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
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
\[ x_{t+1} \leftarrow x_t - \eta g_t \]
\[ x_{t+1} \leftarrow x_t - \eta g_t \]
\( x_{t+1} \leftarrow x_t - \eta g_t + \xi_t \)

\( \mathbb{E}[\xi_t] = 0 \quad \text{Cov}(\xi_t) \propto I \)
$\mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta g_t$

$\mathbb{E}[g_t] = 0 \quad \text{Cov}(g_t) = ???$
\[ x_{t+1} \leftarrow x_t - \eta \mathbb{E}[g_t g_t^T]^{-1/2} g_t \]

\[ \mathbb{E}[g_t] = 0 \quad \text{Cov}(g_t) = ??? \]

\[ \text{Cov}(\mathbb{E}[g_t g_t^T]^{-1/2} g_t) = I \]
\[ x_{t+1} \leftarrow x_t - \eta \mathbb{E}\left[ g_t g_t^T \right]^{-1/2} g_t \]

\[ \mathbb{E}[g_t] = 0 \quad \text{Cov}(g_t) = ??? \]

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

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

$$\hat{G}_t : = \sum_{i=1}^{t} \beta^{t-i} g_i g_i^T$$
\[ x_{t+1} \leftarrow x_t - \eta \hat{G}_t^{-1/2} g_t \]

\[
\hat{G}_t := \sum_{i=1}^{t} \beta^{t-i} g_i g_i^T \\
\approx \mathbb{E}[g_t g_t^T] =: G_t
\]
\[ \mathbf{x}_{t+1} \leftarrow \mathbf{x}_t - \eta \hat{G}_t^{-1/2} \mathbf{g}_t \]

\[
\hat{G}_t : = \sum_{i=1}^{t} \beta^{t-i} \mathbf{g}_i \mathbf{g}_i^T \\
\approx \mathbb{E}[\mathbf{g}_t \mathbf{g}_t^T] =: G_t
\]

(Theorem: w.h.p. if \( \beta \) chosen correctly given \( \eta \))
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 $\beta$ as a function of stepsize $\eta$
Summary

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

Concrete takeaways:
• How to set $\beta$ as a function of stepsizes $\eta$
• How to set the $\epsilon$ in the RMSProp denominator: $\left(\mathbb{E}[g_t g^T_t]^{1/2} + \epsilon I\right)^{-1}$
Summary

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

Concrete takeaways:
• How to set $\beta$ as a function of stepsize $\eta$
• How to set the $\epsilon$ in the RMSProp denominator: 
  \[ (\mathbb{E}[g_t g_t^T]^{1/2} + \epsilon I)^{-1} \]