Escaping Saddle Points with Adaptive Gradient Methods

Matthew Staib[1], Sashank Reddi[2], Satyen Kale[2], Sanjiv Kumar[2], Suvrit Sra[1]

1. MIT EECS
2. Google Research, New York

MIT EECS
ELECTRICAL ENGINEERING AND COMPUTER SCIENCE

Google

Adam, RMSProp and friends

Adam, RMSProp and friends

- Empirically: good non-convex performance

Adam, RMSProp and friends

- Empirically: good non-convex performance
- Limited theory, some non-convergence results [e.g. Reddi et al. ‘18]

Adam, RMSProp and friends

- Empirically: good non-convex performance
- Limited theory, some non-convergence results [e.g. Reddi et al. '18]
- Our take: adaptive methods escape saddles (in words: via isotropic noise), reach SOSPs

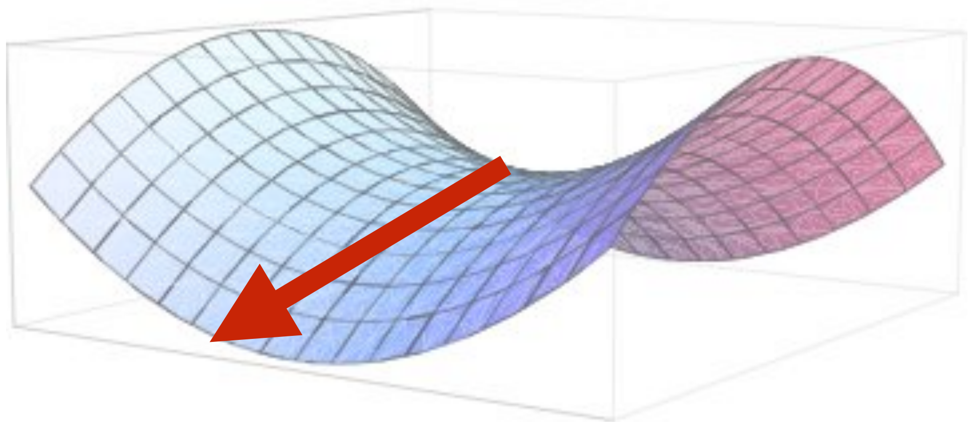
Adam, RMSProp and friends

- Empirically: good non-convex performance
- Limited theory, some non-convergence results [e.g. Reddi et al. '18]
- Our take: adaptive methods escape saddles (in words: via isotropic noise), reach SOSPs

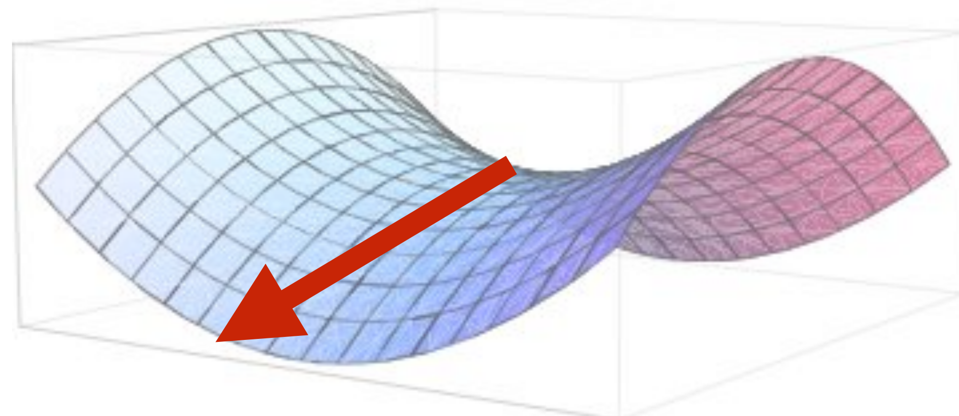
This paper:
The first *second*-order rates for adaptive methods

