Do STOCHASTIC, Feel Noiseless:

Stable Optimization via a Double Momentum Mechanism

Kfir Y. Levy

TECHNION
Israel Institute of Technology

Workshop on regret, optimization & games, Paris 2025

Focus

Stochastic Convex Optimization:

- Captures classical ML problems; e.g. logistic/linear regression
- Theoretical testbed for ML algorithms; e.g. AdaGrad & Adam

Focus

Stochastic Convex Optimization:

- Captures classical ML problems; e.g. logistic/linear regression
- Theoretical testbed for ML algorithms; e.g. AdaGrad & Adam

Today: A new gradient estimator with surprising properties

Stable stochastic optimization

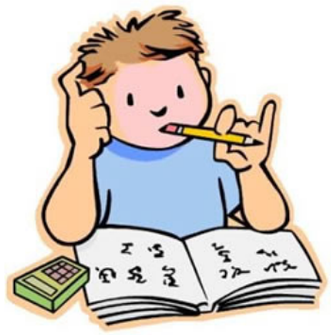
ML as Stochastic Optimization

Unknown distribution $(x, y) \sim \mathcal{D}$

$$f(w; x, y)$$

Loss weight

Learner



ML as Stochastic Optimization

Unknown distribution $(x, y) \sim \mathcal{D}$

Expected Loss:

$$\min_w F(w) \;:=\; \mathbf{E}_{(x,y)\sim\mathcal{D}} f(w; x, y)$$

Loss weight

ML as Stochastic Optimization

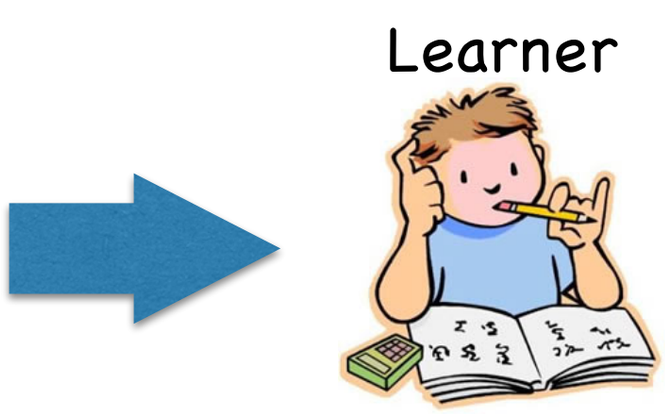
Unknown distribution $(x, y) \sim \mathcal{D}$

Expected Loss:

$$\min_w F(w) \;:=\; \mathbf{E}_{(x,y)\sim\mathcal{D}} f(w; x, y)$$

Training set

$(x_1, y_1), \ldots, (x_T, y_T) \sim \mathcal{D}$



Solution

$\overline{w}_T$

ML as Stochastic Optimization

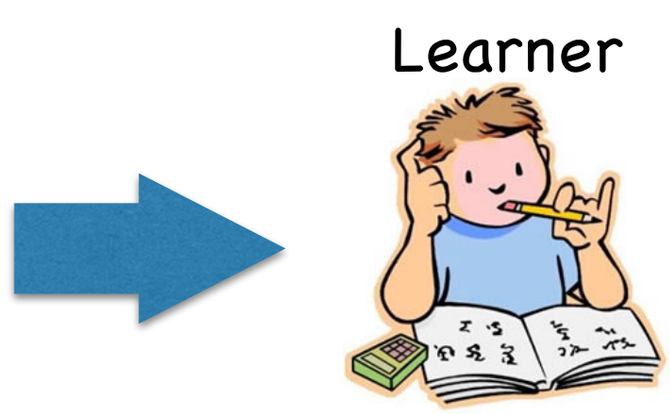
Unknown distribution $(x, y) \sim \mathcal{D}$

Expected Loss:

$$\min_w F(w) \;:=\; \mathbf{E}_{(x,y)\sim\mathcal{D}} f(w; x, y)$$

Training set

$(x_1, y_1), \ldots, (x_T, y_T) \sim \mathcal{D}$



Solution

$\overline{w}_T$

- **Performance Measure:**

$$\text{error}(\overline{w}_T) \;=\; F(\overline{w}_T) - \min_w F(w)$$

ML as Stochastic Optimization

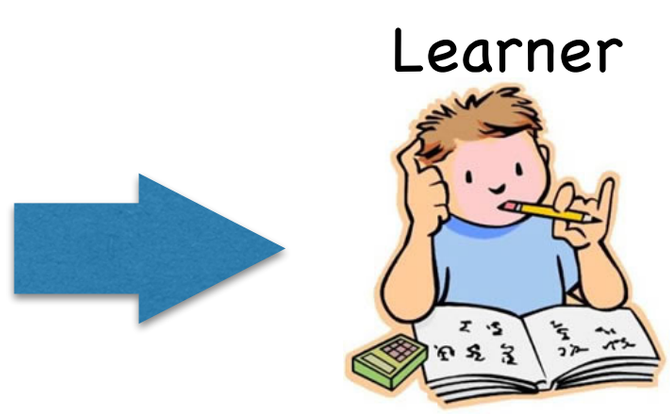
Unknown distribution $(x, y) \sim \mathcal{D}$

