Introduction to stochastic optimization

Anton Rodomanov
Research Fellow at Samsung-HSE Laboratory

28 August 2018
DeepBayes summer school

Part 1: General stochastic optimization

Stochastic optimization

Problem: $\min_{x\in\mathbb{R}^n} f(x)$, where f is a differentiable function.

Main assumption: cannot compute $f(x)$, $\nabla f(x)$ etc. exactly, but we have a stochastic oracle.

Stochastic oracle (SO): Given $x \in \mathbb{R}^n$, it returns a stochastic gradient (SG) $g(x)$:

- $g(x)$ is a random vector in $\mathbb{R}^n$ such that $\mathbb{E}g(x) = \nabla f(x)$ (plus some assumptions on the fluctuations).

Goal: A method for solving the problem given the SO.

Example 1: Stochastic programming

Let ξ be a random variable supported on $\Omega \subseteq \mathbb{R}^d$ and distributed according to a probability measure P. For each $\omega \in \Omega$, let $f_\omega : \mathbb{R}^n \to \mathbb{R}$ be a simple differentiable function, and let

$$f(x) := \mathbb{E} f_\xi(x) = \int_\Omega f_\omega(x) dP(\omega).$$

Main problems:

1. The distribution P may be unknown (machine learning).
2. Even if P is known, to find a ε-approximation of $f(x)$, one needs $O(\varepsilon^{-d})$ computations of $f_\omega(x)$.

Under mild assumptions, $\nabla f(x) = \mathbb{E} \nabla f_\xi(x) = \int_\Omega \nabla f_\omega(x) dP(\omega)$.

Main assumption: It is possible to sample from P efficiently.

SO: Given $x \in \mathbb{R}^n$, generate $\xi \sim P$ and return $g(x) := \nabla f_\xi(x)$.

Example 2: Finite sums

Let $f_1, \ldots, f_m : \mathbb{R}^n \to \mathbb{R}$ be <u>simple</u> differentiable functions, and let

$$f(x) := \frac{1}{m} \sum_{i=1}^{m} f_i(x).$$

Applications: Machine learning with a finite data set.

Main problem: To compute $f(x)$ or $\nabla f(x)$, we need $O(m)$ operations.

SO: Given $x \in \mathbb{R}^n$, generate $i_0 \sim \mathrm{Unif}\{1, \ldots, m\}$, and return $g(x) := \nabla f_{i_0}(x)$.

Complexity: $O(1)$, not depending on m.

Stochastic optimization: goals and complexity

Problem: $f^* := \min_{x\in\mathbb{R}^n} f(x)$, where f is given by an SO.

Goal: Given $\varepsilon > 0$, find a random $\bar{x} \in \mathbb{R}^n$:

- $\mathbb{E}f(\bar{x}) - f^* \leq \varepsilon$ (convex optimization).
- $\mathbb{E}\|\nabla f(\bar{x})\|^2 \leq \varepsilon$ (non-convex optimization).

Complexity measure: Number of calls to the SO.

Main result: $O(\varepsilon^{-2})$ complexity.

NB: We may approximately minimize f without even computing $f(x)$.

NB 2: $O(\varepsilon^{-2})$ is the complexity of Monte-Carlo ε-approximation of $f(x)$ for a <u>single</u> x.
The above $O(\varepsilon^{-2})$ is the complexity of the whole optimization process!

Remark: Same results with high probability under some regularity assumptions.

Stochastic gradient method (SGD) for non-smooth convex optimization

Problem: $f^* := \min_{x \in \mathbb{R}^n} f(x)$, where f is convex.
NB: f may be non-smooth, so instead of gradients we work with subgradients.

Method: Fix $x_0 \in \mathbb{R}^n$, $T \geq 1$, $\alpha > 0$. Repeat for $0 \leq k \leq T - 1$:

1. Generate a stochastic subgradient g_k of f at x_k.
2. Set $x_{k+1} := x_k - \alpha g_k$.

Output: $\bar{x}_T := \frac{1}{T} \sum_{k=0}^{T-1} x_k$.

