 0.9$.

Experiments: ResNet and Vision Transformer

- ResNet20 on CIFAR10 (using both cosine decay and step decay)
- vision transformer (ViT) on ImageNet (cosine decay)

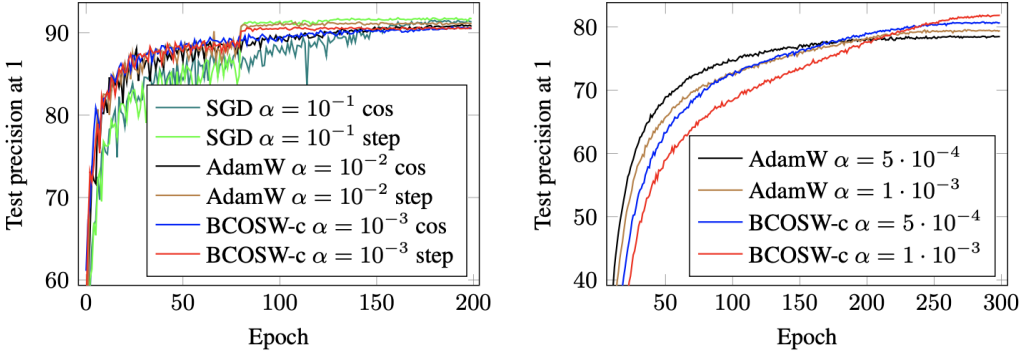


Figure 4: Left: ResNet-20 on CIFAR10. Right: Vision Transformer on ImageNet.

Outline

- background and motivation
- P**BCOS: block coordinate optimal stepsizes**
- instantiations of BCOS
  - **search directions:** stochastic gradient, momentum, preconditioned, ...
  - **2nd moment estimators:** EMA, **conditional estimator**
- **numerical experiments**
- **convergence analysis**

Convergence analysis in two steps

- analysis of **conceptual algorithm**

$$x_{t+1} = (1 - \alpha_t \lambda) x_t - \widetilde{\gamma}_t \odot d_t, \qquad \text{where} \qquad \widetilde{\gamma}_t = \alpha_t \frac{1}{\sqrt{\mathbf{E}_t[d_t^2]}}$$

following classical SA literature in 1950's

Convergence analysis in two steps

- analysis of **conceptual algorithm**

$$x_{t+1} = (1 - \alpha_t \lambda) x_t - \widetilde{\gamma}_t \odot d_t, \qquad \text{where} \qquad \widetilde{\gamma}_t = \alpha_t \frac{1}{\sqrt{\mathbf{E}_t[d_t^2]}}$$

  following classical SA literature in 1950's

- analysis of **practical algorithm**

$$x_{t+1} = (1 - \alpha_t \lambda) x_t - \gamma_t \odot d_t, \qquad \text{where} \qquad \gamma_t = \alpha_t \frac{1}{\sqrt{v_t + \epsilon}}$$

Convergence analysis in two steps

- analysis of **conceptual algorithm**

$$x_{t+1} = (1 - \alpha_t \lambda) x_t - \widetilde{\gamma}_t \odot d_t, \qquad \text{where} \qquad \widetilde{\gamma}_t = \alpha_t \frac{1}{\sqrt{\mathbf{E}_t[d_t^2]}}$$

following classical SA literature in 1950's

- analysis of **practical algorithm**

$$x_{t+1} = (1 - \alpha_t \lambda) x_t - \gamma_t \odot d_t, \qquad \text{where} \qquad \gamma_t = \alpha_t \frac{1}{\sqrt{v_t + \epsilon}}$$

based on bounding the difference between expected updates

$$\left| \mathbf{E}_t[\widetilde{\gamma}_t \odot d_t] - \mathbf{E}_t[\gamma_t \odot d_t] \right| \leq c_t \left| \mathbf{E}_t[\widetilde{\gamma}_t \odot d_t] \right| + \mathcal{O}(\epsilon) + \mathcal{O}(\mathrm{Var}_t(v_t))$$

Classical aiming conditions

- classical stochastic approximation (SA) for solving $\mathbf{E}[g(x_*, \xi)] = 0$

$$x_{t+1} = x_t - \alpha_t \, g(x_t, \xi_t)$$

aiming conditions critical in convergence analysis
  - original SA paper [Robbins and Monro, 1951]
  - multi-dimensional extension [Blum, 1954]
  - many others ...

