Stochastic Gradient Descent with Variance Reduction

Rie Johnson, Tong Zhang
Presenter: Jiawen Yao

March 17, 2015

Outline

Outline

Preliminaries

Recall a few definitions from convex analysis.

Definition 1. A function $f(x)$ is a L-Lipschitz continuous function if

$$\|f(x_1) - f(x_2)\| \leq L\|x_1 - x_2\| \tag{1}$$

for all $x_1, x_2 \in dom(f)$

Definition 2. A convex function $f(x)$ is β-strong convex if there exists a constant $\beta > 0$ and for any $\alpha \in [0, 1]$, it holds:

$$f(\alpha x_1 + (1-\alpha)x_2) \leq \alpha f(x_1) + (1-\alpha)f(x_2) - \frac{1}{2}\alpha(1-\alpha)\beta\|x_1 - x_2\|^2 \tag{2}$$

Preliminaries

When $f(x)$ is differentiable, the strong convexity is equivalent to

$$f(x_1) \geq f(x_2) + < \nabla f(x_2), x_1 - x_2 > + \frac{\beta}{2}\|x_1 - x_2\|^2 \tag{3}$$

Typically, we use the standard Euclidean norm to define Lipschitz and strong convex functions.

Minimizing finite average of convex functions

Let $\psi_1, ..., \psi_n$ be a sequence of vector functions from $\mathbb{R}^d$ to $\mathbb{R}$.

$$minP(\omega), P(\omega) = \frac{1}{n}\sum_{i=1}^{n}\psi_i(\omega) \tag{4}$$

assumptions:

- each $\psi_i(\omega)$ is convex and differentiable on dom(R)
- each $\psi_i(\omega)$ is smooth with Lipschitz constant L

$$\|\nabla\psi_i(\omega) - \nabla\psi_i(\omega')\| \leq L\|\omega - \omega'\| \tag{5}$$

- $P(\omega)$ is strongly convex

$$P(\omega) \geq P(\omega') + \frac{\gamma}{2}\|\omega - \omega'\|^2 + \nabla P(\omega')^{\intercal}(\omega - \omega') \tag{6}$$

Gradient Descent

$$\omega^{(t)} = \omega^{(t-1)} - \eta_t \nabla P(\omega^{(t-1)}) = \omega^{(t-1)} - \frac{\eta_t}{n} \sum_{i=1}^{n} \nabla \psi_i(\omega^{(t-1)}) \quad (7)$$

Stochastic Gradient Descent

Draw i_t randomly from $\{1, ..., n\}$

$$\omega^{(t)} = \omega^{(t-1)} - \eta_t \nabla \psi_{i_t}(\omega^{(t-1)}) \quad (8)$$

SGD

A more general version of SGD is the following

$$\omega^{(t)} = \omega^{(t-1)} - \eta_t g_t(\omega^{(t-1)}, \xi_t) \tag{9}$$

where ξ_t is a random variable that may depend on $\omega^{(t-1)}$, the expectation $\mathbb{E}[g_t(\omega^{(t-1)}, \xi_t)|\omega^{(t-1)}] = \nabla P(\omega^{(t-1)})$

Variance

For general convex optimization, stochastic gradient descent methods can obtain an $\mathcal{O}(1/\sqrt{T})$ convergence rate in expectation.
Randomness introduces large variance if $g_t(\omega^{(t-1)}, \xi_t)$ is very large, it will slow down the convergence.

Outline

Stochastic Average Gradient

SAG method (Le Roux, Schmidt, Bach 2012)

$$\omega_t = \omega_{t-1} - \frac{\eta}{n} \sum_{i=1}^{n} g_/ t^{(i)} \tag{10}$$

where

$$g_t^{(i)} = \begin{cases} \nabla \psi_i(\omega_t), & \text{if } i = i_t \\ g_{t-1}^{(i)}, & \text{otherwise} \end{cases} \tag{11}$$

It needs to store all gradient, not practical for some cases

Outline

SVRG

Motivation

- Reduce the variance
- Stochastic gradient descent has slow convergence asymptotically due to the inherent variance.
- SAG needs to store all gradients

Contribution

- No need to store the intermediate gradients
- The same convergence rate as SAG can obtain
- Under mild assumptions, even work on nonconvex cases

Stochastic variance reduced gradient (SVRG)