Main objects responsible for convergence rate:

- Magnitude of SG: $\mathbb{E}\|g_k\|^2 \leq M^2$ for all $k \geq 0$.
- Distance from x_0 to optimum x^*: $D^2 := \mathbb{E}\|x_0 - x^*\|^2$.

Theorem: For $\alpha := \frac{D}{M\sqrt{T}}$, we have $\mathbb{E} f(\bar{x}_T) - f^* \leq \frac{MD}{\sqrt{T}}$.

Complexity: $O(\varepsilon^{-2})$.

Example: Empirical risk minimization (ERM)

Let $\phi_1, \dots, \phi_m : \mathbb{R} \to \mathbb{R}$ be convex functions, $a_1, \dots, a_m \in \mathbb{R}^n$, and let

$$f(x) := \frac{1}{m} \sum_{i=1}^{m} \phi_i(\langle a_i, x \rangle).$$

SO: Given $x \in \mathbb{R}^n$, generate $i_0 \sim \mathrm{Unif}\{1, \dots, m\}$ and return $g(x) := \phi'_{i_0}(\langle a_{i_0}, x \rangle) a_{i_0}$.

Magnitude of data: $B := \max_{1 \le i \le m} \|a_i\|$.

NB: If $|\phi'_i(t)| \le G$ for all $t \in \mathbb{R}$, $1 \le i \le m$, then $\mathbb{E}\|g(x)\|^2 \le G^2 B^2$. Hence, $M = GB$.

1. (Robust regression) $\phi(t) := |t|$. In this case $\phi'(t) = \mathrm{sign}(t)$, and $G = 1$.
2. (Logistic regression) $\phi(t) := \ln(1 + e^t)$. Here $\phi'(t) = \frac{e^t}{1+e^t}$, hence $G = 1$.
3. (SVM) $\phi(t) := \max\{0, 1 - t\}$. Here $\phi'(t) = \begin{cases} -1 & \text{if } t \le 1 \\ 0 & \text{if } t > 1 \end{cases}$. Thus, $G = 1$.

SGD for non-smooth convex optimization: Proof

Main result: $\frac{1}{T}\sum_{k=0}^{T-1}(\mathbb{E}f(x_k) - f^*) + \frac{\mathbb{E}\|x_T - x^*\|^2}{2\alpha T} \le \frac{\mathbb{E}\|x_0 - x^*\|^2}{2\alpha T} + \frac{\alpha}{2T}\sum_{k=0}^{T-1}\mathbb{E}\|g_k\|^2$.

Proof:

$$
\begin{aligned}
\mathbb{E}\|x_{k+1} - x^*\|^2 &= \mathbb{E}\|x_k - x^* - \alpha g_k\|^2 = \mathbb{E}(\|x_k - x^*\|^2 - 2\alpha\langle g_k, x_k - x^*\rangle + \alpha^2\|g_k\|^2) \\
&= \mathbb{E}\|x_k - x^*\|^2 - 2\alpha\mathbb{E}\langle \nabla f(x_k), x_k - x^*\rangle + \alpha^2\mathbb{E}\|g_k\|^2 \\
&\le \mathbb{E}\|x_k - x^*\|^2 - 2\alpha(\mathbb{E}f(x_k) - f^*) + \alpha^2\mathbb{E}\|g_k\|^2.
\end{aligned}
$$

Hence,

$$
\begin{aligned}
\sum_{k=0}^{T-1}(\mathbb{E}f(x_k) - f^*) &\le \frac{\sum_{k=0}^{T-1}(\mathbb{E}\|x_k - x^*\|^2 - \mathbb{E}\|x_{k+1} - x^*\|^2)}{2\alpha} + \frac{\alpha}{2}\sum_{k=0}^{T-1}\mathbb{E}\|g_k\|^2 \\
&= \frac{1}{2\alpha}(\mathbb{E}\|x_0 - x^*\|^2 - \mathbb{E}\|x_T - x^*\|^2) + \frac{\alpha}{2}\sum_{k=0}^{T-1}\mathbb{E}\|g_k\|^2. \quad \square
\end{aligned}
$$