- simplified presentation (strong aiming condition) [Sakrison, 1966]

$$\mu \|x - x_*\|^2 \leq \big\langle x - x_*, \mathbf{E}[g(x, \xi)] \big\rangle \leq \kappa \|x - x_*\|^2$$

Classical aiming conditions

- classical stochastic approximation (SA) for solving $\mathbf{E}[g(x_*, \xi)] = 0$

$$x_{t+1} = x_t - \alpha_t \, g(x_t, \xi_t)$$

aiming conditions critical in convergence analysis
  - original SA paper [Robbins and Monro, 1951]
  - multi-dimensional extension [Blum, 1954]
  - many others ...

- simplified presentation (strong aiming condition) [Sakrison, 1966]

$$\mu \|x - x_*\|^2 \leq \big\langle x - x_*, \mathbf{E}[g(x, \xi)] \big\rangle \leq \kappa \|x - x_*\|^2$$

- stochastic optimization: minimize $\mathbf{E}[f(x, \xi)]$ and $g(x, \xi) = \nabla f(x, \xi)$

$$\text{(strong) convexity} + \text{smoothness} \quad \Longrightarrow \quad \text{(strong) aiming}$$

Aiming condition with coordinate-wise stepsizes

- expected update for **conceptual BCOS**

$$\mathbf{E}_t[\widetilde{\gamma}_t \odot d_t] = \mathbf{E}_t\left[\alpha_t \frac{d_t}{\sqrt{\mathbf{E}_t[d_t^2]}}\right] = \alpha_t \frac{\mathbf{E}_t[d_t]}{\sqrt{\mathbf{E}_t[d_t^2]}} = \alpha_t \sqrt{\rho_t} \odot \mathrm{sign}(\mathbf{E}_t[d_t])$$

Aiming condition with coordinate-wise stepsizes

- expected update for **conceptual BCOS**

$$\mathbf{E}_t[\widetilde{\gamma}_t \odot d_t] = \mathbf{E}_t\left[\alpha_t \frac{d_t}{\sqrt{\mathbf{E}_t[d_t^2]}}\right] = \alpha_t \frac{\mathbf{E}_t[d_t]}{\sqrt{\mathbf{E}_t[d_t^2]}} = \alpha_t \sqrt{\rho_t} \odot \operatorname{sign}(\mathbf{E}_t[d_t])$$

- **Assumption A (Aiming):**
  - there exists $x_* \in \mathbf{R}^n$ such that the following holds for all $t \geq 0$ almost surely

$$\left\langle x_t - x_*,\ \sqrt{\rho_t} \odot \operatorname{sign}(\mathbf{E}_t[d_t]) + \lambda x_t \right\rangle \geq \lambda \|x_t - x_*\|^2$$

  - if $\mathbf{E}_t[d_t] = \mathbf{E}[d_t|x_t]$, then it suffices to have it for every $x \in \mathbf{R}^n$

Aiming condition with coordinate-wise stepsizes

- expected update for **conceptual BCOS**

$$\mathbf{E}_t[\widetilde{\gamma}_t \odot d_t] = \mathbf{E}_t\left[\alpha_t \frac{d_t}{\sqrt{\mathbf{E}_t[d_t^2]}}\right] = \alpha_t \frac{\mathbf{E}_t[d_t]}{\sqrt{\mathbf{E}_t[d_t^2]}} = \alpha_t \sqrt{\rho_t} \odot \operatorname{sign}(\mathbf{E}_t[d_t])$$

- **Assumption A (Aiming):**
  - there exists $x_* \in \mathbf{R}^n$ such that the following holds for all $t \geq 0$ almost surely

$$\left\langle x_t - x_*,\ \sqrt{\rho_t} \odot \operatorname{sign}(\mathbf{E}_t[d_t]) + \lambda x_t \right\rangle \geq \lambda \|x_t - x_*\|^2$$

  - if $\mathbf{E}_t[d_t] = \mathbf{E}[d_t|x_t]$, then it suffices to have it for every $x \in \mathbf{R}^n$
- compare with classical aiming: one-sided bound only (due to $\rho_{t,k} \leq 1$)
- assumes neither convexity nor smoothness, but overlapping with convexity

Understanding the aiming condition

- **block-coordinate weighted aiming** (simplified version with $\lambda = 0$)

$$\left\langle x_t - x_*,\ \sqrt{\rho_t} \odot \operatorname{sign}(\mathbf{E}_t[d_t]) \right\rangle \geq 0$$

  - sign of expected search direction weighted by SiF: $\rho_{t,k} = \dfrac{\|\mathbf{E}_t[d_{t,k}]\|^2}{\mathbf{E}_t[\|d_{t,k}\|^2]}$
  - less contribution from noisy directions