- SVRG (Johnson & Zhang, NIPS 2013)
  - update form

$$\omega^{(t)} = \omega^{(t-1)} - \eta_t(\nabla\psi_{i_t}(\omega^{(t-1)}) - \nabla\psi_{i_t}(\tilde{\omega}) + \nabla P(\tilde{\omega})) \qquad (12)$$

  - update $\tilde{\omega}$ periodically (every m SGD iterations)

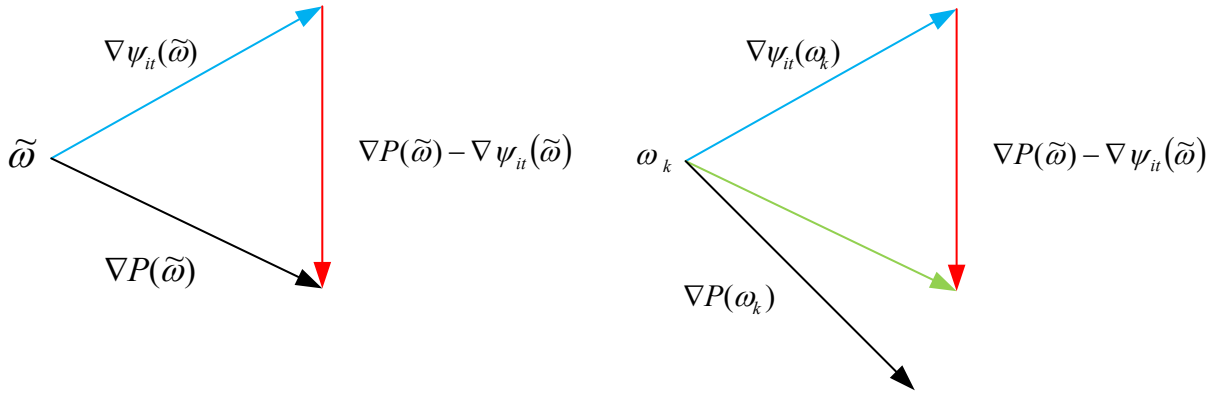

Figure: Intuition of variance reduction

Procedure SVRG

input: update frequency m and learning rate η

initialization: $\tilde{\omega}_0$

for s=1,2,... **do**

$\tilde{\omega} = \tilde{\omega}_{s-1}$

$\tilde{\mu} = \nabla P(\tilde{\omega}) = \frac{1}{n}\sum_{i=1}^{n} \nabla \psi_i(\tilde{\omega})$

$\omega_0 = \tilde{\omega}$

Randomly pick $i_t \in \{1, ..., n\}$ and update weight, repeat m times

$\omega_t = \omega_{t-1} - \eta_t(\nabla \psi_{i_t}(\omega_{t-1}) - \nabla \psi_{i_t}(\tilde{\omega}) + \nabla P(\tilde{\omega}))$

option I: set $\tilde{\omega}_s = \omega_m$

option II: set $\tilde{\omega}_s = \omega_t$ for randomly chosen $t \in \{0, ..., m-1\}$

end for

Convergence for SVRG

Theorem

Consider SVRG with option II. Assume that all $\psi_i(\omega)$ are convex and smooth, $P(\omega)$ is strongly convex. Let $\omega_* = \text{argmin}_\omega P(\omega)$. Assume that m is sufficiently large so that

$$\alpha = \frac{1}{\gamma\eta(1-2L\eta)m} + \frac{2L\eta}{1-2L\eta} < 1$$

then we have geometric convergence in expectations for SVRG

$$\mathbb{E}P(\tilde{\omega}_s) \leq \mathbb{E}P(\tilde{\omega}_*) + \alpha^s[P(\tilde{\omega}_0) - P(\omega_*)]$$

Proof

Given any i, consider

$$g_i(\omega) = \psi_i(\omega) - \psi_i(\omega_*) - \nabla\psi_i(\omega_*)^T(\omega - \omega_*) \tag{13}$$

where $g_i(\omega_*) = argmin_\omega g_i(\omega)$ and $\nabla g_i(\omega_*) = 0$

$$\begin{aligned} 0 = g_i(\omega_*) &\leq min_\eta[g_i(\omega - \eta\nabla g_i(\omega))] \\ &\leq min_\eta[g_i(\omega) - \eta\|\nabla g_i(\omega)\|^2 + 0.5L\eta^2\|\nabla g_i(\omega)\|^2] \end{aligned} \tag{14}$$

