13. Stochastic and online algorithms

• stochastic gradient method

• online optimization and dual averaging method

• minimizing finite average
Stochastic optimization problem

\[
\min_{x \in X} \left\{ F(x) \overset{\text{def}}{=} \mathbb{E}_{\xi} f(x, \xi) \right\}
\]

- \(X \subset \mathbb{R}^n\) is a (bounded) closed convex set
- \(\xi\) is a random vector whose distribution \(P\) is supported on set \(\Xi \subset \mathbb{R}^d\)
- \(f : X \times \Xi \to \mathbb{R}\), and the expectation

\[
\mathbb{E}_{\xi} f(x, \xi) = \int_{\Xi} f(x, \xi) dP(\xi)
\]

is well defined and has finite value for every \(x \in X\)

- \(F(\cdot)\) continuous and convex on \(X\), and optimal value \(F^*\) attained at \(x^*\)
  (e.g., \(F(\cdot)\) is convex if \(f(\cdot, \xi)\) is convex for every \(\xi \in \Xi\)
Sample average approximation

\[
\minimize_{x \in X} \left\{ \hat{F}_N(x) \overset{\text{def}}{=} \frac{1}{N} \sum_{j=1}^{N} f(x, \xi_j) \right\}
\]

- **assumption**: \(\{\xi_j\}_{j=1}^{N}\) is a sequence of independent random outcomes
- reasonably efficient when solved by appropriate (deterministic) algorithm
- **sample complexity**: suppose \(f\) has bounded variation, and let

\[
V = \max \{ f(x_1, \xi_1) - f(x_2, \xi_2) \mid x_1, x_2 \in X, \ \xi_1, \xi_2 \in \Xi \}
\]

then for any \(\epsilon > 0\) and \(\rho \in (0, 1)\), sample size \(N = \left\lceil \frac{V^2}{2\epsilon^2 \ln \frac{2}{\rho}} \right\rceil\) guarantees

\[
\text{prob}\left( |\hat{F}_N(x) - F(x)| \leq \epsilon \right) \geq 1 - \rho, \quad \forall x \in X
\]

(proved using Hoeffding inequality in probability theory)
Stochastic approximation

choose $x^{(1)} \in X$, and iterate for $k = 1, 2, \ldots$

$$x^{(k+1)} = \pi_X \left( x^{(k)} - t_k \, g(x^{(k)}, \xi_k) \right)$$

- $g(x, \xi)$ is a stochastic subgradient, i.e., $g(x, \xi) \in \partial_x f(x, \xi)$ and
  $$F'(x) \overset{\text{def}}{=} \mathbb{E}_\xi g(x, \xi) \in \partial F(x)$$

assumption: there exist a constant $G$ such that

$$\mathbb{E}_\xi \left[ \| g(x, \xi) \|_2^2 \right] \leq G^2, \quad \forall x \in X$$

- $\pi_X(\cdot)$ denotes projection onto $X$:
  $$\pi_X(x) = \operatorname{argmin}_{y \in X} \| y - x \|_2^2$$
Convergence analysis

consider squared distance to $x^*$, and let $r_k = \mathbf{E}[\|x^{(k)} - x^*\|_2^2]$

$$\|x^{(k+1)} - x^*\|_2^2 = \|\pi_X(x^{(k)} - t_k g(x^{(k)}, \xi_k)) - \pi_X(x^*)\|_2^2$$

$$\leq \|x^{(k)} - t_k g(x^{(k)}, \xi_k) - x^*\|_2^2$$

$$= \|x^{(k)} - x^*\|_2^2 - 2t_k(x^{(k)} - x^*)^T g(x^{(k)}, \xi_k) + t_k^2\|g(x^{(k)}, \xi_k)\|_2^2$$

since $x^{(k)}$ is a function of $\xi_{[k-1]} = (\xi_0, \ldots, \xi_{k-1})$, it is independent of $\xi_k$

$$\mathbf{E}[(x^{(k)} - x^*)^T g(x^{(k)}, \xi_k)] = \mathbf{E}\{\mathbf{E}[(x^{(k)} - x^*)^T g(x^{(k)}, \xi_k) | \xi_{[k-1]}]\}$$