Understanding the aiming condition

- **block-coordinate weighted aiming** (simplified version with $\lambda = 0$)

$$\left\langle x_t - x_*,\ \sqrt{\rho_t} \odot \operatorname{sign}(\mathbf{E}_t[d_t]) \right\rangle \geq 0$$

  - sign of expected search direction weighted by SiF: $\rho_{t,k} = \dfrac{\|\mathbf{E}_t[d_{t,k}]\|^2}{\mathbf{E}_t[\|d_{t,k}\|^2]}$
  - less contribution from noisy directions

- **aiming condition versus convexity**
  - full-block version recovers classical condition (which follows from convexity)
  - can tolerate coordinate-wise nonconvexity
  - **cannot handle non-diagonal ill-conditioning** (not surprising)

Almost-sure convergence of conceptual BCOSW

- **lemma:** under **Assumption A**, $0 \leq \alpha_t \lambda < 1$, and $c_* = n + \lambda^2 \|x_*\|^2 + 2\lambda \|x_*\|_1$,

$$\mathbf{E}_t\left[\|x_{t+1} - x_*\|^2\right] \leq (1 - \alpha_t \lambda)^2 \|x_t - x_*\|^2 + \alpha_t^2 c_*,$$

  thus for sufficiently small α_t,

$$\mathbf{E}_t\left[\|x_{t+1} - x_*\|^2\right] \leq \|x_t - x_*\|^2$$

Almost-sure convergence of conceptual BCOSW

- **lemma:** under **Assumption A**, $0 \leq \alpha_t \lambda < 1$, and $c_* = n + \lambda^2 \|x_*\|^2 + 2\lambda \|x_*\|_1$,

$$\mathbf{E}_t\left[\|x_{t+1} - x_*\|^2\right] \leq (1 - \alpha_t \lambda)^2 \|x_t - x_*\|^2 + \alpha_t^2 c_*,$$

  thus for sufficiently small α_t,

$$\mathbf{E}_t\left[\|x_{t+1} - x_*\|^2\right] \leq \|x_t - x_*\|^2$$

- **theorem:** suppose $\alpha_t \geq 0$ and $0 \leq \alpha_t \lambda \leq 1$ for all $t \geq 0$ and

$$\sum_{t=0}^{\infty} \alpha_t = \infty, \qquad \sum_{t=0}^{\infty} \alpha_t^2 < \infty,$$

  then **Assumption A** implies $\|x_t - x_*\| \to 0$ **almost surely**

  (by "almost supermargingale" lemma of [Robbins and Siegmund, 1971])

Convergence rates of conceptual BCOSW

- **corollary:** suppose $\alpha_t = \alpha > 0$ and $\alpha\lambda < 1$ then **Assumption A** implies

$$\mathbf{E}\left[\|x_t - x_*\|^2\right] \leq (1-\alpha\lambda)^{2t}\, \mathbf{E}\left[\|x_0 - x_*\|^2\right] + \frac{\alpha^2 c_*}{1-(1-\alpha\lambda)^2}$$

Convergence rates of conceptual BCOSW

- **corollary:** suppose $\alpha_t = \alpha > 0$ and $\alpha\lambda < 1$ then **Assumption A** implies

$$\mathbf{E}\left[\|x_t - x_*\|^2\right] \leq (1-\alpha\lambda)^{2t}\, \mathbf{E}\left[\|x_0 - x_*\|^2\right] + \frac{\alpha^2 c_*}{1-(1-\alpha\lambda)^2}$$

- **theorem:** suppose $\alpha_t = \alpha/(t+1)$ with $1/2 < \alpha\lambda < 1$, then **Assumption A** implies for all $t \geq 1$

$$\mathbf{E}\left[\|x_t - x_*\|^2\right] \leq \frac{\alpha^2\left(c_*' + \lambda^2 \mathbf{E}[\|x_0 - x_*\|^2]\right)}{2\alpha\lambda - 1}\frac{1}{t} + \mathcal{O}\left(\frac{1}{t^{2\alpha\lambda}}\right)$$

(by applying Chung's lemma [Chung, 1954])

Analysis of practical BCOSW

- **Assumption B (Bias):** there exist $\tau > 0$ and $\epsilon > 0$ such that for all $t \geq 0$

$$\left|\mathbf{E}_t[v_t] - \mathbf{E}_t[d_t^2]\right| \leq \tau \mathbf{E}_t[d_t^2] + \epsilon$$

Analysis of practical BCOSW

- **Assumption B (Bias):** there exist $\tau > 0$ and $\epsilon > 0$ such that for all $t \geq 0$

$$\left|\mathbf{E}_t[v_t] - \mathbf{E}_t[d_t^2]\right| \leq \tau \mathbf{E}_t[d_t^2] + \epsilon$$

