# Robust Learning Rate Selection for Stochastic Optimization Using Splitting Diagnostic

Matteo Sordello

JSM 2019

Joint work with Weijie Su

# Stochastic Optimization

Problem: find the minimizer $\theta^*$ of a function $F(\theta)$ when $\nabla F(\theta)$ is unknown

# Stochastic Optimization

Problem: find the minimizer $\theta^*$ of a function $F(\theta)$ when $\nabla F(\theta)$ is unknown

- Online Learning: $F(\theta) = \mathbb{E}\left[f(\theta, Z)\right]$

# Stochastic Optimization

Problem: find the minimizer $\theta^*$ of a function $F(\theta)$ when $\nabla F(\theta)$ is unknown

- Online Learning: $F(\theta) = \mathbb{E}\left[f(\theta, Z)\right]$
- Empirical Risk Minimization: finite population of size $n$, which is *extremely large*, and $F(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(\theta, z_i)$

# Stochastic Optimization

Problem: find the minimizer $\theta^*$ of a function $F(\theta)$ when $\nabla F(\theta)$ is unknown

- Online Learning: $F(\theta) = \mathbb{E}\left[f(\theta, Z)\right]$
- Empirical Risk Minimization: finite population of size $n$, which is *extremely large*, and $F(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(\theta, z_i)$

Noisy gradient $g(\theta, Z) = \nabla_\theta f(\theta, Z)$, satisfying $\mathbb{E}\left[g(\theta, Z)\right] = \nabla F(\theta)$

# Stochastic Optimization

Problem: find the minimizer $\theta^*$ of a function $F(\theta)$ when $\nabla F(\theta)$ is unknown

- Online Learning: $F(\theta) = \mathbb{E}\left[f(\theta, Z)\right]$
- Empirical Risk Minimization: finite population of size $n$, which is *extremely large*, and $F(\theta) = \frac{1}{n}\sum_{i=1}^{n} f(\theta, z_i)$

Noisy gradient $g(\theta, Z) = \nabla_\theta f(\theta, Z)$, satisfying $\mathbb{E}\left[g(\theta, Z)\right] = \nabla F(\theta)$

SGD (Robbins and Monro (1951)):

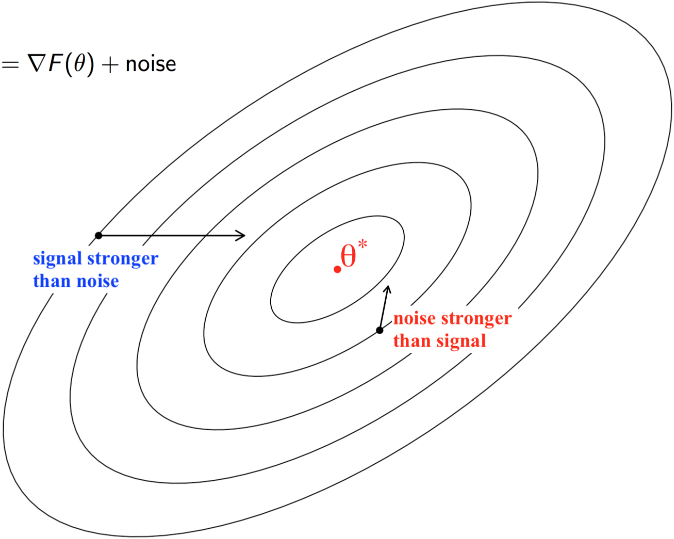
From a starting point $\theta_0$, SGD recursively updates

$$\theta_{t+1} = \theta_t - \eta_t \cdot g(\theta_t, Z_{t+1})$$

$\eta_t$ is the **learning rate**.

# How to Select the Learning Rate

$g(\theta, Z) = \nabla F(\theta) + \text{noise}$



# Popular Choices for the Learning Rate

- $\eta_t = \eta$
  - convergence is **not guaranteed!** [Moulines and Bach (2011)]

# Popular Choices for the Learning Rate

- $\eta_t = \eta$
  - convergence is **not guaranteed!** [Moulines and Bach (2011)]
- $\eta_t \propto t^{-\alpha}$ with $\alpha \in (0.5, 1)$ [Robbins and Monro (1951)]
  - heavily dependent on the initial learning rate

