Fully Stochastic Trust Region Algorithms Without Ratio Tests

Frank E. Curtis, Lehigh University

joint work with

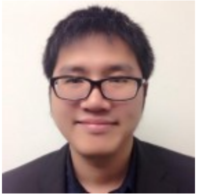
Katya Scheinberg, Cornell University
Rui Shi, Lehigh University

presented at

International Conference on Stochastic Programming
Trondheim, Norway

August 1, 2019

References

⋆ F. E. Curtis, K. Scheinberg, and R. Shi.
A Stochastic Trust Region Algorithm Based on Careful Step Normalization.
INFORMS Journal on Optimization, https://doi.org/10.1287/ijoo.2018.0010, 2019.

Outline

Outline

Motivation

First-order TRish

Second-order TRish

Summary

Ideals

Ideal features of optimization algorithms:

- good worst-case complexity / convergence rate
- function / variable scale invariance

This talk focuses on the importance of the latter.

Goal: Design stochastic optimization algorithms whose

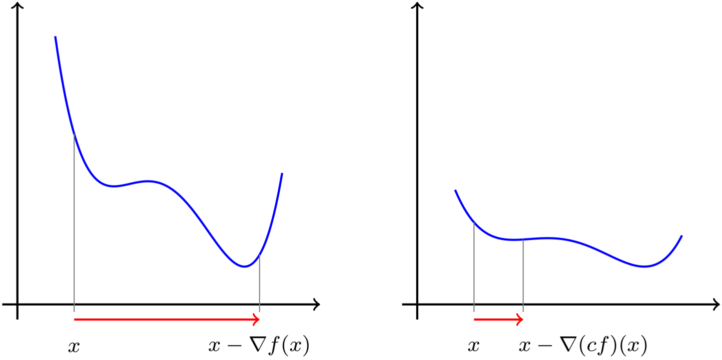
- theoretical performance is comparable to that of stochastic gradient (SG);*
- practical performance is more stable (in fully stochastic regime).

Ideas can also be used with variance reduction, second-order techniques, etc.

*Robbins and Monro (former Lehigh faculty member!) (1951)

Function scale independence

Gradient step has no natural scaling.



This is NOT handled by stepsize tuning!

Function / variable scale independence

Consider the minimization problem and gradient descent iteration:

$$\min_{x\in\mathbb{R}^n} \; f(x) \qquad \Longrightarrow \qquad x_{k+1} \leftarrow x_k - \nabla f(x_k).$$

Considering the equivalent problem

$$\min_{\hat{x}\in\mathbb{R}^n} \{\hat{f}(\hat{x}) \equiv cf(A\hat{x})\}, \;\; \text{where} \;\; (A,c) \in \mathbb{R}^{n\times n} \times \mathbb{R}_{>0} \;\; \text{with} \;\; A \succ 0$$

leads to the different gradient descent iteration

$$\begin{aligned} & & \hat{x}_{k+1} &\leftarrow \hat{x}_k - \nabla \hat{f}(\hat{x}_k) \\ & & &= \hat{x}_k - cA\nabla f(A\hat{x}_k) \\ \Longrightarrow & & A\hat{x}_{k+1} &\leftarrow A\hat{x}_k - cA^2\nabla f(A\hat{x}_k) \\ \Longrightarrow & & x_{k+1} &\leftarrow x_k - cA^2\nabla f(x_k). \end{aligned}$$

By contrast, Newton's method leads to the equivalent iterations

$$\begin{aligned} & & x_{k+1} &\leftarrow x_k - (\nabla^2 f(x_k))^{-1}\nabla f(x_k) \\ \Longleftrightarrow & & A\hat{x}_{k+1} &\leftarrow A\hat{x}_k - (\nabla^2 f(A\hat{x}_k))^{-1}\nabla f(A\hat{x}_k). \end{aligned}$$

Trust region methods

Trust region methods have proved to be effective for nonconvex optimization.

- The trust region constraint imposes scale on step length.
- However, these methods traditionally rely on a ratio test involving

$$\rho_k := \frac{\text{actual reduction}}{\text{predicted reduction}} \equiv \frac{f(x_k) - f(x_k + s_k)}{m_k(x_k) - m_k(x_k + s_k)}.$$

Even stochastic trust region methods rely on ρ_k estimates.

- Blanchet, Cartis, Menickelly, and Scheinberg (2016)
- Chen, Menickelly, and Scheinberg (2018)
- Wang and Yuan (2019)

These require approximate function evaluations (and complicated analyses).