Here it uses a well-known inequality for a function with 1/L-Lipschitz continuous gradient

$$f(x) - f(y) - \langle\nabla f(y), x - y\rangle \leq \frac{L}{2}\|x - y\|^2$$

Proof

From (14), we can get $\eta = 1/L$, then

$$0 = g_i(\omega_*) \leq g_i(\omega) - \frac{1}{2L}\|\nabla g_i(\omega)\|^2 \tag{15}$$

It can be rewrite as

$$\|\nabla g_i(\omega)\|^2 \leq 2L g_i(\omega) \tag{16}$$

using the definition of $g_i(\omega)$ and $\nabla g_i(\omega) = \nabla \psi_i(\omega) - \nabla \psi_i(\omega_*)$, the (16) will be

$$\|\nabla \psi_i(\omega) - \nabla \psi_i(\omega_*)\|^2 \leq 2L[\psi_i(\omega) - \psi_i(\omega_*) - \nabla \psi_i(\omega_*)^T(\omega - \omega_*)] \tag{17}$$

Proof

By summing the inequality (17) over $i = \{1, ..., n\}$, the fact that $P(\omega) = \frac{1}{n}\sum_{i=1}^{n} \nabla\psi_i(\omega)$ and $\nabla P(\omega_*) = 0$, we can get

$$n^{-1}\sum_{i=1}^{n} \|\nabla\psi_i(\omega) - \nabla\psi_i(\omega_*)\|^2 \leq 2L[P(\omega) - P(\omega_*)] \tag{18}$$

Use $\tilde{\mu} = \nabla P(\tilde{\omega})$ and let $v_t = \nabla\psi_{i_t}(\omega_{t-1}) - \nabla\psi_{i_t}(\tilde{\omega}) + \tilde{\mu}$, v_t is the approximate gradient of SVRG.

Proof

With respect to i_t, expectation can be obtained as

$$\begin{aligned}
\mathbb{E}\|v_t\|^2 &= \mathbb{E}\|\nabla\psi_{i_t}(\omega_{t-1}) - \nabla\psi_{i_t}(\tilde{\omega}) + \tilde{\mu}\|^2 \\
&\leq 2\mathbb{E}\|\nabla\psi_{i_t}(\omega_{t-1}) - \nabla\psi_{i_t}(\omega_*)\|^2 + 2\mathbb{E}\|[\nabla\psi_{i_t}(\tilde{\omega}) - \nabla\psi_{i_t}(\omega_*)] - \tilde{\mu}\|^2 \\
&= 2\mathbb{E}\|\nabla\psi_{i_t}(\omega_{t-1}) - \nabla\psi_{i_t}(\omega_*)\|^2 + 2\mathbb{E}\|[\nabla\psi_{i_t}(\tilde{\omega}) - \nabla\psi_{i_t}(\omega_*)] \\
&- \mathbb{E}[\nabla\psi_{i_t}(\tilde{\omega}) - \nabla\psi_{i_t}(\omega_*)]\|^2 \\
&\leq 2\mathbb{E}\|\nabla\psi_{i_t}(\omega_{t-1}) - \nabla\psi_{i_t}(\omega_*)\|^2 + 2\mathbb{E}\|\nabla\psi_{i_t}(\tilde{\omega}) - \nabla\psi_{i_t}(\omega_*)\|^2 \\
&\leq 4L[P(\omega_{t-1}) - P(\omega_*) + P(\tilde{\omega}) - P(\omega_*)] \qquad (19)
\end{aligned}$$

The first inequality uses $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$
The second inequality uses $\mathbb{E}\|X - \mathbb{E}X\|^2 \leq \mathbb{E}\|X\|^2$
The third one uses (18)

Proof

The update form of SVRG is $\omega_t = \omega_{t-1} - \eta v_t$, conditioned on ω_{t-1}

$$\begin{aligned}\mathbb{E}\|\omega_t - \omega_*\|^2 &= \mathbb{E}\|\omega_{t-1} - \omega_* - \eta v_t\|^2 \\ &= \|\omega_{t-1} - \omega_*\|^2 - 2\eta(\omega_{t-1} - \omega_*)^\intercal \mathbb{E} v_t + \eta^2 \mathbb{E}\|v_t\|^2\end{aligned}$$