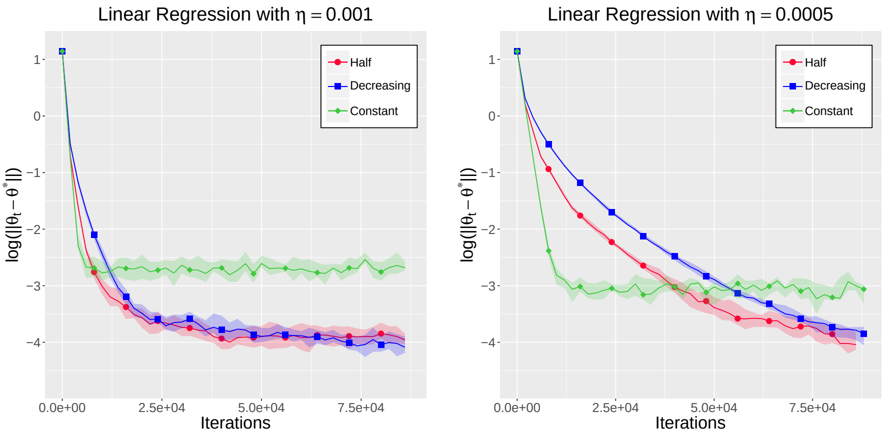
# Popular Choices for the Learning Rate

- $\eta_t = \eta$
  - convergence is **not guaranteed!** [Moulines and Bach (2011)]
- $\eta_t \propto t^{-\alpha}$ with $\alpha \in (0.5, 1)$ [Robbins and Monro (1951)]
  - heavily dependent on the initial learning rate
- $\eta_t = \eta$ for the first $t_1$ iterations, then it gets halved and so on. This procedure is called $\text{SGD}^{1/2}$. [Bottou et al. (2018)]
  - also not robust

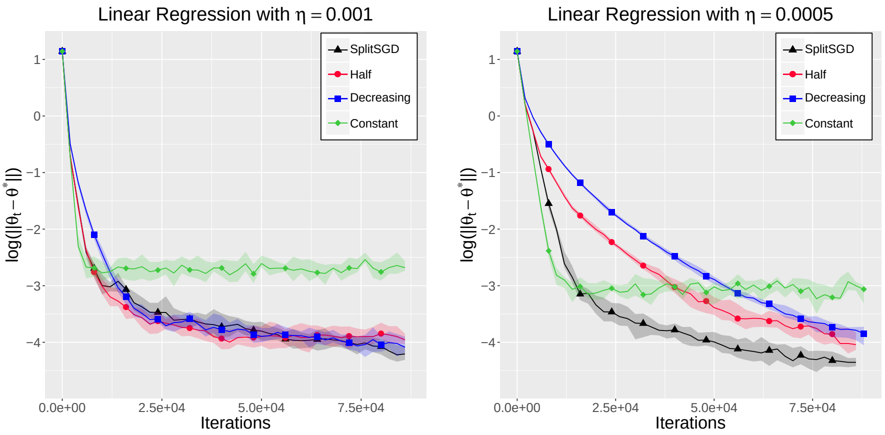
# Popular Choices for the Learning Rate

- $\eta_t = \eta$
  - convergence is **not guaranteed!** [Moulines and Bach (2011)]
- $\eta_t \propto t^{-\alpha}$ with $\alpha \in (0.5, 1)$ [Robbins and Monro (1951)]
  - heavily dependent on the initial learning rate
- $\eta_t = \eta$ for the first $t_1$ iterations, then it gets halved and so on. This procedure is called SGD$^{1/2}$. [Bottou et al. (2018)]
  - also not robust
- Adaptive methods:
  - `pflug` Diagnostic [Chee, Toulis (2018)]
  - AdaGrad [Duchi, Hazan, Singer (2011)]
  - Adam [Kingma, Ba (2015)]

# Behavior of Classic Methods



# Robustness of Our Method



# Outline

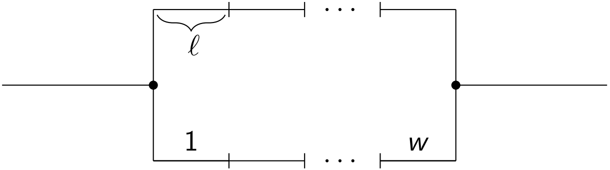
- Introduce the Splitting Diagnostic
- Theoretical guarantees for Splitting Diagnostic
- Introduce procedure SplitSGD

# Splitting Diagnostic

**Goal:**

Detect the **phase transition**, so we can keep $\eta$ constant until stationarity.