- **lemma: Assumptions B** implies

$$\left|\frac{\mathbf{E}_t[d_t]}{\sqrt{\mathbf{E}_t[d_t^2]}} - \mathbf{E}_t\left[\frac{d_t}{\sqrt{v_t+\epsilon}}\right]\right| \leq c_t \left|\frac{\mathbf{E}_t[d_t]}{\sqrt{\mathbf{E}_t[d_t^2]}}\right| + \mathcal{O}(\epsilon) + \mathcal{O}(\mathrm{Var}_t(v_t)),$$

where

$$c_t = \frac{4\tau + 3\tau^2}{8} + \frac{8 + 4\tau + 3\tau^2}{16}\left(\frac{1}{\mathrm{SNR}_t(v_t+\epsilon)} + \frac{1}{\sqrt{\mathrm{SNR}_t(d_t)}\sqrt{\mathrm{SNR}_t(v_t+\epsilon)}}\right)$$

- $\mathrm{SNR}_t(\cdot)$ denotes **conditional Signal-to-Noise Ratio**, e.g.,

$$\mathrm{SNR}_t(d_t) = \frac{\mathbf{E}_t[d_t]^2}{\mathrm{Var}_t(d_t)} = \frac{\rho_t}{1-\rho_t}$$

Almost sure convergence of practical BCOSW

- **theorem** (almoost sure convergence to a neighborhood of x_*)**:** suppose
  - Assumptions **A (Aiming)** and **B (Bias)** hold
  - $\|d_t\|$ bounded almost surely
  - $0 \leq \alpha_t \lambda \leq 1$ for all $t \geq 0$ and $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$

  let δ be the smallest constant such that, for all $t \geq 0$,

$$2c_t \|\sqrt{\rho_t}\| + \mathcal{O}(\epsilon) + \mathcal{O}(\mathrm{Var}_t(v_t)) \leq \lambda\delta,$$

  then

$$\limsup_{t \to \infty} \|x_t - x_*\| \leq \delta \qquad \text{a.s.}$$

  (by Dvoretzky's theorem and extensions [Dvoretzky, 1956] [Venter, 1966])

Almost sure convergence of practical BCOSW

- **theorem** (almoost sure convergence to a neighborhood of x_*)**:** suppose
  - Assumptions **A (Aiming)** and **B (Bias)** hold
  - $\|d_t\|$ bounded almost surely
  - $0 \leq \alpha_t \lambda \leq 1$ for all $t \geq 0$ and $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$

  let δ be the smallest constant such that, for all $t \geq 0$,

  $$2c_t \|\sqrt{\rho_t}\| + \mathcal{O}(\epsilon) + \mathcal{O}(\mathrm{Var}_t(v_t)) \leq \lambda\delta,$$

  then

  $$\limsup_{t \to \infty} \|x_t - x_*\| \leq \delta \qquad \text{a.s.}$$

  (by Dvoretzky's theorem and extensions [Dvoretzky, 1956] [Venter, 1966])

- definition of c_t reflect **bias-variance tradeoff** (simplified for $\tau < 1$)

$$c_t < \tau + \frac{1}{\mathrm{SNR}_t(v_t + \epsilon)} + \frac{1}{\sqrt{\mathrm{SNR}_t(d_t)}\sqrt{\mathrm{SNR}_t(v_t + \epsilon)}}$$

Bias-variance tradeoff

a general framework for convergence analysis

Bias-variance tradeoff

a general framework for convergence analysis

- **SGD** with $d_t = g_t$ or $d_t = m_t$ (using common constant v_t for all coordinates)
  - high bias: $|\mathbf{E}_t[v_t] - \mathbf{E}_t[d_t^2]| = |v - \mathbf{E}_t[d_t^2]|$ for some constant v
  - zero variance

Bias-variance tradeoff

a general framework for convergence analysis

- **SGD** with $d_t = g_t$ or $d_t = m_t$ (using common constant v_t for all coordinates)
  - high bias: $|\mathbf{E}_t[v_t] - \mathbf{E}_t[d_t^2]| = |v - \mathbf{E}_t[d_t^2]|$ for some constant v
  - zero variance
- **sign-SGD** with $d_t = g_t$ or $d_t = m_t$ (effectively using $v_t = d_t^2$)
  - zero bias: $\mathbf{E}_t[v_t] = \mathbf{E}[d_t^2]$
  - high variance: $\mathrm{Var}_t(v_t) = \mathrm{Var}_t(d_t)$