Expected Loss:

$$\min_w F(w) \;:=\; \mathbf{E}_{(x,y)\sim\mathcal{D}} f(w; x, y)$$

Training set

$$(x_1, y_1), \ldots, (x_T, y_T) \sim \mathcal{D}$$



Solution

$$\overline{w}_T$$

- **Performance Measure:**

$$\text{error}(\overline{w}_T) \;=\; F(\overline{w}_T) - \min_w F(w)$$

*Generalization
#samples to ensure error $\leq \epsilon$

*Computational Complexity

ML as Stochastic Optimization

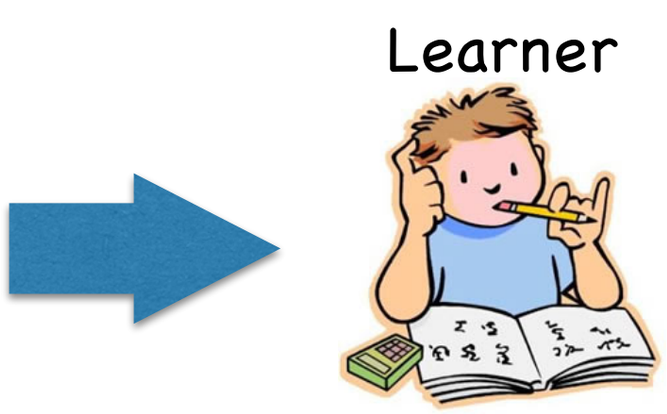
Unknown distribution $z \sim \mathcal{D}$

Expected Loss:

$$\min_w F(w) \;\; := \mathbf{E}_{z\sim\mathcal{D}}[f(w;z)]$$

Training set

$z_1, \ldots, z_T \sim \mathcal{D}$



Solution

$\overline{w}_T$

ML as Stochastic Optimization

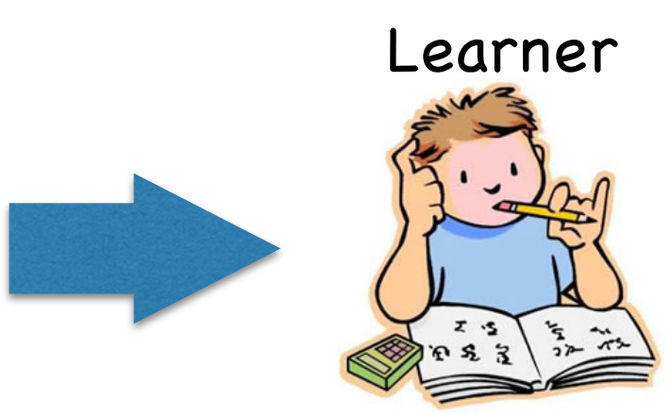
Unknown distribution $z \sim \mathcal{D}$

Expected Loss:

$$\min_w F(w) \ := \mathbf{E}_{z\sim\mathcal{D}}[f(w;z)]$$

Training set

$z_1, \ldots, z_T \sim \mathcal{D}$



Solution

$\overline{w}_T$

*Assumption: Smoothness

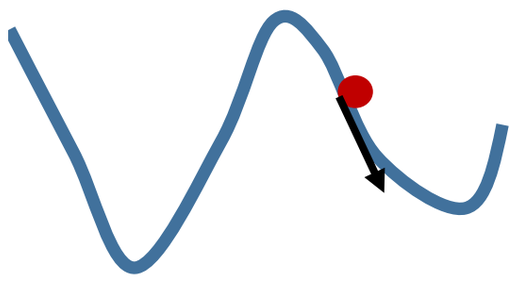
$$\|\nabla f(w;z) - \nabla f(u;z)\| \leq L\|w-u\|$$

Gradient Descent (GD)

$$\min_w F(w) \;:= \mathbf{E}_{z\sim\mathcal{D}}[f(w;z)]$$

- Update,

$$w_{t+1} \leftarrow w_t - \eta \nabla F(w_t)$$



Gradient Descent (GD)

GD is quite simple:

- Can use **fixed** learning rate $\eta = 1/L$

Gradient Descent (GD)

GD is quite simple:

- Can use **fixed** learning rate $\eta = 1/L$
- May use $\|\nabla F(w_t)\|$ as a stopping criteria

Gradient Descent (GD)

GD is quite simple:

- Can use **fixed** learning rate $\eta = 1/L$
- May use $\|\nabla F(w_t)\|$ as a stopping criteria
- Compare between solutions by simply measuring $F(w_t)$

Stochastic Gradient Descent (SGD)

$$\min_w F(w) \;:= \mathbf{E}_{z\sim\mathcal{D}}[f(w;z)]$$

In each round t :

- Draw sample $z_t \sim \mathcal{D}$
- Update,

$$w_{t+1} \;\leftarrow\; w_t - \eta \nabla f(w_t; z_t)$$

Stochastic Gradient Descent (SGD)

$$\min_w F(w) \;:= \mathbf{E}_{z\sim\mathcal{D}}[f(w;z)]$$

In each round t :

- Draw sample $z_t \sim \mathcal{D}$
- Update,

$$w_{t+1} \;\leftarrow\; w_t - \eta \nabla f(w_t; z_t)$$

Stochastic Gradient Descent (SGD)

Issues with SGD:

- Requires careful tuning of learning rate

$$\eta_t = \frac{1}{L + (\sigma/D)\sqrt{t}}$$

Stochastic Gradient Descent (SGD)

Issues with SGD:

- Requires careful tuning of learning rate

$$\eta_t = \frac{1}{L + (\sigma/D)\sqrt{t}}$$

- May **not** use $\|\nabla f(w_t; z_t)\|$ as a stopping criteria

Stochastic Gradient Descent (SGD)

Issues with SGD:

- Requires careful tuning of learning rate

$$\eta_t = \frac{1}{L + (\sigma/D)\sqrt{t}}$$

- May **not** use $\|\nabla f(w_t; z_t)\|$ as a stopping criteria

- Requires **Test/Validation** set -> evaluation/hyperparameter-tuning
  Ideally like to avoid this:
  *Small datasets: like to use all data for training
  *Huge datasets: higher computational burden