From the main result, using $\alpha = \frac{D}{M\sqrt{T}}$, we obtain

$$
\mathbb{E}f(\bar{x}_T) - f^* \le \frac{1}{T}\sum_{k=0}^{T-1}(\mathbb{E}f(x_k) - f^*) \le \underbrace{\frac{D^2}{2\alpha T}}_{\to 0} + \underbrace{\frac{\alpha M^2}{2}}_{\text{noise}} = \frac{MD}{\sqrt{T}}. \quad \square
$$

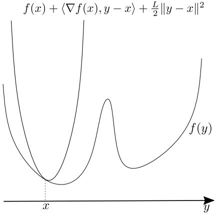
Smooth functions in optimization

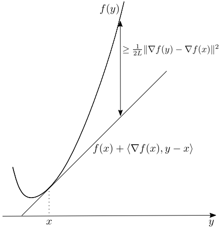
Def: A differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called <u>L-smooth</u> if $f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y-x\|^2$ for all $x, y \in \mathbb{R}^n$.
Sufficient condition: $\nabla^2 f(x) \preceq LI$ for all $x \in \mathbb{R}^n$.
Example (ERM): Let $f(x) := \phi(\langle a, x\rangle)$, where $\phi : \mathbb{R} \to \mathbb{R}$, $a \in \mathbb{R}^n$. Then

$$\nabla^2 f(x) = \phi''(\langle a, x\rangle) aa^T \preceq HB^2 I,$$

where $H := \sup_{t\in\mathbb{R}} \phi''(t)$, $B := \|a\|$. Hence, $L = HB^2$.

- (Least squares) $\phi(t) := \frac{1}{2}t^2$. Here $\phi''(t) = 1$ and hence $H = 1$.
- (Logistic regression) $\phi(t) := \ln(1 + e^t)$. Here $\phi''(t) = \frac{e^t}{(1+e^t)^2}$, and $H = \frac{1}{4}$.





Important fact: If f is convex and L-smooth, then for all $x, y \in \mathbb{R}^n$, we have

$$f(y) \ge f(x) + \langle \nabla f(x), y - x\rangle + \frac{1}{2L}\|\nabla f(y) - \nabla f(x)\|^2.$$

Variance of stochastic gradients

Let $\mathbb{E}g(x) = \nabla f(x)$.

Previous assumption: $\mathbb{E}\|g(x)\|^2 \le M^2$ (bounded second moment).

New assumption: f is L-smooth and $\mathbb{E}\|g(x) - \nabla f(x)\|^2 \le \sigma^2$ (bounded variance).

NB: Since $\mathbb{E}\|g(x) - \nabla f(x)\|^2 = \mathbb{E}\|g(x)\|^2 - \|\nabla f(x)\|^2$, we always have $\sigma \le M$. However, sometimes σ can be much smaller than M.

Example 1: Let $g(x) := \nabla f(x) + \xi(x)$, where $\mathbb{E}\xi(x) = 0$, $\mathbb{E}\|\xi(x)\|^2 \le \sigma^2$. Then $\mathbb{E}\|g(x) - \nabla f(x)\|^2 \le \sigma^2$, while $\mathbb{E}\|g(x)\|^2 = \|\nabla f(x)\|^2 + \mathbb{E}\|\xi(x)\|^2 \le \|\nabla f(x)\|^2 + \sigma^2$.

Example 2: Mini-batching (in a couple of slides).

SGD for smooth convex optimization [cf. Ghadimi-Lan, 2013]

Problem: $\min_{x\in\mathbb{R}^n} f(x)$, where f is convex L-smooth and given by an SO.

Method: Fix $x_0 \in \mathbb{R}^n$, $T \geq 1$, $\alpha > 0$. Repeat for $0 \leq k \leq T-1$:

1. Generate a stochastic gradient g_k of f at x_k.
2. Set $x_{k+1} := x_k - \alpha g_k$.

Output: $\bar{x}_T := \frac{1}{T}\sum_{k=0}^{T-1} x_k$.

Theorem: For $\alpha := \frac{1}{L+\frac{\sigma\sqrt{T}}{D}}$, we have $\mathbb{E}f(\bar{x}_T) - f^* \leq \underbrace{\frac{LD^2}{T}}_{\text{deterministic}} + \underbrace{\frac{3\sigma D}{2\sqrt{T}}}_{\text{stochastic}}$.