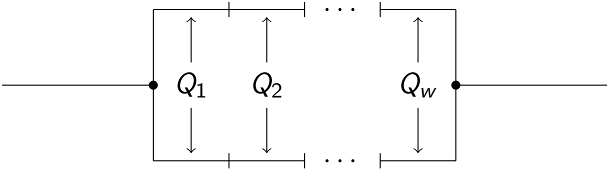
- Run two SGD threads (related to HiGrad [Su and Zhu (2018)])
- If the noisy gradients point on average in the same direction, we are still approaching the minimizer
- If not, we reached stationarity



# Splitting Diagnostic

**Definition:**

- $\bar{g}_i^{(k)} :=$ is the average noisy gradient in window $i$ and thread $k$
- $Q_i(\theta_0, \eta, l) = \langle \bar{g}_i^{(1)}, \bar{g}_i^{(2)} \rangle$ is the **gradient coherence**, which is positive if $\bar{g}_i^{(1)}$ and $\bar{g}_i^{(2)}$ have approximately the same direction



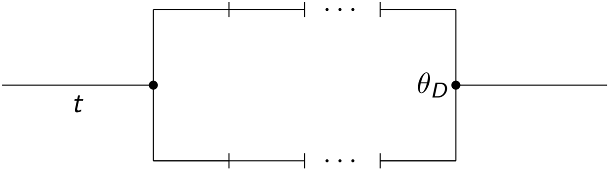
# Output

A binary value that tells if stationarity is detected,

$$T_D = \begin{cases} \text{STATIONARY} & \text{if there are enough negative } Q_i \\ \text{NON STATIONARY} & \text{otherwise} \end{cases}$$

and the average of the last iterates in the two threads

$$\theta_D := \frac{\theta^{(1)}_{t+w\cdot\ell} + \theta^{(2)}_{t+w\cdot\ell}}{2}$$



# Theoretical Guarantees for $\eta$ small

We want the diagnostic to say that stationarity has not been reached yet.

# Theoretical Guarantees for $\eta$ small

We want the diagnostic to say that stationarity has not been reached yet.

Theorem:

If $F(\theta)$ is $L$-smooth, and $\mathbb{E}\left[||g(\theta, Z)||^2\right] \leq G^2$, then for any fixed $t \in \mathbb{N}$ and $i \in \{1, ..., w\}$ we can set $\eta$ small enough such that

$$T_D = \text{NON STATIONARY}$$

with high probability.

# Theoretical Guarantees for $\eta$ small

We prove it by showing that

$$\mathrm{sd}(Q_i) \leq C_1(\eta, \ell) \cdot \mathbb{E}[Q_i]$$

where $C_1(\eta, \ell)$ can be made arbitrarily small. When $\eta$ is small, the gradient coherence is positive.

# Theoretical Guarantees for $t$ large

If $t \to \infty$, we want stationarity to be detected.

# Theoretical Guarantees for $t$ large

If $t \to \infty$, we want stationarity to be detected.

Theorem:

If $F(\theta)$ is $\mu$-strongly convex and $L$-smooth, and $\mathbb{E}\left[||g(\theta, Z)||^2\right] \leq G^2$, then for any $\eta \leq \mu/L^2$, $\ell \in \mathbb{N}$ and $i \in \{1, ..., w\}$, as $t \to \infty$ we have that

$$T_D = \text{STATIONARY}$$

with high probability.

# Theoretical Guarantees for $t$ large

We prove it by showing that

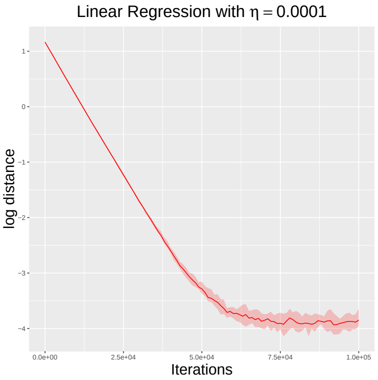
$$|\mathbb{E}[Q_i]| \leq C_2(\eta) \cdot \text{sd}(Q_i)$$

where $C_2(\eta) = C_2 \cdot \eta + o(\eta)$. When $t$ is large, the gradient coherence is distributed around 0.