Here $\mathbb{E}v_t = \mathbb{E}[\nabla\psi_{i_t}(\omega_{t-1}) - \nabla\psi_{i_t}(\tilde{\omega}) + \tilde{\mu}] = \nabla P(\omega_{t-1})$
Using (19) then we can get

$$\begin{aligned}&\mathbb{E}\|\omega_t - \omega_*\|^2 \\ \leq&\|\omega_{t-1} - \omega_*\|^2 - 2\eta(\omega_{t-1} - \omega_*)^\intercal \nabla P(\omega_{t-1}) \\ &+ 4L\eta^2[P(\omega_{t-1}) - P(\omega_*) + P(\tilde{\omega}) - P(\omega_*)] \qquad (20)\end{aligned}$$

By convexity of $P(\omega)$ that

$$-(\omega_{t-1} - \omega_*)^\intercal \nabla P(\omega_{t-1}) \leq P(\omega_*) - P(\omega_{t-1}) \qquad (21)$$

Proof

$$
\begin{aligned}
&\mathbb{E}\|\omega_t - \omega_*\|^2 \\
\leq &\|\omega_{t-1} - \omega_*\|^2 - 2\eta[P(\omega_*) - P(\omega_{t-1})] \\
&+ 4L\eta^2[P(\omega_{t-1}) - P(\omega_*) + P(\tilde{\omega}) - P(\omega_*)] \\
= &\|\omega_{t-1} - \omega_*\|^2 - 2\eta(1 - 2L\eta)[P(\omega_{t-1}) - P(\omega_*)] + 4L\eta^2[P(\tilde{\omega}) - P(\omega_*)]
\end{aligned} \tag{22}
$$

In each fixed stage s, $\tilde{\omega} = \tilde{\omega}_{s-1}$ and $\tilde{\omega}_s$ is selected after all updates have completed. By summing the inequality over $t = 1, ..., m$, taking expectation with all the history

$$
\begin{aligned}
&\mathbb{E}\|\omega_m - \omega_*\|^2 + 2\eta(1 - 2L\eta)m\mathbb{E}[P(\tilde{\omega}_s) - P(\omega_*)] \\
\leq &\mathbb{E}\|\tilde{\omega} - \omega_*\|^2 + 4Lm\eta^2\mathbb{E}[P(\tilde{\omega}) - P(\omega_*)] \\
\leq &\frac{2}{\gamma}\mathbb{E}[P(\tilde{\omega}) - P(\omega_*)] + 4Lm\eta^2\mathbb{E}[P(\tilde{\omega}) - P(\omega_*)]
\end{aligned} \tag{23}
$$

Proof

From above inequality, we can have

$$\begin{aligned} & 2\eta(1-2L\eta)m\mathbb{E}[P(\tilde{\omega}_s)-P(\omega_*)] \\ & \leq \frac{2}{\gamma}\mathbb{E}[P(\tilde{\omega})-P(\omega_*)]+4Lm\eta^2\mathbb{E}[P(\tilde{\omega})-P(\omega_*)] \end{aligned} \tag{24}$$

which can be also rewrite as

$$\mathbb{E}[P(\tilde{\omega}_s)-P(\tilde{\omega}_*)] \leq \alpha\mathbb{E}[P(\tilde{\omega}_{s-1})-P(\omega_*)] \tag{25}$$

where

$$\alpha = \frac{1}{\gamma\eta(1-2L\eta)m}+\frac{2L\eta}{1-2L\eta} \tag{26}$$

Proof

From (25), we can get the desired bound in the Theorem

$$\mathbb{E}[P(\tilde{\omega}_s) - P(\tilde{\omega}_*)] \leq \alpha^s \mathbb{E}[P(\tilde{\omega}_0) - P(\omega_*)] \tag{27}$$

The bound in Theorem 1 is comparable to Le Roux et al. [2012] and Shalev-Shwartz and Zhang [2012].
The convergence rate of SVRG is $\mathcal{O}(1/T)$ which improves the standard SGD convergence rate of $\mathcal{O}(1/\sqrt{T})$