$$x_{t+1} \leftarrow x_t - \eta g_t$$

$$x_{t+1} \leftarrow x_t - \eta g_t$$

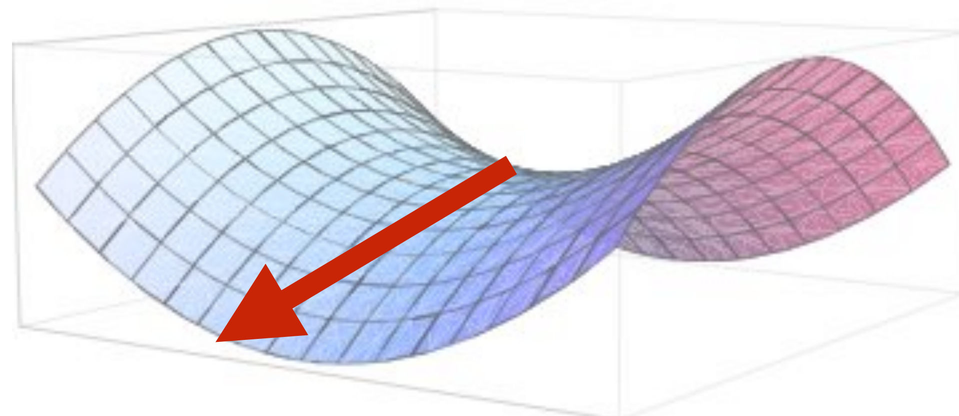


$$x_{t+1} \leftarrow x_t - \eta g_t + \xi_t$$



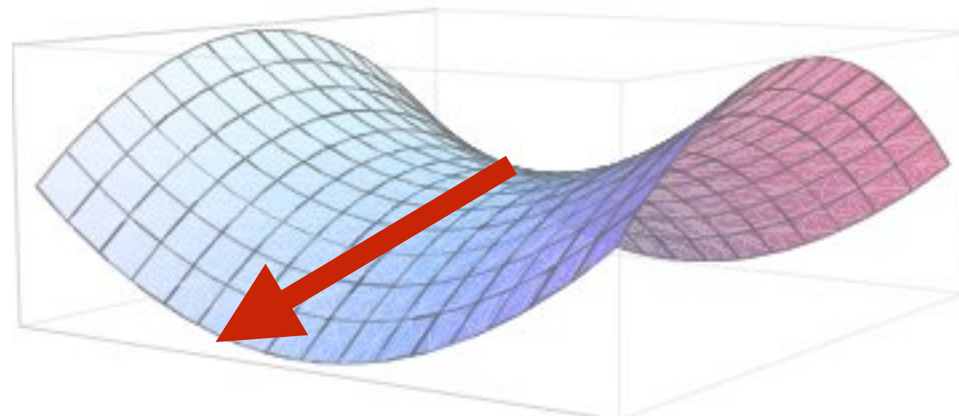
$$\mathbb{E}[\xi_t] = 0 \qquad \mathrm{Cov}(\xi_t) \propto I$$

$$x_{t+1} \leftarrow x_t - \eta g_t$$



$$\mathbb{E}[g_t] = 0 \qquad \mathrm{Cov}(g_t) = ???$$

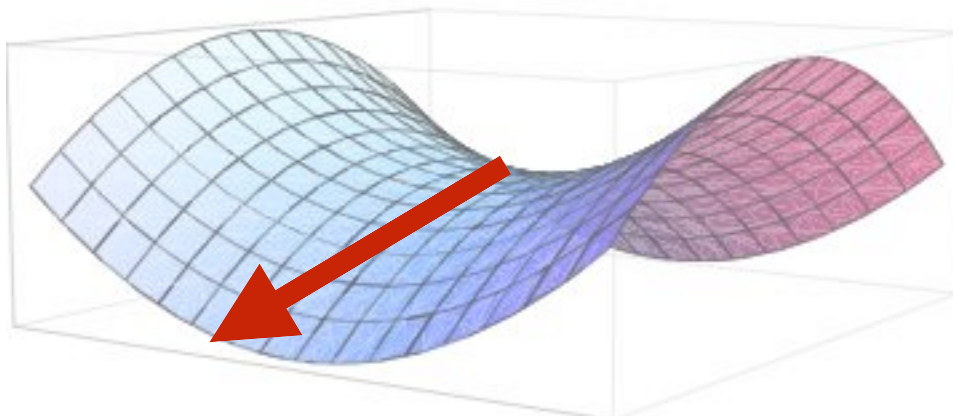
$$x_{t+1} \leftarrow x_t - \eta \mathbb{E}[g_t g_t^T]^{-1/2} g_t$$



$$\mathbb{E}[g_t] = 0 \qquad \mathrm{Cov}(g_t) = ???$$

$$\mathrm{Cov}(\mathbb{E}[g_t g_t^T]^{-1/2} g_t) = I$$

$$x_{t+1} \leftarrow x_t - \eta \mathbb{E}[g_t g_t^T]^{-1/2} g_t$$



$$\mathbb{E}[g_t] = 0 \qquad \mathrm{Cov}(g_t) = ???$$

$$\mathrm{Cov}(\mathbb{E}[g_t g_t^T]^{-1/2} g_t) = I$$

RMSProp

$$x_{t+1} \leftarrow x_t - \eta \hat{G}_t^{-1/2} g_t$$

$$x_{t+1} \leftarrow x_t - \eta \hat{G}_t^{-1/2} g_t$$

$$\hat{G}_t := \sum_{i=1}^{t} \beta^{t-i} g_i g_i^T$$

$$x_{t+1} \leftarrow x_t - \eta \hat{G}_t^{-1/2} g_t$$

$$\hat{G}_t := \sum_{i=1}^{t} \beta^{t-i} g_i g_i^T$$

$$\approx \mathbb{E}[g_t g_t^T] =: G_t$$

$$x_{t+1} \leftarrow x_t - \eta \hat{G}_t^{-1/2} g_t$$

$$\hat{G}_t := \sum_{i=1}^{t} \beta^{t-i} g_i g_i^T$$

$$\approx \mathbb{E}[g_t g_t^T] =: G_t$$

(**Theorem:** w.h.p. if β chosen correctly given η)

Theorem (informal):
RMSProp converges to a $(\tau, \tau^{1/2})$-stationary point in time $O(\tau^{-5})$.

Summary

- New approach: theory for general preconditioners

Summary

- New approach: theory for general preconditioners
- Also works for standard diagonal approx.

Summary

- New approach: theory for general preconditioners
- Also works for standard diagonal approx.

Concrete takeaways:

Summary

- New approach: theory for general preconditioners
- Also works for standard diagonal approx.

Concrete takeaways:

- How to set β as a function of stepsize η

Summary

- New approach: theory for general preconditioners
- Also works for standard diagonal approx.

Concrete takeaways:

- How to set β as a function of stepsize η
- How to set the ε in the RMSProp denominator: $(\mathbb{E}[g_t g_t^T]^{1/2} + \varepsilon I)^{-1}$

Summary

- New approach: theory for general preconditioners
- Also works for standard diagonal approx.

Concrete takeaways:

- How to set β as a function of stepsize η
- How to set the ε in the RMSProp denominator: $(\mathbb{E}[g_t g_t^T]^{1/2} + \varepsilon I)^{-1}$

Poster #98