$$= \mathbf{E}\{(x^{(k)} - x^*)^T \mathbf{E}[g(x^{(k)}, \xi_k) | \xi_{[k-1]}]\}$$

$$= \mathbf{E}\{(x^{(k)} - x^*)^T F'(x^{(k)})\}$$

therefore

$$r_{k+1} \leq r_k - 2t_k \mathbf{E}[(x^{(k)} - x^*)^T F'(x^{(k))}] + t_k^2 G^2 \quad (1)$$
by convexity of $F$, it holds $F(x^*) \geq F(x^{(k)}) + (x^* - x^{(k)})^T F'(x^{(k)})$, hence

$$E[(x^{(k)} - x^*)^T F'(x^{(k)})] \geq E[F(x^{(k)}) - F^*]$$

combining with (1) gives

$$t_k E[F(x^{(k)}) - F^*] \leq \frac{1}{2}(r_k - r_{k+1} + t_k^2 G^2)$$

summing over $j = 1, \ldots, k$ yields

$$\sum_{j=1}^k t_j E[F(x^{(j)}) - F^*] \leq \frac{1}{2}\left(r_1 - r_{k+1} + G^2 \sum_{j=1}^k t_j^2\right) \leq \frac{1}{2}\left(r_1 + G^2 \sum_{j=1}^k t_j^2\right)$$

let $\nu_j^{(k)} = \frac{t_j}{\sum_{i=1}^k t_i}$ and $\tilde{x}^{(k)} = \sum_{j=1}^k \nu_j^{(k)} x^{(j)}$ (note $\sum_{j=1}^k \nu_j^{(k)} = 1$), then

$$E[F(\tilde{x}^{(k)}) - F^*] \leq E\left[\sum_{j=1}^k \nu_j^{(k)} F(x^{(j)}) - F^*\right] \leq \frac{r_1 + G^2 \sum_{j=1}^k t_j^2}{2 \sum_{j=1}^k t_j}$$
Fixed step size

suppose the number of iterations $N$ is known in advance, then

$$
E[F(\tilde{x}^{(k)}) - F^*] \leq \frac{D^2 + G^2 N t^2}{2Nt}
$$

where $D = \max_{x \in X} \|x - x^*\|_2$, so that $r_1 = E\|x^{(1)} - x^*\|_2^2 \leq D^2$

● minimizing upper bound over $t > 0$ gives $t = \frac{D}{G \sqrt{N}}$ and

$$
E[F(\tilde{x}^{(k)}) - F^*] \leq \frac{DG}{\sqrt{N}}
$$

● if $t = \frac{\theta D}{G \sqrt{N}}$ for some constant $\theta > 0$, then

$$
E[F(\tilde{x}^{(k)}) - F^*] \leq \max\{\theta, \theta^{-1}\} \frac{DG}{\sqrt{N}}
$$

therefore, $O(1/\sqrt{N})$ convergence robust against step size choices
Diminishing step size

following the halving trick in deterministic subgradient method, redefine

$$\tilde{x}(k) = \frac{\sum_{k/2 \leq j \leq k} t_j x(j)}{\sum_{k/2 \leq j \leq k} t_j}$$

if the step sizes are chosen as

$$t_k = \frac{\theta D}{G \sqrt{k}}$$

then the following holds with a constant $C > 1$

$$\mathbb{E}[F(\tilde{x}(k)) - F^*] \leq C \max\{\theta, \theta^{-1}\} \frac{DG}{\sqrt{k}}$$

$O(1/\sqrt{k})$ convergence rate is optimal for general convex functions
Analysis for strongly convex functions

assume $F = \mathbb{E}_\xi f(x, \xi)$ is differentiable and strongly convex

$$F(y) \geq F(x) + \nabla F(x)^T(y - x) + \frac{\mu}{2}\|y - x\|^2_2 \quad \forall x, y \in X$$

or equivalently

$$(x - y)^T(\nabla F(x) - \nabla F(y)) \geq \mu\|x - y\|^2_2, \quad \forall x, y \in X$$

by optimality of $x^*$,

$$(x - x^*)^T \nabla F(x^*) \geq 0, \quad \forall x \in X$$

therefore

$$(x - x^*)^T \nabla F(x) \geq \mu\|x - x^*\|^2_2, \quad \forall x \in X$$  \hspace{1cm} (2)$$
combining (1) and (2) gives

\[ r_{k+1} \leq (1 - 2 \mu t_k) r_k + t_k^2 G^2 \]

let’s take step size \( t_k = \theta/k \) for some constant \( \theta > 1/(2 \mu) \), then

\[ r_{k+1} \leq (1 - 2 \mu \theta/k) r_k + \theta^2 G^2 / k^2 \]