Experiments

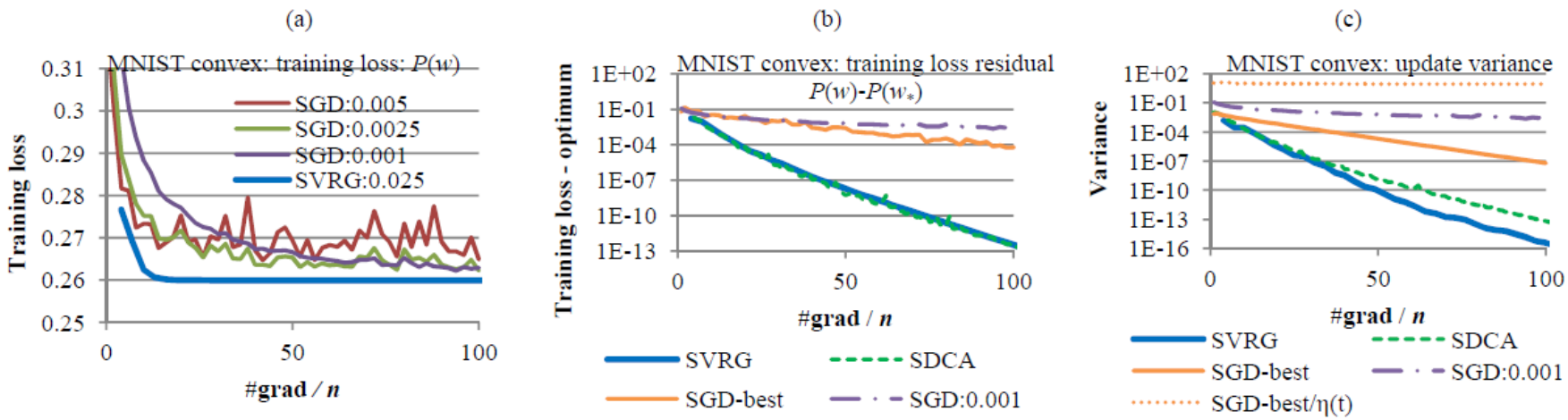

Figure: (a) Training loss comparison with SGD with fixed learning rates. (b) Training loss residual $P(\omega) - P(\omega_*)$ (c) Variance of weight update

It is hard to find a good η for SGD. Use a single relatively large value of η, SVRG smoothly goes down faster than SGD.