Bias-variance tradeoff

a general framework for convergence analysis

- **SGD** with $d_t = g_t$ or $d_t = m_t$ (using common constant v_t for all coordinates)
  - high bias: $|\mathbf{E}_t[v_t] - \mathbf{E}_t[d_t^2]| = |v - \mathbf{E}_t[d_t^2]|$ for some constant v
  - zero variance
- **sign-SGD** with $d_t = g_t$ or $d_t = m_t$ (effectively using $v_t = d_t^2$)
  - zero bias: $\mathbf{E}_t[v_t] = \mathbf{E}[d_t^2]$
  - high variance: $\mathrm{Var}_t(v_t) = \mathrm{Var}_t(d_t)$
- **Adam(W)**
  - bias: no closely form (better bound by assuming smoothness)
  - variance: $\mathrm{Var}_t(v_t) = (1-\beta_2)^2\mathrm{Var}_t(g_t^2)$

Bias-variance tradeoff

a general framework for convergence analysis

- **SGD** with $d_t = g_t$ or $d_t = m_t$ (using common constant v_t for all coordinates)
  - high bias: $|\mathbf{E}_t[v_t] - \mathbf{E}_t[d_t^2]| = |v - \mathbf{E}_t[d_t^2]|$ for some constant v
  - zero variance
- **sign-SGD** with $d_t = g_t$ or $d_t = m_t$ (effectively using $v_t = d_t^2$)
  - zero bias: $\mathbf{E}_t[v_t] = \mathbf{E}[d_t^2]$
  - high variance: $\mathrm{Var}_t(v_t) = \mathrm{Var}_t(d_t)$
- **Adam(W)**
  - bias: no closely form (better bound by assuming smoothness)
  - variance: $\mathrm{Var}_t(v_t) = (1-\beta_2)^2\mathrm{Var}_t(g_t^2)$
- **BCOS(W)-c**
  - bias: $\mathbf{E}_t[v_t] - \mathbf{E}_t[d_t^2] = 2\beta(1-\beta)m_{t-1}\big(m_{t-1} - \mathbf{E}_t[g_t]\big)$
  - variance: $\mathrm{Var}_t(v_t) = (1-\beta)^4\mathrm{Var}_t(g_t^2)$ (same as Adam $\beta_2 = 1-(1-\beta)^2$)

Summary

BCOS: a family of block-coordinate optimal stepsizes

- **basic idea:** minimize expected distance of next iterate to an optimal point
- instantiations of BCOS
  - **search directions:** gradient, momentum, preconditioned, spectral, ...
  - **2nd moment estimators:** EMA, **conditional estimator**
- convergence analysis
  - two-step convergence analysis: **from conceptual to practical**
  - **SiF weighted aiming condition** (neither convexity nor smoothness)
  - decoupled weight decay: **stronger convergence guarantees**
  - **bias-variance tradeoff** for practical algorithms (better with smoothness)
- more empirical study to understand full potential

References I

J. R. Blum. Multidimensional Stochastic Approximation Methods. *The Annals of Mathematical Statistics*, 25(4):737 – 744, 1954. doi: 10.1214/aoms/1177728659. URL https://doi.org/10.1214/aoms/1177728659.

K. L. Chung. On a stochastic approximation method. *The Annals of Mathematical Statistics*, pages 463–483, 1954.

A. Dvoretzky. On stochastic approximation. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 39–55. University of California Press, 1956.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *Proceedings of International Conference on Learning Representations (ICLR)*, 2015. arXiv:1412.6980.

References II

I. Loshchilov and F. Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations (ICLR)*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407, 1951.

H. Robbins and D. Siegmund. A convergence theorem for non negative almost supermartingales and some applications. In J. S. Rustagi, editor, *Optimizing Methods in Statistics*, pages 233–257. Academic Press, 1971. ISBN 978-0-12-604550-5. doi: https://doi.org/10.1016/B978-0-12-604550-5.50015-8. URL https://www.sciencedirect.com/science/article/pii/B9780126045505500158.

References III

D. J. Sakrison. Stochastic approximation: A recursive method for solving regression problems. In A. Balakrishnan, editor, *Advances in Communication Systems*, volume 2, pages 51–106. Elsevier, 1966. doi: https://doi.org/10.1016/B978-1-4832-2939-3.50008-9. URL https://www.sciencedirect.com/science/article/pii/B9781483229393500089.

T. Tieleman and G. Hinton. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26–31, 2012.

J. H. Venter. On Dvoretzky Stochastic Approximation Theorems. *The Annals of Mathematical Statistics*, 37(6):1534 – 1544, 1966. doi: 10.1214/aoms/1177699145. URL https://doi.org/10.1214/aoms/1177699145.