NB: For $\sigma = 0$, we recover the $\frac{LD^2}{T}$ convergence rate of the gradient descent (GD).

Previous result: For $\alpha := \frac{D}{M\sqrt{T}}$, we have $\mathbb{E}f(\bar{x}_T) - f^* \leq \frac{MD}{\sqrt{T}}$.
Complexity: Still $O(\varepsilon^{-2})$.

SGD in the smooth convex optimization: Proof

Let $\delta_k := g_k - \nabla f(x_k)$. Using $\|\nabla f(x_k)\|^2 \le L\langle \nabla f(x_k), x_k - x^*\rangle$, we have

$$
\begin{aligned}
\mathbb{E}\|x_{k+1} - x^*\|^2 - \mathbb{E}\|x_k - x^*\|^2 &\le -2\alpha\mathbb{E}\langle \nabla f(x_k), x_k - x^*\rangle + \alpha^2\mathbb{E}\|g_k\|^2 \\
&= -2\alpha\mathbb{E}\langle \nabla f(x_k), x_k - x^*\rangle + \alpha^2\mathbb{E}(\|\nabla f(x_k)\|^2 + 2\langle \nabla f(x_k), \delta_k\rangle + \|\delta_k\|^2) \\
&\le -\alpha(2 - L\alpha)\mathbb{E}\langle \nabla f(x_k), x_k - x^*\rangle + \alpha^2\sigma^2 \le -\alpha(2 - L\alpha)(\mathbb{E}f(x_k) - f^*) + \alpha^2\sigma^2
\end{aligned}
$$

Hence, using $\alpha = \frac{1}{L + \frac{\sigma\sqrt{T}}{D}} \le \frac{1}{L}$, we obtain

$$
\begin{aligned}
\mathbb{E}f(\bar{x}_T) - f^* &\le \frac{1}{T}\sum_{k=0}^{T-1}(\mathbb{E}f(x_k) - f^*) \le \frac{\sum_{k=0}^{T-1}(\mathbb{E}\|x_k - x^*\|^2 - \mathbb{E}\|x_{k+1} - x^*\|^2)}{\alpha(2 - L\alpha)T} + \frac{\alpha\sigma^2}{2 - L\alpha} \\
&\le \frac{D^2}{\alpha(2 - L\alpha)T} + \frac{\alpha\sigma^2}{2 - L\alpha} \le \frac{D^2(L + \frac{\sigma\sqrt{T}}{D})}{T} + \frac{\sigma D}{2\sqrt{T}} = \frac{LD^2}{T} + \frac{3\sigma D}{2\sqrt{T}}. \quad \square
\end{aligned}
$$

Minibatching

Idea: Given a point $x \in \mathbb{R}^n$, call the SO N times to obtain the i.i.d. $g_1(x), \dots, g_N(x)$, and set $g(x) := \frac{1}{N}\sum_{i=1}^N g_i(x)$. Especially efficient when $g_1(x), \dots, g_N(x)$ are computed in parallel.

Proposition: $\mathbb{E}\|g(x) - \nabla f(x)\|^2 \le \frac{\sigma^2}{N}$.

Proof:

$$\mathbb{E}\|g(x) - \nabla f(x)\|^2 = \mathbb{E}\left\|\frac{1}{N}\sum_{i=1}^N (g_i(x) - \nabla f(x))\right\|^2$$

$$= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\|g_i(x) - \nabla f(x)\|^2 + \frac{2}{N^2}\sum_{1 \le i < j \le N} \mathbb{E}\langle g_i(x) - \nabla f(x), g_j(x) - \nabla f(x)\rangle$$

$$= \frac{1}{N^2}\sum_{i=1}^N \mathbb{E}\|g_i(x) - \nabla f(x)\|^2 \le \frac{\sigma^2}{N}. \quad \square$$

Result: Instead of $\mathbb{E}f(\bar{x}_T) - f^* \le \frac{LD^2}{T} + \frac{3\sigma D}{2\sqrt{T}}$, we get $\mathbb{E}f(\bar{x}_T) - f^* \le \frac{LD^2}{T} + \frac{3\sigma D}{2\sqrt{N}\sqrt{T}}$.