Experiments

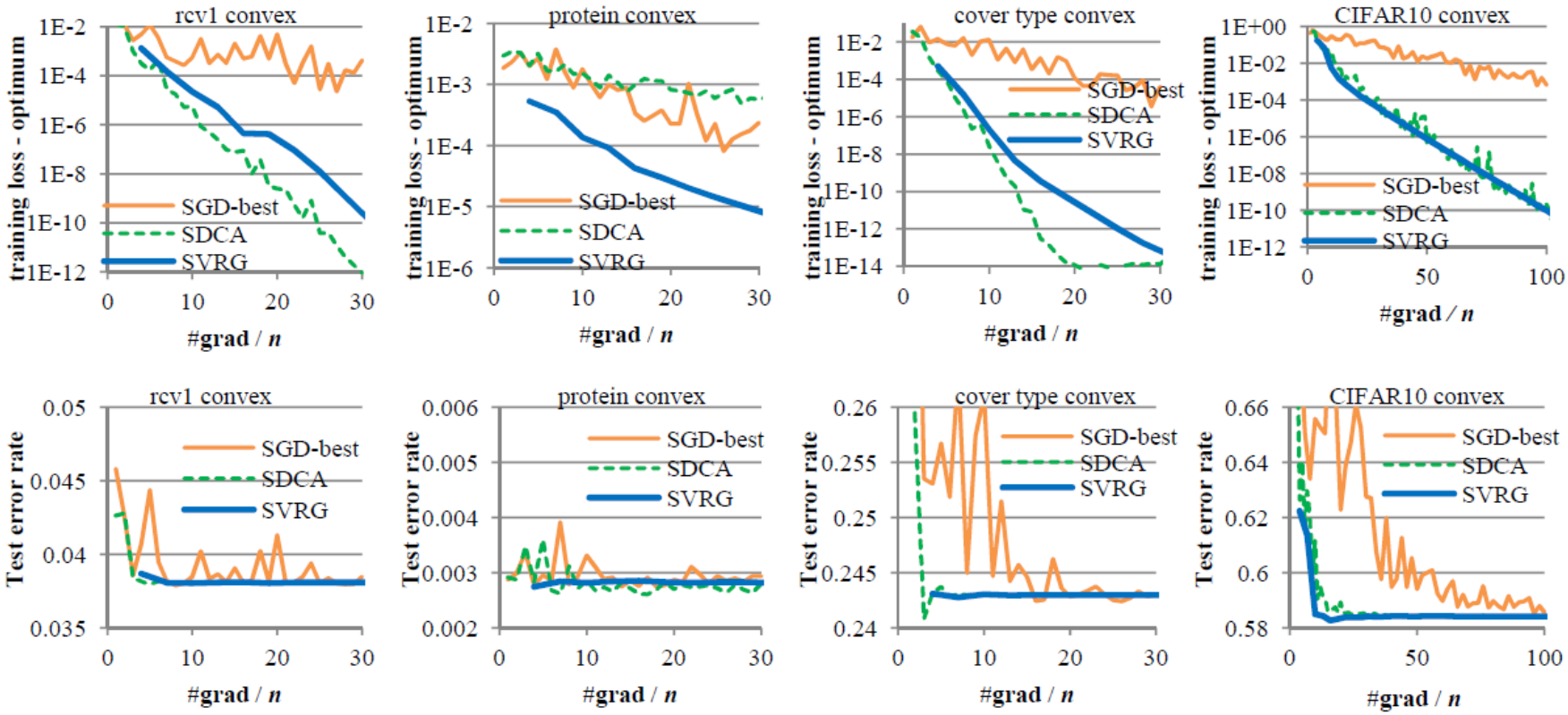

Figure: More convex-case results. Loss residual $P(\omega) - P(\omega_*)$ (top) and test error rates (down)

SVRG is competitive with SDCA and better than the best-tuned SGD.

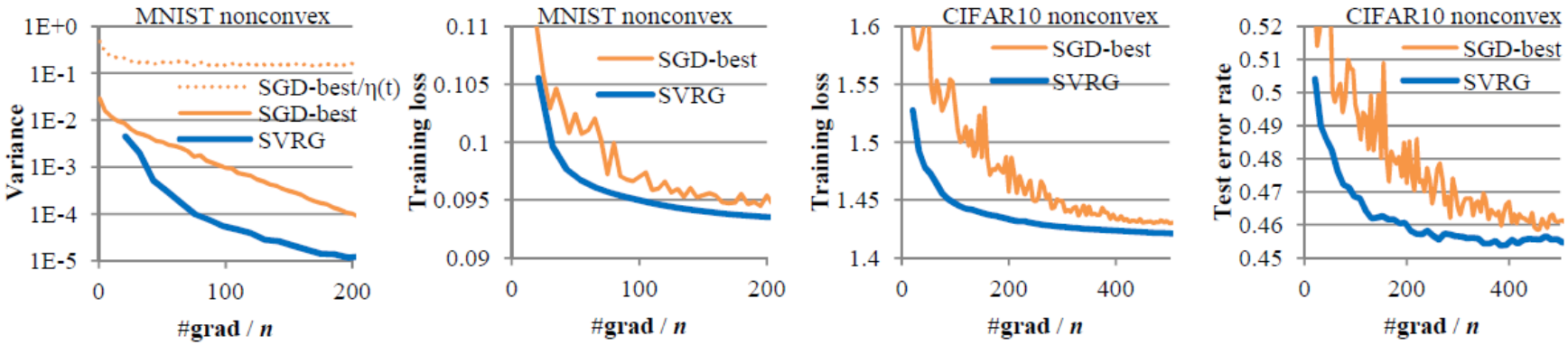

Figure: Neural net results (nonconvex)

For nonconvex problems, it is useful to start with an initial vector $\tilde{\omega}_0$ that is close to a local minimum. Results show that SVRG reduces the variance and smoothly converges faster than the best-tuned SGD.

Conclusion

- For smooth and strongly convex functions, we prove SVRG enjoys the same fast convergence rate as SAG
- Unlike SAG, no requirement of the storage of gradients
- Unlike SAG, it is more easily applicable to complex problems

Thank you!