# Comparison with `pflug` Diagnostic

$$\theta^* = (1, ..., 1)$$
$$\theta_0 = (\epsilon_1, ..., \epsilon_d)$$

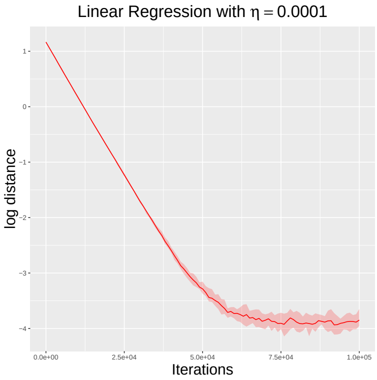
where $\epsilon_i \sim N(0, 0.1)$. We run multiple SGD threads and "eye-ball" the elbow of the distance with $\theta^*$.



# Comparison with `pflug` Diagnostic

$$\theta^* = (1, ..., 1)$$
$$\theta_0 = (\epsilon_1, ..., \epsilon_d)$$

where $\epsilon_i \sim N(0, 0.1)$. We run multiple SGD threads and "eye-ball" the elbow of the distance with $\theta^*$.



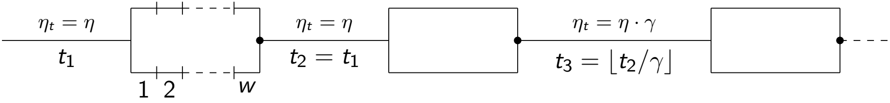

- Splitting Diagnostic declares stationarity after 47.000 iterations
- The `pflug` Diagnostic consistently estimates more than a million

# Comparison with pflug Diagnostic

Number of iterations before stationarity, multiply by 1000.

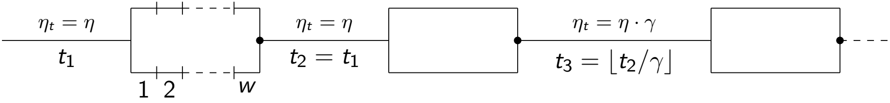
| | | Eyeballing | | pflug | | Splitting | |
|---|---|---|---|---|---|---|---|
| | start / $\eta$ | close | far | close | far | close | far |
| Linear | 0.001 | 4.0 | 5.0 | 4.7 | 717.6 | 6.1 | 10.3 |
| | 0.0001 | 30.0 | 50.0 | 65.3 | 1000.0 | 14.6 | 47.1 |
| Logistic | 0.01 | 5.0 | 10.0 | 0.8 | 51.5 | 15.7 | 17.1 |
| | 0.001 | 30.0 | 100.0 | 3.5 | 452.2 | 20.1 | 57.2 |

# SplitSGD



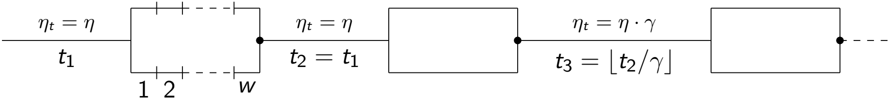
$\eta_t = \eta$
$t_1$
1
2
$w$
$\eta_t = \eta$
$t_2 = t_1$
$\eta_t = \eta \cdot \gamma$
$t_3 = \lfloor t_2/\gamma \rfloor$

# SplitSGD




- Run SGD with fixed learning rate $\eta$ on a single thread. The number of iterations $t_1$ if decided in advance.
- From $\theta_{t_1}$ split the single thread into two and start the Splitting Diagnostic.