SGD for smooth non-convex optimization [Ghadimi-Lan, 2013]

Problem: $\min_{x \in \mathbb{R}^n} f(x)$, where $f : \mathbb{R}^n \to \mathbb{R}$ is L-smooth, possibly non-convex.

Method: Fix $x_0 \in \mathbb{R}^n$, $T \geq 1$, $\alpha > 0$. Repeat for $0 \leq k \leq T-1$:

1. Generate a stochastic gradient g_k of f at x_k.
2. Set $x_{k+1} := x_k - \alpha g_k$.

Output: $y_T \sim \mathrm{Unif}((x_k)_{0 \leq k \leq T-1})$.

Theorem: For $\alpha := \frac{1}{L + \frac{\sigma\sqrt{T}}{D_f}}$, we have $\frac{1}{L}\mathbb{E}\|\nabla f(y_T)\|^2 \leq \frac{LD_f^2}{T} + \frac{3\sigma D_f}{2\sqrt{T}}$, where $D_f := (\frac{2}{L}(\mathbb{E}f(x_0) - f^*))^{\frac{1}{2}}$.

NB: When $\sigma = 0$, we recover the $O(\frac{1}{T})$ convergence rate of the standard GD.

Remark: If f is convex, then for $\alpha := \frac{1}{L + \frac{\sigma\sqrt{T}}{D}}$, we have $\mathbb{E}f(y_T) - f^* \leq \frac{LD^2}{T} + \frac{3\sigma D}{2\sqrt{T}}$, where $D^2 := \mathbb{E}\|x_0 - x^*\|^2$.

SGD for smooth non-convex optimization: Proof

Let $\delta_k := g_k - \nabla f(x_k)$. By L-smoothness, we have

$$\begin{aligned}
\mathbb{E} f(x_{k+1}) &\le \mathbb{E}\left(f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k \rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2 \right) \\
&= \mathbb{E} f(x_k) - \alpha \mathbb{E}\langle \nabla f(x_k), g_k \rangle + \frac{L\alpha^2}{2}\mathbb{E}\|g_k\|^2 \\
&= \mathbb{E} f(x_k) - \alpha \mathbb{E}\|\nabla f(x_k)\|^2 + \frac{L\alpha^2}{2}\mathbb{E}(\|\nabla f(x_k)\|^2 + 2\langle \nabla f(x_k), \delta_k \rangle + \|\delta_k\|^2) \\
&= \mathbb{E} f(x_k) - \frac{\alpha(2 - L\alpha)}{2}\mathbb{E}\|\nabla f(x_k)\|^2 + \frac{L\alpha^2\sigma^2}{2}.
\end{aligned}$$

Hence,

$$\frac{1}{L}\mathbb{E}\|\nabla f(y_T)\|^2 = \frac{1}{LT}\sum_{k=0}^{T-1}\mathbb{E}\|\nabla f(x_k)\|^2 \le \frac{2\sum_{k=0}^{T-1}(\mathbb{E} f(x_k) - \mathbb{E} f(x_{k+1}))}{L\alpha(2 - L\alpha)T} + \frac{L\alpha\sigma^2}{2 - L\alpha}$$

$$= \frac{2(\mathbb{E} f(x_0) - f^*)}{L\alpha(2 - L\alpha)T} + \frac{L\alpha\sigma^2}{2 - L\alpha} = \frac{D_f^2}{\alpha(2 - L\alpha)T} + \frac{L\alpha\sigma^2}{2 - L\alpha} = \frac{LD_f^2}{T} + \frac{3\sigma D_f}{2\sqrt{T}}. \quad \square$$

Part 2: Noise reduction for finite sums

Overview

Problem: $\min_{x\in\mathbb{R}^n} f(x)$, where $f(x) := \frac{1}{m}\sum_{i=1}^m f_i(x)$ with smooth $f_1, \ldots, f_m$.

Goal: Given $\varepsilon > 0$, find $\bar{x} \in \mathbb{R}^n$:

- $\mathbb{E}f(\bar{x}) - f^* \leq \varepsilon$ (convex optimization).
- $\mathbb{E}\|\nabla f(\bar{x})\|^2 \leq \varepsilon$ (non-convex optimization).