Contributions

Stochastic trust region algorithms with

- no ratio tests;
- no function evaluation estimates;
- good behavior in fully stochastic regime;
- convergence theory comparable to that for SG;
- practical behavior more stable than SG;
- first- and second-order variants;
- exact subproblem solutions not needed.

Outline

Problem description

Consider the stochastic optimization problem

$$\min_{x \in \mathbb{R}^n} f(x), \quad \text{where} \quad f(x) = \mathbb{E}_\xi[F(x, \xi)]. \tag{1}$$

A special case is the finite-sum problem

$$f(x) = \frac{1}{N} \sum_{i=1}^{N} f_i(x); \tag{2}$$

such an objective might also arise in a sample average approximation of (1).

The stochastic gradient method (SG) uses stochastic gradients defined by

$$g_k = \nabla_x F(x_k, \xi_k) \quad \text{for (1)}$$

where ξ_k is a realization of the random variable ξ, or

$$g_k = \nabla_x f_{i_k}(x_k) \quad \text{for (2)}$$

where i_k is chosen randomly as an index in $\{1, \ldots, N\}$. Simply: $g_k \approx \nabla f(x_k)$.

First-order trust region subproblem

Consider the trust region subproblem

$$\min_{s \in \mathbb{R}^n} \; g_k^T s \quad \text{s.t.} \quad \|s\|_2 \leq \alpha_k. \tag{3}$$

Solution:

$$s_k = -\alpha_k g_k / \|g_k\|_2.$$

Using this formula for s_k, the algorithm might not be convergent!

Related work:

- Normalized gradient descent; Hazan, Levy, and Shalev-Shwartz (2015)
- Batch normalization; Ioffe and Szegedy (2015)

We provide convergence guarantees under weaker assumptions.

Example

Suppose that, at a point $x_k \in \mathbb{R}$, one has $\nabla f(x_k) = 1$ and

$$g_k = \begin{cases} 6 \text{ with probability } \frac{1}{3} \\ -\frac{3}{2} \text{ with probability } \frac{2}{3}. \end{cases}$$

However, this means that the normalized stochastic gradient satisfies

$$\frac{g_k}{\|g_k\|_2} = \begin{cases} 1 \text{ with probability } \frac{1}{3} \\ -1 \text{ with probability } \frac{2}{3}, \end{cases}$$

from which it follows that $s_k = -\alpha_k g_k / \|g_k\|_2$ is twice more likely to be a direction of ascent for f at x_k than it is to be a direction of descent for f at x_k.

First-order TRish

Central idea:

- Only take normalized step when norm is in certain range.
- Take constant multiple of stochastic gradient step in other cases.

Algorithm 1 Trust-region-ish (TRish) algorithm

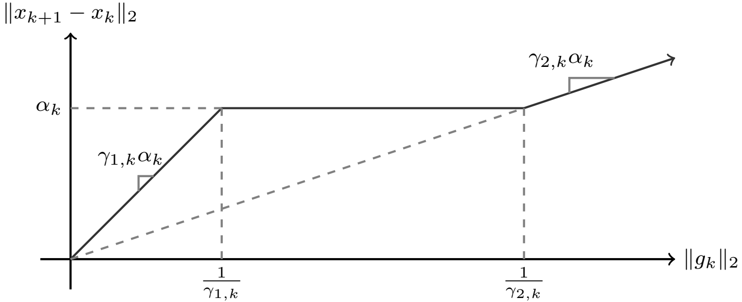
---

choose positive stepsizes $\{\alpha_k\}$
choose positive sequences $\{\gamma_{1,k}\}$ and $\{\gamma_{2,k}\}$ with $\gamma_{1,k} > \gamma_{2,k} > 0$ for all $k \in \mathbb{N}$
for all $k \in \mathbb{N} := \{1, 2, \dots\}$ **do**
    generate a stochastic gradient $g_k \approx \nabla f(x_k)$
    set

$$x_{k+1} \leftarrow x_k - \begin{cases} \gamma_{1,k}\alpha_k g_k & \text{if } \|g_k\|_2 \in [0, \frac{1}{\gamma_{1,k}}) \\ \alpha_k g_k / \|g_k\|_2 & \text{if } \|g_k\|_2 \in [\frac{1}{\gamma_{1,k}}, \frac{1}{\gamma_{2,k}}] \\ \gamma_{2,k}\alpha_k g_k & \text{if } \|g_k\|_2 \in (\frac{1}{\gamma_{2,k}}, \infty) \end{cases}$$

end for

Illustration of iterate displacement



Assumption

Our main assumption is exactly the same as for standard SG.