Stochastic Gradient Descent (SGD)

Issues with SGD:

- Requires careful tuning of learning rate

$$\eta_t = \frac{1}{\lambda t}$$

- M

- Re

Ideally like to avoid this:

*Small datasets: like to use all data for training

*Huge datasets: higher computational burden

Reason:

$$\|\nabla f(w_t; z_t) - \nabla F(w_t)\| = O(1)$$

$$\|f(w_t; z_t) - F(w_t)\| = O(1)$$

Stochastic Gradient Descent (SGD)

Issues with SGD:

- Requires careful tuning of learning rate

$$\eta_t = \frac{1}{t}$$

- M

- Re

Ideally like to avoid this:

*Small datasets: like to use all data for training

*Huge datasets: higher computational burden

Q: can we design an SGD variant with similar properties to GD?

μ^2-Stochastic Gradient Descent (SGD)

A new SGD variant:

- Same convergence rate as optimal SGD

μ^2-Stochastic Gradient Descent (SGD)

A new SGD variant:

- Same convergence rate as optimal SGD

- Allows using a fixed learning rate $\eta = 1/4L$, irrespective of the Noise!

μ^2-Stochastic Gradient Descent (SGD)

A new SGD variant:

- Same convergence rate as optimal SGD

- Allows using a fixed learning rate $\eta = 1/4L$, irrespective of the Noise!

- May use gradient estimates as stopping criteria

μ^2-Stochastic Gradient Descent (SGD)

A new SGD variant:

- Same convergence rate as optimal SGD

- Allows using a fixed learning rate $\eta = 1/4L$, irrespective of the Noise!

- May use gradient estimates as stopping criteria

- No need for Test/Validation set: can evaluate well on-the-fly

μ^2-Stochastic Gradient Descent (SGD)

A new SGD variant:

- Same convergence rate as optimal SGD

- Allows using a fixed learning rate $\eta = 1/4L$, irrespective of the Noise!

- May use gradient estimates as stopping criteria

- No need for Test/Validation set: can evaluate well on-the-fly

- Stable: Achieves **optimal** convergence for $\eta \in [\eta_{\min}, \eta_{\max}]$

$$\eta_{\max}/\eta_{\min} \approx (\sigma/L)\sqrt{T}$$

μ^2-ExtraSGD

A new accelerated SGD variant:

- Optimal (accelerated) convergence rate

- Allows using a fixed learning rate $\eta = 1/4L$, irrespective of the Noise!

- May use gradient estimates as stopping criteria

- No need for Test/Validation set: can evaluate well on-the-fly

- Stable: Achieves **optimal** convergence for $\eta \in [\eta_{\min}, \eta_{\max}]$

$$\eta_{\max}/\eta_{\min} \approx (\sigma/L)T^{3/2}$$

μ^2-SGD & μ^2-ExtraSGD

New gradient estimator:

$$\|d_t - \nabla F(x_t)\| = O(1/\sqrt{t})$$

μ^2-SGD & μ^2-ExtraSGD

A new accelerated SGD variant:

- Optimal (accelerated) convergence rates

- Allows using a fixed learning rate $\eta = 1/4L$, irre

- May

- No e-fly

- Stable: Achieves **optimal** convergence for $\eta \in [\eta_{\min}, \eta_{\max}]$

$$\eta_{\max}/\eta_{\min} \approx (\sigma/L)T^{3/2}$$

New gradient/value estimators:

$$\|d_t - \nabla F(x_t)\| = O(1/\sqrt{t})$$

$$\|\hat{f}_t - F(x_t)\| = O(1/\sqrt{t})$$

Algorithmic Technique

Combination of two mechanisms -> μ^2

Algorithmic Technique

Combination of two mechanisms -> μ^2

Anytime
gradient mechanism

Momentum

[Cutkosky, 2019]

[Kavis L Bach & Cevher, 2019]

Algorithmic Technique

Combination of two mechanisms -> μ^2

Corrected
Momentum Mechanism

Momentum

[Cutkosky & Orabona, 2019]

Anytime
gradient mechanism

Momentum

[Cutkosky, 2019]
[Kavis L Bach & Cevher, 2019]

Algorithmic Technique

Combination of two mechanisms -> μ^2

Corrected
Momentum Mechanism

Momentum

[Cutkosky & Orabona, 2019]

Anytime
gradient mechanism

Momentum

[Cutkosky, 2019]
[Kavis L Bach & Cevher, 2019]

Mechanism I: Anytime

Averaging **Query Points:**

SGD:

$$w_{t+1} \leftarrow w_t - \eta \nabla f(w_t; z_t)$$

Mechanism I: Anytime

Averaging **Query Points:**

Anytime-SGD:

$$w_{t+1} \leftarrow w_t - \eta \nabla f(x_t; z_t)$$

Averaging:

$$x_t = \frac{1}{t} \sum_{\tau=1}^{t} w_\tau$$

[Cutkosky, 2019] [Kavis L Bach & Cevher, 2019]

Mechanism I: Anytime

Averaging **Query Points:**

Anytime-SGD:

$$w_{t+1} \leftarrow w_t - \eta \nabla f(x_t; z_t)$$

- Output: x_T

- Enjoys the same guarantees as Optimal SGD
- Can be directly related to momentum
- Highly related to acceleration
- Gradient error is still O(1)

Mechanism II: Corrected Momentum

Averaging **Gradients+correcting bias:**

Momentum SGD:

$$w_{t+1} \leftarrow w_t - \eta \, d_t$$

- Momentum:

$$d_t \leftarrow a_t \nabla f(w_t; z_t) + (1 - a_t) d_t$$

Mechanism II: Corrected Momentum

Averaging **Gradients+correcting bias:**

Momentum SGD:

$$w_{t+1} \leftarrow w_t - \eta \, d_t$$

- Momentum:

$$d_t \leftarrow a_t \nabla f(w_t; z_t) + (1 - a_t) d_t$$

Biased estimate of $\nabla F(w_t)$!