• it follows by induction that (Nemirovski et al. 2009)

\[
E \left[ \| x^{(k)} - x^* \|_2^2 \right] = r_k \leq \frac{Q(\theta)}{k} 
\]

where \( Q(\theta) = \max\{ \theta^2 G^2 (2 \mu \theta - 1)^{-1}, \| x^{(1)} - x^* \|_2^2 \} \)

• if in addition \( \nabla F \) is Lipschitz continuous with constant \( L > 0 \), then

\[
E \left[ F(x^{(k)}) - F^* \right] \leq \frac{L}{2} E \left[ \| x^{(k)} - x^* \|_2^2 \right] \leq \frac{LQ(\theta)}{2k} 
\]
Sensitivity to priori knowledge of $\mu$

example: let $F(x) = x^2/10$, $X = [-1, 1]$, $\mu = 0.2$, and there is no noise

- if $\theta = 1$ (which violates the condition $\theta > 1/(2\mu)$), then

$$x^{(k+1)} = x^{(k)} - \frac{1}{k}F'(x^{(k)}) = \left(1 - \frac{1}{5k}\right)x^{(k)}$$

starting with $x^{(1)} = 1$ leads to

$$x^{(k)} > 0.8k^{-1/5}$$

error is larger than 0.015 even after $10^9$ iterations!

- if $\theta = 1/\mu = 5$, then $x^* = 0$ is obtained in one iteration

- step size $t_k = \theta/k$ too small if $F$ is not strongly convex
Outline

- stochastic approximation
- online optimization and dual averaging method
- minimizing finite average
Online convex optimization

• explained as online game: for $k = 1, 2, 3, \ldots$,
  – player chooses $x^{(k)} \in X$ based on previous information
  – adversary reveals cost function $f_k$, and player incurs loss $f_k(x^{(k)})$

assumptions: $f_k$ convex; $X$ bounded, closed and convex

• player wants to minimize regret:

$$R_N \triangleq \sum_{k=1}^{N} (f_k(x^{(k)})) - \min_{x \in X} \left\{ \sum_{k=1}^{N} f_k(x) \right\}$$

• online subgradient method

$$x^{(k+1)} = \pi_X (x^{(k)} - t_k g^{(k)}), \quad g^{(k)} \in \partial f_k(x^{(k)})$$

with appropriate step size, can show $R_N \leq O(\sqrt{N})$
Connection to stochastic approximation

- a more general framework without stochastic assumptions

- suppose $f_k(x) \overset{\text{def}}{=} f(x, \xi_k)$, and let $\bar{x}^{(N)} = \frac{1}{N} \sum_{k=1}^{N} x^{(k)}$, then

$$F(\bar{x}^{(N)}) - F^* \leq \frac{1}{N} \mathbb{E}[R_N]$$

proof:

$$F(\bar{x}^{(N)}) - F^* \leq \frac{1}{N} \sum_{k=1}^{N} \left( F(x^{(k)}) - F^* \right)$$

$$= \frac{1}{N} \sum_{k=1}^{N} \left( \mathbb{E}[f(x^{(k)}, \xi_k)] - \min_x \mathbb{E}[f(x, \xi_k)] \right)$$

$$= \frac{1}{N} \mathbb{E} \left[ \sum_{k=1}^{N} \left( f(x^{(k)}, \xi_k) - f(x^*, \xi_k) \right) \right]$$
**Dual averaging method (Nesterov)**

initialize: choose $x^{(1)} \in \mathbb{R}^n$ and set $s^{(0)} = 0$

iterate for $k = 0, 1, 2, \ldots$

1. compute $g^{(k)} \in \partial f_k(x^{(k)})$ and set

   $$s^{(k)} = s^{(k-1)} + g^{(k)}$$

2. update:  