Assumption 1

The objective function $f : \mathbb{R}^n \to \mathbb{R}$ *satisfies:*

- f *is continuously differentiable*
- f *is bounded below by* $f_* = \inf_{x \in \mathbb{R}^n} f(x) \in \mathbb{R}$
- *there exists* $L \in \mathbb{R}$ *(independent of* k*) such that*

$$f(x) \leq f(\bar{x}) + \nabla f(\bar{x})^T (x - \bar{x}) + \tfrac{1}{2} L \|x - \bar{x}\|_2^2 \ \ \textit{for all} \ \ (x, \bar{x}) \in \mathbb{R}^n \times \mathbb{R}^n$$

In addition, for all $k \in \mathbb{N}$*, the stochastic gradient* g_k *satisfies*

- $\mathbb{E}_k[g_k] = \nabla f(x_k)$
- *there exists* $(M_1, M_2) \in (0, \infty)^2$ *(independent of* k*) such that*

$$\mathbb{E}_k[\|g_k\|_2^2] \leq M_1 + M_2 \|\nabla f(x_k)\|_2^2.$$

* $\mathbb{E}_k[\cdot]$ and $\mathbb{P}_k[\cdot]$ are expectation and probability conditioned on history to x_k.

Fundamental lemmas

Lemma 1

Under Assumption 1, for all $k \in \mathbb{N}$, one finds

$$\begin{aligned} &\mathbb{E}_k[f(x_{k+1})] - f(x_k) \\ \leq\ & \underbrace{-\gamma_{1,k}\alpha_k(1 - \tfrac{1}{2}\gamma_{1,k}LM_2\alpha_k)\|\nabla f(x_k)\|_2^2}_{\text{deterministic decrease}} \\ & \underbrace{+(\gamma_{1,k} - \gamma_{2,k})\alpha_k\mathbb{P}_k[E_k]\mathbb{E}_k[\nabla f(x_k)^T g_k | E_k]}_{\text{conditional increase}}\ \underbrace{+\tfrac{1}{2}\gamma_{1,k}^2 LM_1\alpha_k^2}_{\text{increase from noise}}\ . \end{aligned}$$

where E_k is the event that $\nabla f(x_k)^T g_k \geq 0$.

Lemma 2

Under Assumption 1, for all $k \in \mathbb{N}$, one finds

$$\mathbb{P}_k[E_k]\mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] \leq h_1 + h_2\|\nabla f(x_k)\|_2^2$$

for any (h_1, h_2) such that $h_1 \geq \frac{1}{2}\sqrt{M_1}$ and $h_2 \geq \frac{1}{2}\sqrt{M_1} + \sqrt{M_2}$.

Example result for nonconvex f

Theorem 3 (Nonconvex f, fixed parameters and stepsize)

For all $k \in \mathbb{N}$, suppose $(\gamma_{1,k}, \gamma_{2,k}) = (\gamma_1, \gamma_2)$ with $\frac{\gamma_1}{\gamma_2} < \frac{h_2}{h_2 - 1}$ and $\alpha_k = \alpha$ with

$$0 < \alpha \leq \frac{\gamma_1 - h_2(\gamma_1 - \gamma_2)}{\gamma_1 L M_2}.$$

Then, there exists (θ_1, θ_2) (for which we provide formulas) such that

$$\mathbb{E}\left[\frac{1}{K}\sum_{k=1}^{K} \|\nabla f(x_k)\|_2^2\right] \leq \frac{\theta_2}{\alpha\theta_1} + \frac{f(x_1) - f_*}{K\alpha\theta_2} \xrightarrow{K \to \infty} \frac{\theta_2}{\alpha\theta_1}.$$

Also, for diminishing stepsizes, expected average gradient vanishes, implying

$$\liminf_{k \to \infty} \mathbb{E}[\|\nabla f(x_k)\|_2^2] = 0.$$

Example result under the Polyak-Łojasiewicz (P-L) condition

Theorem 4 (P-L condition, diminishing stepsizes)

Suppose f satisfies the P-L condition and, for all $k \in \mathbb{N}$,

$$\gamma_{1,k} = \gamma_1 > 0 \quad \textit{and} \quad \gamma_{2,k} = \gamma_1(1 - \tfrac{1}{2}\eta\alpha_k) \quad \textit{for some} \quad \eta \in (0,1),$$

and, for appropriately chosen (a,b) (see paper),

$$\alpha_k = \frac{a}{b+k} \quad \textit{with} \quad \alpha_1 \in \left(0, \min\left\{\frac{1}{\eta}, \frac{1}{\eta h_2 + \gamma_1 L M_2}\right\}\right].$$