Mechanism II: Corrected Momentum

Averaging **Gradients+correcting bias:**

Momentum SGD:

$$w_{t+1} \leftarrow w_t - \eta \, d_t$$

- Momentum:

$$d_t \leftarrow a_t \nabla f(w_t; z_t) + (1 - a_t) d_{t-1}$$

- Corrected Momentum:

$$d_t \leftarrow \nabla f(w_t; z_t) + (1 - a_t)(d_{t-1} - \nabla f(w_{t-1}; z_t))$$

[Cutkosky & Orabona, 2019]

Mechanism II: Corrected Momentum

Averaging **Gradients+correcting bias:**

Momentum SGD:

$$w_{t+1} \leftarrow w_t - \eta \, d_t$$

- Momentum:

$$d_t \leftarrow a_t \nabla f(w_t; z_t) + (1 - a_t) d_{t-1}$$

- Corrected Momentum:

$$d_t \leftarrow \nabla f(w_t; z_t) + (1 - a_t)(d_{t-1} - \nabla f(w_{t-1}; z_t))$$

Unbiased estimate of $\nabla F(w_t)$

Mechanism II: Corrected Momentum

Averaging **Gradients+correcting bias:**

Momentum SGD:

$$w_{t+1} \leftarrow w_t - \eta \, d_t$$

- Momentum:

$$d_t \leftarrow a_t \nabla f(w_t; z_t) + (1 - a_t) d_{t-1}$$

- Corrected Momentum:

$$d_t \leftarrow \nabla f(w_t; z_t) + (1 - a_t)(d_{t-1} - \nabla f(w_{t-1}; z_t))$$

Variance reduction effect $\propto \|w_t - w_{t-1}\|^2$

Mechanism II: Corrected Momentum

Averaging **Gradients+correcting bias:**

STochastic Recursive Momentum (STORM) SGD:

$$w_{t+1} \leftarrow w_t - \eta \, d_t$$

- Corrected Momentum:

$$d_t \leftarrow \nabla f(w_t; z_t) + (1 - a_t)(d_{t-1} - \nabla f(w_{t-1}; z_t))$$

- Originally designed for Non-convex problems
- Enjoys the same guarantees as Optimal SGD
- Gradient error depends on learning rate

μ^2-SGD

Combination of **STORM+Anytime**:

μ^2-SGD:

$$w_{t+1} \leftarrow w_t - \eta \, d_t$$

Estimate for $\nabla F(x_t)$

$$x_t = \frac{1}{t} \sum_{\tau=1}^{t} w_\tau$$

μ^2-SGD

Combination of **STORM+Anytime**:

μ^2-SGD:

$$w_{t+1} \leftarrow w_t - \eta \, d_t$$

- Corrected Momentum:

$$d_t \leftarrow \nabla f(x_t; z_t) + (1 - a_t)(d_{t-1} - \nabla f(x_{t-1}; z_t))$$

μ^2-SGD

Combination of **STORM+Anytime**:

μ^2-SGD:

$$w_{t+1} \leftarrow w_t - \eta \, d_t$$

- Corrected Momentum:

$$d_t \leftarrow \nabla f(x_t; z_t) + (1 - a_t)(d_{t-1} - \nabla f(x_{t-1}; z_t))$$

Variance reduction effect $\propto \|x_t - x_{t-1}\|^2$

μ^2-SGD

Combination of **STORM+Anytime**:

μ^2-SGD:

$$w_{t+1} \leftarrow w_t - \eta\, d_t$$

- Corrected Momentum:

$$d_t \leftarrow \nabla f(x_t; z_t) + (1 - a_t)(d_{t-1} - \nabla f(x_{t-1}; z_t))$$

Variance reduction effect $\propto \|x_t - x_{t-1}\|^2$

Intuitively: $\|x_t - x_{t-1}\|^2 \ll \|w_t - w_{t-1}\|^2$

μ^2-SGD

Combination of **STORM+Anytime**:

μ^2-SGD:

$$w_{t+1} \leftarrow w_t - \eta\, d_t$$

- Corrected Momentum:

$$d_t \leftarrow \nabla f(x_t; z_t) + (1-a_t)(d_{t-1} - \nabla f(x_{t-1}; z_t))$$

Indeed: $\|d_t - \nabla F(x_t)\| = O(1/\sqrt{t})$

μ^2-SGD

Combination of **STORM+Anytime**:

μ^2-SGD:

$$w_{t+1} \leftarrow w_t - \eta\, d_t$$

- Corrected Momentum:

$$d_t \leftarrow \nabla f(x_t; z_t) + (1 - a_t)(d_{t-1} - \nabla f(x_{t-1}; z_t))$$

Indeed: $\|d_t - \nabla F(x_t)\| = O(1/\sqrt{t})$

Can similarly design **Value** estimates:

$$\|\hat{f}_t - F(x_t)\| = O(1/\sqrt{t})$$

μ^2-Stochastic Gradient Descent (SGD)

A new SGD variant:

- Same convergence rate as optimal SGD

- Allows using a fixed learning rate $\eta = 1/4L$, irrespective of the Noise!