   $$x^{(k+1)} = \arg\min_{x \in X} \left\{ \langle s^{(k)}, x \rangle + \frac{\beta_k}{2} \| x - x^{(0)} \|_2^2 \right\}$$

   $$= \pi_X \left( x^{(0)} - \frac{1}{\beta_k} s^{(k)} \right)$$

- choice of $\{\beta_k\}$: e.g., $\beta_k = \gamma \sqrt{k}$ with $\gamma > 0$

- can also work with composite objectives: minimize $f(x) + \Psi(x)$
A soft support function

for any $\beta \geq 0$ and any $x^{(0)} \in X$, define

$$V_\beta(s) = \max_{x \in X} \left\{ \langle s, x - x^{(0)} \rangle - \frac{\beta}{2} \|x - x^{(0)}\|_2^2 \right\}$$

- $V_\beta(s) \geq 0$ for any $\beta \geq 0$; if $\beta_2 \geq \beta_1 > 0$, then $V_{\beta_2}(s) \leq V_{\beta_1}(s)$
- $V_\beta(\cdot)$ is convex and differentiable
- $\nabla V_\beta$ is Lipschitz continuous with constant $1/\beta$

$$\|\nabla V_\beta(s_1) - \nabla V_\beta(s_2)\|_2 \leq \frac{1}{\beta} \|s_1 - s_2\|_2, \quad \forall s_1, s_2 \in \mathbb{R}^n$$

therefore

$$V_\beta(s + \delta) \leq V_\beta(s) + \langle \delta, \nabla V_\beta(s) \rangle + \frac{1}{2\beta} \|\delta\|_2^2$$
**Lemma:** let \( D = \max_{x \in X} \| x - x^{(0)} \|_2 \), then

\[
\max_{x \in X} \langle s, x - x^{(0)} \rangle \leq \frac{\beta D^2}{2} + V_\beta(s)
\]

**Proof:**

\[
\max_{x \in X} \langle s, x - x^{(0)} \rangle = \max_{x \in X} \left\{ \langle s, x - x^{(0)} \rangle : \frac{1}{2} \| x - x^{(0)} \|_2^2 \leq \frac{1}{2} D^2 \right\}
\]

\[
= \max_{x \in X} \min_{\beta \geq 0} \left\{ \langle s, x - x^{(0)} \rangle + \frac{\beta}{2} \left( D^2 - \| x - x^{(0)} \|_2^2 \right) \right\}
\]

\[
\leq \min_{\beta \geq 0} \left\{ \max_{x \in X} \left\{ \langle s, x - x^{(0)} \rangle + \frac{\beta}{2} \left( D^2 - \| x - x^{(0)} \|_2^2 \right) \right\} \right\}
\]

\[
\leq \max_{x \in X} \left\{ \langle s, x - x^{(0)} \rangle + \frac{\beta}{2} \left( D^2 - \| x - x^{(0)} \|_2^2 \right) \right\}
\]

\[
\leq \frac{\beta D^2}{2} + V_\beta(s)
\]
Convergence analysis

\[ V_{\beta_k}(-s^{(k)}) \leq V_{\beta_{k-1}}(-s^{(k)}) \]
\[ \leq V_{\beta_{k-1}}(-s^{(k-1)}) + \langle -g^{(k)}, \nabla V_{\beta_{k-1}}(-s^{(k-1)}) \rangle + \frac{1}{2\beta_{k-1}} \|g^{(k)}\|_2^2 \]
\[ = V_{\beta_{k-1}}(-s^{(k-1)}) - \langle g^{(k)}, x^{(k)} - x^{(0)} \rangle + \frac{1}{2\beta_{k-1}} \|g^{(k)}\|_2^2 \]

therefore

\[ \langle g^{(k)}, x^{(k)} - x^{(0)} \rangle \leq V_{\beta_{k-1}}(-s^{(k-1)}) - V_{\beta_k}(-s^{(k)}) + \frac{1}{2\beta_{k-1}} \|g^{(k)}\|_2^2 \]