Complexity measure: Number of computations of $\nabla f_i(x)$.

Special algorithm: SVRG (Stochastic Variance Reduced Gradient method).

	Convex	Non-convex
GD	$O(m\varepsilon^{-1})$	$O(m\varepsilon^{-1})$
SGD	$O(\varepsilon^{-2})$	$O(\varepsilon^{-2})$
SVRG	$O(\varepsilon^{-1} + m\log\varepsilon^{-1})$	$O(m + m^{\frac{2}{3}}\varepsilon^{-1})$.

Noise reduction for convex optimization

Recall the convergence rate of SGD:

$$\mathbb{E}f(\bar{x}_T) - f^* \leq \frac{D^2}{2\alpha T} + \frac{\alpha M^2}{2},$$

where $\mathbb{E}\|g_k\|^2 \leq M^2$.

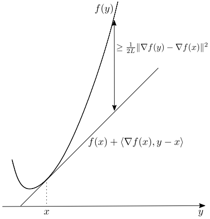
This gives $O(\frac{1}{\sqrt{T}})$ convergence rate for $\alpha = \Theta(\frac{1}{\sqrt{T}})$.

NB: When $\mathbb{E}\|g_k\|^2 \to 0$, we obtain $\mathbb{E}f(\bar{x}_T) - f^* \leq O(\frac{1}{T})$ for α not depending on T.

Example: For $g_k := \nabla f(x_k)$, we have $\mathbb{E}\|g_k\|^2 = \|\nabla f(x_k)\|^2 \to 0$.

Main question: How to ensure $\mathbb{E}\|g_k\|^2 \to 0$ in the presence of noise?

Key idea of SVRG



Let $f := \frac{1}{m}\sum_{i=1}^m f_i$. Let $x, \tilde{x} \in \mathbb{R}^n$, $i_0 \sim \text{Unif}\{1, \ldots, m\}$,

$$g(x) := \underbrace{\nabla f_{i_0}(x)}_{\mathbb{E}=\nabla f(x)} + \underbrace{(\nabla f(\tilde{x}) - \nabla f_{i_0}(\tilde{x}))}_{\mathbb{E}=\nabla f(\tilde{x})-\nabla f(\tilde{x})=0}.$$

Key lemma: Let $f_1, \ldots, f_m$ be convex and L-smooth. Then

$$\mathbb{E}\|g(x)\|^2 \leq 4L(\mathbb{E}f(x) - f^* + \mathbb{E}f(\tilde{x}) - f^*) \to 0 \text{ when } f(x),\ f(\tilde{x}) \to f^*.$$

Proof: Using the inequalities $\|u+v\|^2 \leq 2\|u\|^2 + 2\|v\|^2$, $\mathbb{E}\|\xi - \mathbb{E}\xi\|^2 \leq \mathbb{E}\|\xi\|^2$ and the important fact about convex L-smooth functions, we obtain

$$\begin{aligned}
\mathbb{E}\|g(x)\|^2 &= \mathbb{E}\|\nabla f_{i_0}(x) + (\nabla f(\tilde{x}) - \nabla f_{i_0}(\tilde{x}))\|^2 \\
&\leq 2\mathbb{E}\|\nabla f_{i_0}(x) - \nabla f_{i_0}(x^*)\|^2 + 2\mathbb{E}\|\nabla f_{i_0}(x^*) - \nabla f_{i_0}(\tilde{x}) - (\nabla f(x^*) - \nabla f(\tilde{x}))\|^2 \\
&\leq 2\mathbb{E}\|\nabla f_{i_0}(x) - \nabla f_{i_0}(x^*)\|^2 + 2\mathbb{E}\|\nabla f_{i_0}(\tilde{x}) - \nabla f_{i_0}(x^*)\|^2 \\
&\leq 4L\mathbb{E}(f_{i_0}(x) - f_{i_0}(x^*) - \langle \underbrace{\nabla f_{i_0}(x^*)}_{\mathbb{E}=0}, x - x^*\rangle) + 4L\mathbb{E}(f_{i_0}(\tilde{x}) - f_{i_0}(x^*) - \langle \underbrace{\nabla f_{i_0}(x^*)}_{\mathbb{E}=0}, \tilde{x} - x^*\rangle) \\
&= 4L(\mathbb{E}f(x) - f^*) + 4L(\mathbb{E}f(\tilde{x}) - f^*). \quad \square
\end{aligned}$$

SVRG for convex optimization [cf. Allen-Zhu and Yuan, 2016]

Problem: $\min_{x\in\mathbb{R}^n} f(x)$, where $f(x) := \frac{1}{m}\sum_{i=1}^m f_i(x)$ and f_i are convex.