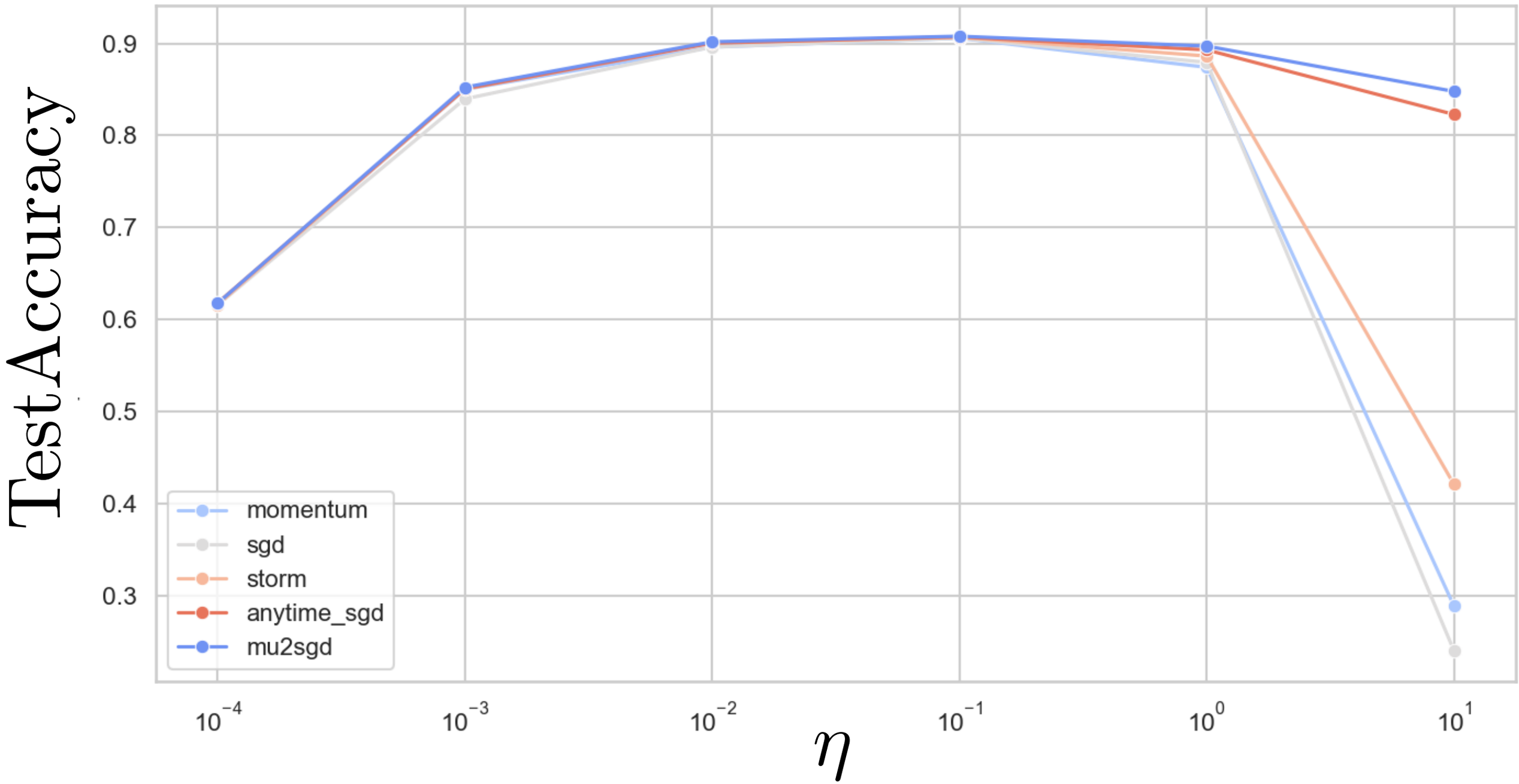
- May use gradient estimates as stopping criteria

- No need for Test/Validation set: can evaluate well on-the-fly

- Stable: Achieves **optimal** convergence for $\eta \in [\eta_{\min}, \eta_{\max}]$