# SplitSGD

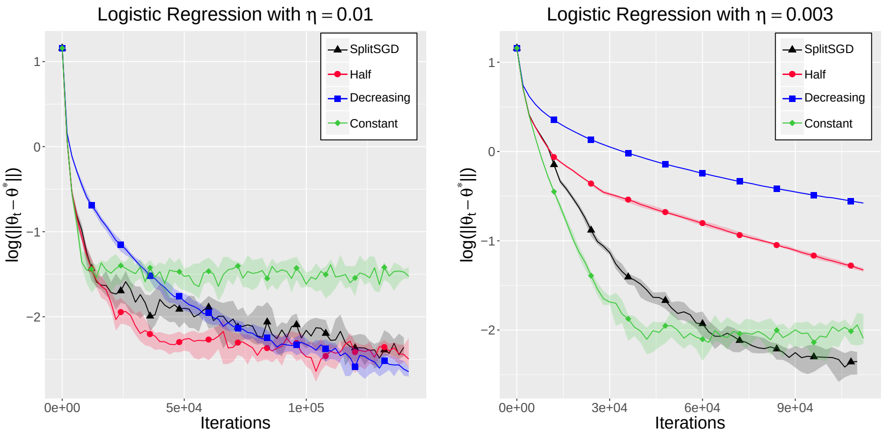



From the output of the diagnostic, $\theta_D$, restart a new single thread

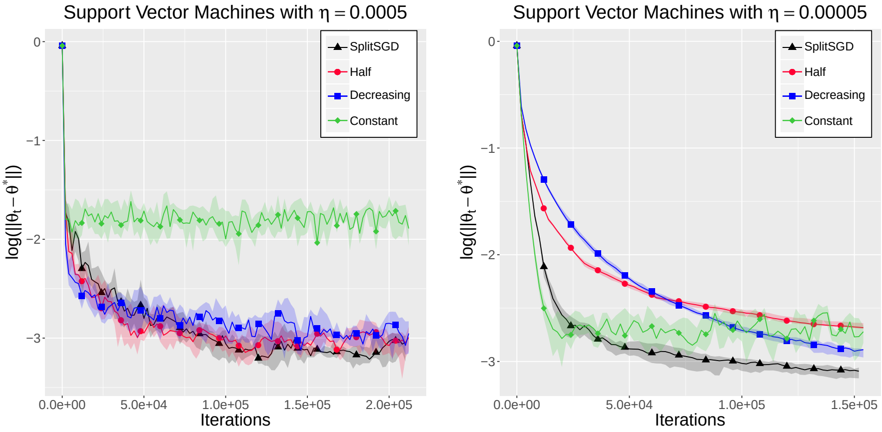
- of length $t_1$ and with learning rate $\eta$ if $T_D =$ NON STATIONARY.
- of length $\lfloor t_1/\gamma \rfloor$ and with learning rate $\eta \cdot \gamma$ if $T_D =$ STATIONARY.

$\gamma$ is the **discount factor**.

# Comparison with other SGDs (Logistic Regression)

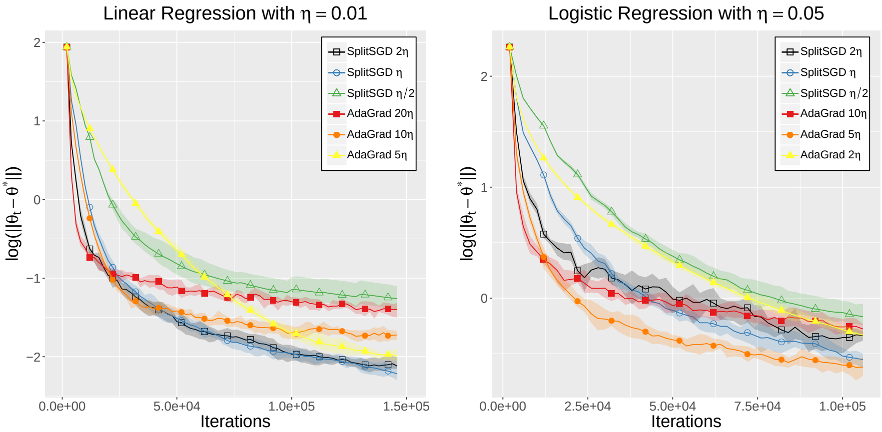


# Comparison with other SGDs (SVM)



# Comparison with AdaGrad

We used a **sparse** feature matrix. In Adagrad $\eta_t = \frac{\eta}{\sqrt{G_t^2 + \epsilon}} \in \mathbb{R}^d$



# Conclusion

We developed an efficient optimization method using a diagnostic that detects when SGD with constant learning rate has reached stationarity.

# Conclusion

We developed an efficient optimization method using a diagnostic that detects when SGD with constant learning rate has reached stationarity.

SplitSGD:

- no more computational cost than the standard SGD
- **robust** to the choice of the initial learning rate
- **robust** to the choice of the starting point

# Conclusion

We developed an efficient optimization method using a diagnostic that detects when SGD with constant learning rate has reached stationarity.

SplitSGD:

- no more computational cost than the standard SGD
- **robust** to the choice of the initial learning rate
- **robust** to the choice of the starting point

Future Work:

- Performance in nonconvex settings
- Incorporate into other existing methods (momentum, ...)

# References

[1] Chee, J. and Toulis, P. (2018). Convergence diagnostics for stochastic gradient descent with constant learning rate. *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, 1476-1485. PMLR.

[2] Su, W.J., and Zhu, Y. (2018). Uncertainty Quantification for Online Learning and Stochastic Approximation via Hierarchical Incremental Gradient Descent. *arXiv preprint arXiv:1802.04876*

Thank you!