Method: Fix $x_0 \in \mathbb{R}^n$, $\alpha > 0$. Set $\tilde{x}_0 := x_0$, $x_0^0 := x_0$. Repeat for $0 \le s \le S-1$:

- Compute $\tilde{g}_s := \nabla f(\tilde{x}_s) = \frac{1}{m}\sum_{i=1}^m \nabla f_i(\tilde{x}_s)$.
- Repeat for $0 \le k \le K_s - 1$:
  - Set $x_{k+1}^s := x_k^s - \alpha g_k^s$, where $g_k^s := \nabla f_{i_k^s}(x_k^s) + (\tilde{g}_s - \nabla f_{i_k^s}(\tilde{x}_s))$, $i_k^s \sim \mathrm{Unif}\{1,\dots,m\}$.
- Set $\tilde{x}_{s+1} := \frac{1}{K_s}\sum_{k=0}^{K_s-1} x_k^s$ and $x_0^{s+1} := x_{K_s}^s$.

Output: $\tilde{x}_S$.

Complexity: $O(\sum_{s=0}^{S-1} K_s + mS)$.

Theorem: Let $\Delta := \mathbb{E}f(x_0) - f^*$, $D^2 := \mathbb{E}\|x_0 - x^*\|^2$. For $\alpha := \frac{1}{6L}$, $K_s := 2^s K_0$, $K_0 := \frac{9LD^2}{\Delta}$, $S := \log_2 \frac{2\Delta}{\varepsilon}$, we have $\mathbb{E}f(\tilde{x}_S) - f^* \le \varepsilon$. Complexity: $O(\frac{LD^2}{\varepsilon} + m\log\frac{\Delta}{\varepsilon})$.

Gain: $O(\varepsilon^{-1})$ instead of the $O(\varepsilon^{-2})$ of SGD.

SVRG for convex optimization: Proof

It suffices to prove that $\mathbb{E}f(\tilde{x}_S) - f^* \le \frac{\Delta + \frac{9LD^2}{K_0}}{2^S}$ (*), and then plug in K_0 and S.

By the main result of SGD and the key lemma of SVRG, we have

$$\begin{aligned}\frac{1}{K_s}\sum_{k=0}^{K_s-1}(\mathbb{E}f(x_k^s) - f^*) + \frac{1}{2\alpha K_s}\mathbb{E}\|x_{K_s}^s - x^*\|^2 &\le \frac{1}{2\alpha K_s}\mathbb{E}\|x_0^s - x^*\|^2 + \frac{\alpha}{2K_s}\sum_{k=0}^{K_s-1}\mathbb{E}\|g_k^s\|^2 \\ &\le \frac{1}{2\alpha K_s}\mathbb{E}\|x_0^s - x^*\|^2 + 2L\alpha(\mathbb{E}f(\tilde{x}_s) - f^*) + \frac{2L\alpha}{K_s}\sum_{k=0}^{K_s-1}(\mathbb{E}f(x_k^s) - f^*).\end{aligned}$$

Hence,

$$\frac{1}{K_s}\sum_{k=0}^{K_s-1}(\mathbb{E}f(x_k^s) - f^*) + \frac{\mathbb{E}\|x_{K_s}^s - x^*\|^2}{2\alpha K_s(1-2L\alpha)} \le \frac{1}{2}\left(\frac{4L\alpha}{1-2L\alpha}(\mathbb{E}f(\tilde{x}_s) - f^*) + \frac{\mathbb{E}\|x_0^s - x^*\|^2}{\alpha K_s(1-2L\alpha)}\right)$$