$$\eta_{\max}/\eta_{\min} \approx (\sigma/L)\sqrt{T}$$

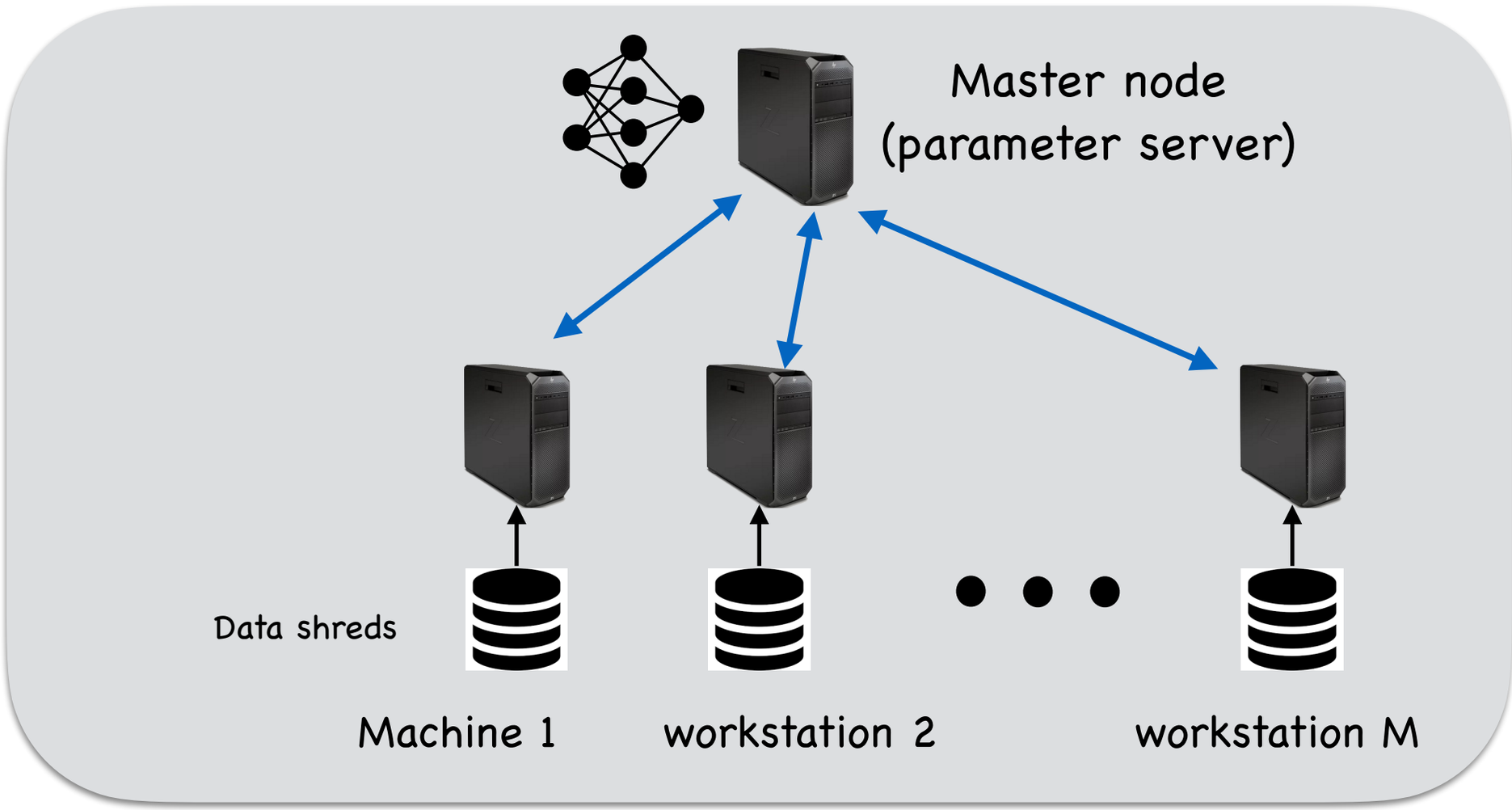
Experiment: CIFAR10, Resnet18



Applications

Distributed Learning

Centralized Distributed Training

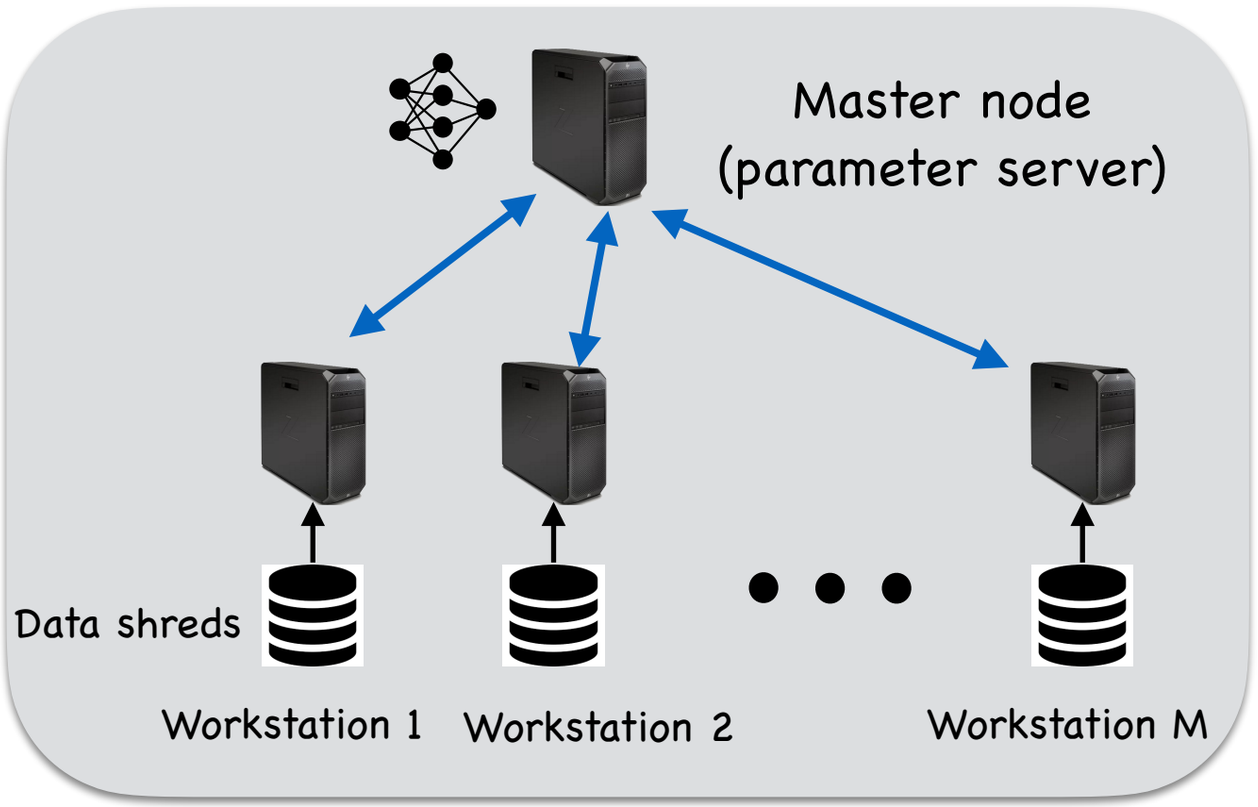


Parallelization via. **Minibatch SGD**

Minibatch SGD

In each round t :