Then, for all $k \in \mathbb{N}$, one finds

$$\mathbb{E}[f(x_k)] - f_* \leq \frac{\phi}{b+k},$$

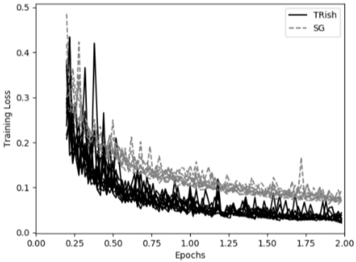
where

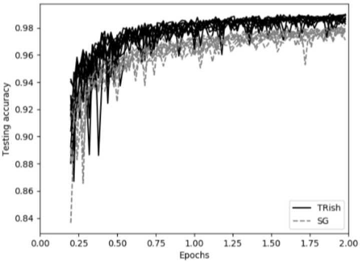
$$\phi := \max\left\{\frac{a^2\delta}{ac\gamma_1 - 1}, (b+1)(f(x_1) - f_*)\right\} > 0$$

$$\textit{and} \quad \delta := \tfrac{1}{2}\gamma_1(\eta h_1 + \gamma_1 L M_1) > 0.$$

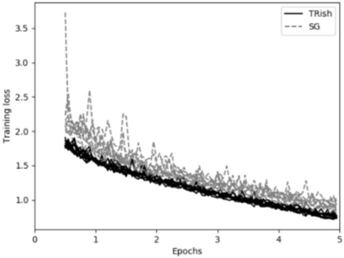
Also, with variance reduction, linear rate of convergence for $\alpha_k = \alpha$ small.

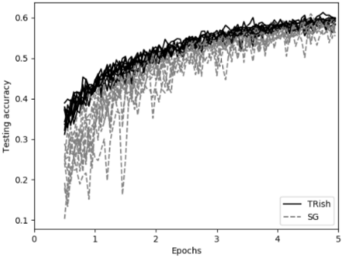
DNN training on mnist





DNN training on cifar-10





Outline

Primary challenge

Introducing stochastic second-order information is complicated here!

- Recall Lemma 2, which gave

$$\mathbb{P}_k[E_k]\mathbb{E}_k[\nabla f(x_k)^T g_k | E_k] \leq h_1 + h_2\|\nabla f(x_k)\|_2^2$$

- A similar bound on the conditional expectation of

$$\nabla f(x_k)^T s_k$$

is no longer straightforward with s_k influenced by $H_k \approx \nabla^2 f(x_k)$.

Second-order TRish

Algorithm 2 Second-order TRish algorithm

1: choose positive stepsizes $\{\alpha_k\}$
2: choose positive sequences $\{\gamma_{1,k}\}$ and $\{\gamma_{2,k}\}$ with $\gamma_{1,k} > \gamma_{2,k} > 0$ for all $k \in \mathbb{N}$
3: **for all** $k \in \mathbb{N} := \{1, 2, \dots\}$ **do**
4: generate a stochastic gradient $g_k \approx \nabla f(x_k)$
5: generate a stochastic Hessian $H_k \approx \nabla^2 f(x_k)$
6: **if** $\|g_k\|_2 \in [\frac{1}{\gamma_{1,k}}, \frac{1}{\gamma_{2,k}}]$, then approximately solve

$$\min_{s \in \mathbb{R}^n} \; g_k^T s + \tfrac{1}{2} s^T H_k s \quad \text{s.t.} \quad \|s_k\|_2 \leq \alpha_k$$

7: **else** approximately solve (with $\gamma_k = \gamma_{1,k}$ or $\gamma_k = \gamma_{2,k}$ depending on $\|g_k\|_2$)

$$\min_{s \in \mathbb{R}^n} \; g_k^T s + \tfrac{1}{2} s^T H_k s \quad \text{s.t.} \quad \|s_k\|_2 \leq \gamma_k \alpha_k \|g_k\|_2$$

8: **end for**

Requiring only Cauchy decrease and with standard assumptions, convergence guarantees of all the same types as for first-order TRish. Numerics forthcoming.

Outline

Motivation

First-order TRish

Second-order TRish

Summary

Contributions

Stochastic trust region algorithms with

- no ratio tests;
- no function evaluation estimates;
- good behavior in fully stochastic regime;
- convergence theory comparable to that for SG;
- practical behavior more stable than SG;
- first- and second-order variants;
- exact subproblem solutions not needed.

⋆ F. E. Curtis, K. Scheinberg, and R. Shi.
A Stochastic Trust Region Algorithm Based on Careful Step Normalization.
INFORMS Journal on Optimization, https://doi.org/10.1287/ijoo.2018.0010, 2019.