Using $\alpha := \frac{1}{6L}$, $K_{s+1} := 2K_s$, $\tilde{x}_{s+1} := \frac{1}{K_s}\sum_{k=0}^{K_s-1} x_k$ and $x_0^{s+1} := x_{K_s}^s$, we obtain

$$\mathbb{E}f(\tilde{x}_{s+1}) - f^* + \frac{\mathbb{E}\|x_0^{s+1} - x^*\|^2}{\alpha K_{s+1}(1-2L\alpha)} \le \frac{1}{2}\left(\mathbb{E}f(\tilde{x}_s) - f^* + \frac{\mathbb{E}\|x_0^s - x^*\|^2}{\alpha K_s(1-2L\alpha)}\right).$$

Now (*) follows by induction. $\square$

SVRG for non-convex optimization [Reddi et al., 2016]

Objective: $f(x) := \frac{1}{m}\sum_{i=1}^{m} f_i(x)$, where f_i are L-smooth but possibly non-convex.

Method: Same as before but now $K_s := K$ (constant number of inner iterations) and $\tilde{x}_{s+1} := x_0^{s+1} := x_K^s$ (last iterate).
Output: $y_T \sim \mathrm{Unif}((x_k^s)_{0\le k\le K-1; 0\le s\le S-1})$.

Theorem: Let $T \ge 1$. For $\alpha := \Theta(\frac{1}{Lm^{2/3}})$, $K := \Theta(m)$ and $S := T/K$, we have $\frac{1}{L}\mathbb{E}\|\nabla f(y_T)\|^2 = O(\frac{m^{2/3}}{T}(\mathbb{E}f(x_0) - f^*))$ with complexity is $O(m+T)$.

Corollary: SVRG complexity is $O(m + m^{2/3}\varepsilon^{-1})$.

Complexity of SGD: $O(\varepsilon^{-2})$. **Complexity of GD:** $O(m\varepsilon^{-1})$.

Practical performance [Reddi et al., 2016]

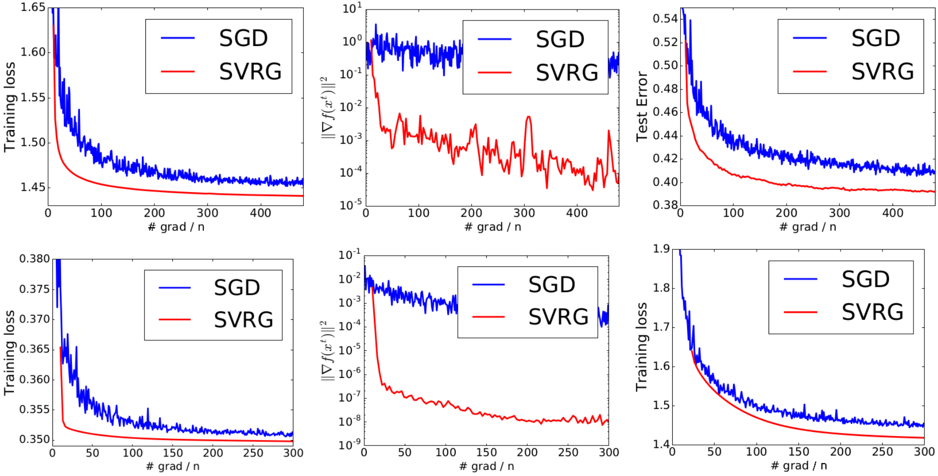


Figure: Neural network results for CIFAR-10, MNIST and STL-10 datasets.

Conclusion

Part 1: General stochastic optimization:

- Use random unbiased estimates $g(x)$ of the true gradient $\nabla f(x)$.
- Main method: SGD. Complexity: $O(\varepsilon^{-2})$.
- Important characteristics:
  - Magnitude of stochastic gradients: $\mathbb{E}\|g(x)\|^2 \leq M^2$.
  - Variance of stochastic gradients: $\mathbb{E}\|g(x) - \nabla f(x)\|^2 \leq \sigma^2$.

Part 2: Noise reduction for finite sums

- When $f(x) = \frac{1}{m}\sum_{i=1}^{m} f_i(x)$, both M and σ can be dynamically reduced provided that one can evaluate objective several times.
- Gain: More efficient method SVRG. Complexity: $O(\varepsilon^{-1})$.