- Compute $g_t \leftarrow \text{Average}\left(g_t^{(1)}, \ldots, g_t^{(M)}\right)$
- Update, $w_{t+1} \leftarrow w_t - \eta g_t$



Parallelization via. **Minibatch Mu2 SGD**

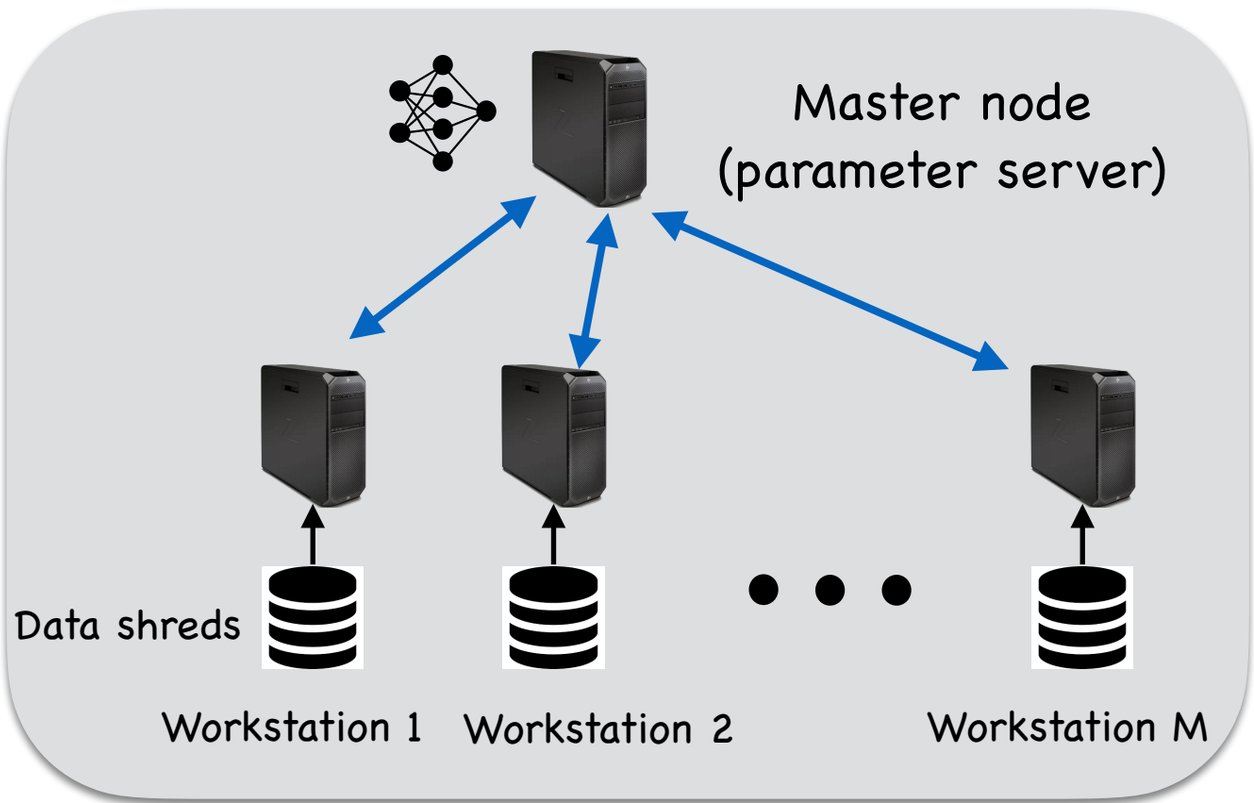
Minibatch Mu2 SGD

In each round t :

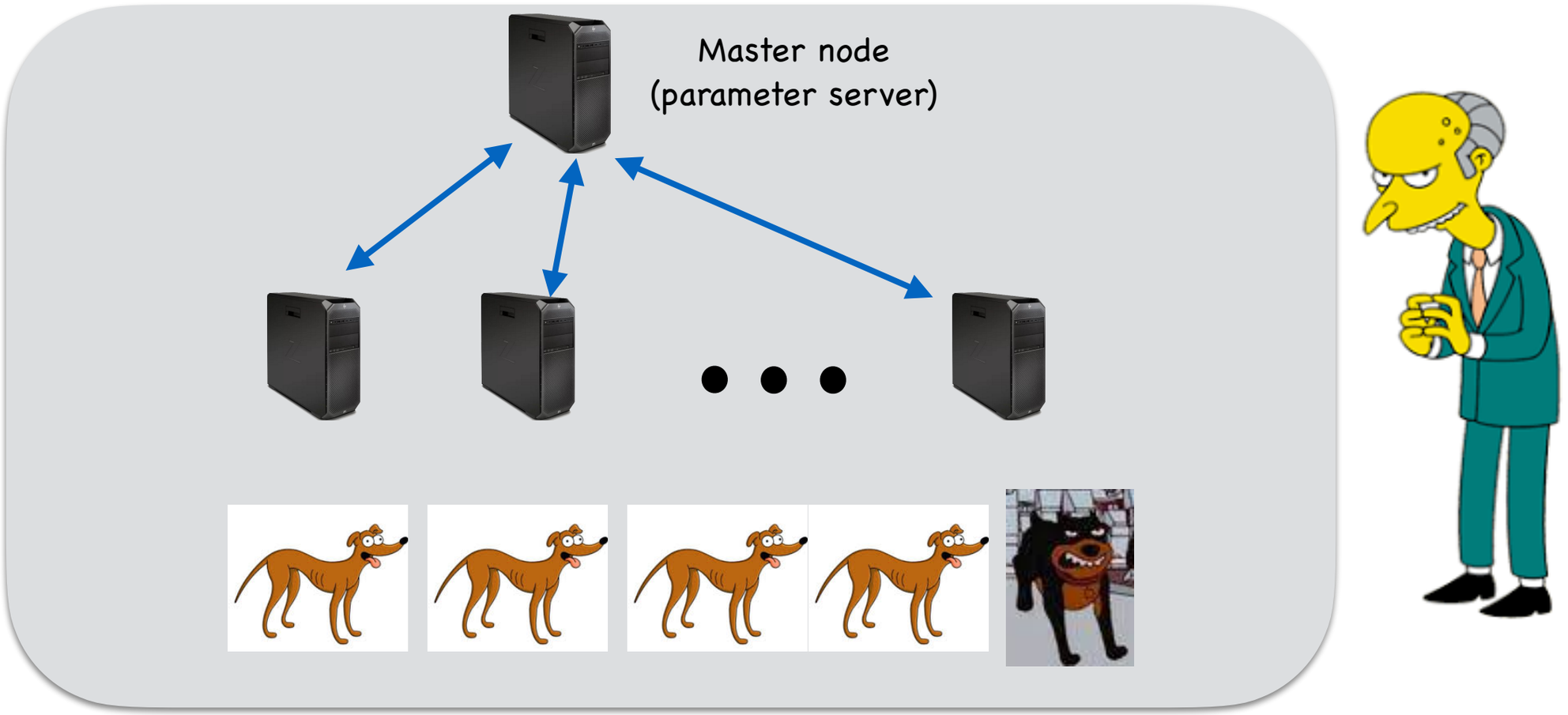
- Compute $d_t \leftarrow \text{Average}\left(d_t^{(1)}, \ldots, d_t^{(M)}\right)$
- Update,

$$w_{t+1} \leftarrow w_t - \eta d_t$$

$$\& \quad x_{t+1} \leftarrow \text{Average}(w_1, \ldots, w_{t+1})$$



Challenge I: Byzantine Training



[Dahan & L, 2024a]

[Dahan & L, 2024b]

Challenge II: Distributed Training with Differential Privacy

Learning with Privacy Guarantees

***Learning should not uncover private data**

[Reshef & L, 2024]

[Reshef & L, 2025]

Extensions & Future Directions

Extension:

-Can ensure $O(1/\sqrt{T})$ error along all rounds,

only $\log T$ increase in sample complexity

Future directions:

-Other interesting applications? Stochastic Control?


What about games?

Convex-concave Zero-sum Games

Goal of U-player

$$\min_{u \in U} \max_{v \in V} \; M(u, v)$$

Goal of V-player

$$\max_{v \in V} \min_{u \in U} \; M(u, v)$$

Duality-gap:

$$\mathrm{DualityGap}(u, v) := \max_{v_0 \in V} M(u, v_0) - \min_{u_0 \in U} M(u_0, v)$$

Goal: find (approximate) Equilibrium,

$$\mathrm{DualityGap}(u^*, v^*) := 0$$

Games in ML: Boosting, GANs, Constrained Opt.

Classical Reduction: Freund & Schapire

$$\min_{u \in U} \max_{v \in V} \; M(u, v)$$

U-Player: $\{\ell_t(u) := M(u, v_t)\}_{t=1,\ldots,T}$

V-Player: $\{r_t(v) := M(u_t, v)\}_{t=1,\ldots,T}$

$$\mathrm{DualityGap}(\bar{u}_T, \bar{v}_T) \leq \frac{1}{T}\left(\mathrm{Reg}_T^U + \mathrm{Reg}_T^V\right)$$

Classical Reduction: Freund & Schapire

$$\min_{u \in U} \max_{v \in V} \; M(u, v)$$

U-Player: $\{\ell_t(u) := \nabla_u M(u_t, v_t) \cdot u\}_{t=1,\ldots,T}$

V-Player: $\{r_t(v) := M(u_t, v)\}_{t=1,\ldots,T}$

$$\text{DualityGap}(\bar{u}_T, \bar{v}_T) \leq \frac{1}{T}\left(\text{Reg}_T^U + \text{Reg}_T^V\right)$$

New Anytime Reduction (Asymmetric)

$$\min_{u \in U} \max_{v \in V} \; M(u, v)$$

U-Player: $\{\ell_t(u) := \nabla_u M(\bar{u}_t, v_t) \cdot u\}_{t=1,\ldots,T}$

V-Player: $\{r_t(v) := tM(\bar{u}_t, v) - (t-1)M(\bar{u}_{t-1}, v)\}_{t=1,\ldots,T}$

$$\mathrm{DualityGap}(\bar{\mathrm{u}}_\mathrm{T}, \bar{\mathrm{v}}_\mathrm{T}) \leq \frac{1}{\mathrm{T}} \left(\mathrm{Reg}_\mathrm{T}^\mathrm{U} + \mathrm{Reg}_\mathrm{T}^\mathrm{V}\right)$$

Future Directions

Future directions:

MinMax Games with Stochastic Feedback?

- Extension to Multi-Agent Games?
- Interesting Structure for Bandit Feedback on Y Player

Merci Beaucoup!