summing over \( k = 2, \ldots, N \) and choose \( x^{(0)} = x^{(1)} \) results in

\[ \sum_{k=1}^{N} \langle g^{(k)}, x^{(k)} - x^{(0)} \rangle \leq V_{\beta_1}(-s^{(1)}) - V_{\beta_N}(-s^{(N)}) + \sum_{k=2}^{N} \frac{1}{2\beta_{k-1}} \|g^{(k)}\|_2^2 \]
\[ \delta_N \overset{\text{def}}{=} \max_{x \in X} \sum_{k=1}^{N} \langle g^{(k)}, x^{(k)} - x \rangle \]

\[ = \sum_{k=1}^{N} \langle g^{(k)}, x^{(k)} - x^{(0)} \rangle + \max_{x \in X} \sum_{k=1}^{N} \langle g^{(k)}, x^{(0)} - x \rangle \]

\[ = \sum_{k=1}^{N} \langle g^{(k)}, x^{(k)} - x^{(0)} \rangle + \max_{x \in X} \langle -s^{(N)}, x - x^{(0)} \rangle \]

\[ \leq V_{\beta_1}(-s^{(1)}) - V_{\beta_N}(-s^{(N)}) + \sum_{k=2}^{N} \frac{1}{2\beta_{k-1}} \|g^{(k)}\|_2^2 + \frac{\beta_N D^2}{2} + V_{\beta_N}(-s^{(N)}) \]

\[ \leq \frac{1}{2\beta_1} \|g^{(1)}\|_2^2 + \sum_{k=2}^{N} \frac{1}{2\beta_{k-1}} \|g^{(k)}\|_2^2 + \frac{\beta_N D^2}{2} \]

\[ \leq \frac{\beta_N D^2}{2} + \sum_{k=0}^{N-1} \frac{G^2}{2\beta_k} \quad (\text{for convenience, define } \beta_0 = \beta_1) \]
by convexity,

\[ \delta_N \stackrel{\text{def}}{=} \max_{x \in X} \sum_{k=1}^{N} \langle g^{(k)}, x^{(k)} - x \rangle \]

\[ \geq \max_{x \in X} \sum_{k=1}^{N} (f_k(x^{(k)}) - f_k(x)) \]

\[ = \sum_{k=1}^{N} f_k(x^{(k)}) - \min_{x \in X} \sum_{k=1}^{N} f_k(x) \]

therefore, \( R_N \leq \delta_N \), so

\[ R_N \stackrel{\text{def}}{=} \sum_{k=1}^{N} f_k(x^{(k)}) - \min_{x \in X} \sum_{k=1}^{N} f_k(x) \leq \frac{\beta_N D^2}{2} + \sum_{k=0}^{N-1} \frac{G^2}{2\beta_k} \]
choose parameters
\[ \beta_k = \gamma \sqrt{k}, \quad k \geq 1 \]

and let \( \beta_0 = \beta_1 \), then

\[
\sum_{k=0}^{N-1} \frac{G^2}{2\beta_k} = \frac{G^2}{2\gamma} \left( 1 + \sum_{k=1}^{N-1} \frac{1}{\sqrt{k}} \right) \leq \frac{G^2}{2\gamma} \left( 2 + \int_1^N \frac{1}{\sqrt{t}} dt \right) = \frac{G^2 \sqrt{N}}{\gamma}
\]

finally,

\[ R_N \leq \left( \gamma \frac{D^2}{2} + \frac{G^2}{\gamma} \right) \sqrt{N} \]

upper bound is minimized by choosing

\[ \gamma^* = \sqrt{2} \frac{G}{D} \]

which yields

\[ R_N \leq \sqrt{2} GD \sqrt{N} \]
Outline

• stochastic approximation

• online optimization and dual averaging method

• {

• minimizing finite average
Minimizing finite average of convex functions

problem

$$\text{minimize} \quad F(x) \overset{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} f_i(x)$$

stochastic gradient method: pick \( i_k \in \{1, \ldots, n\} \) randomly and update

$$x_{k+1} = x_k - \eta_k \nabla f_{i_k}(x_k)$$

two perspectives:

- **stochastic optimization**: viewed as trying to minimize \( \mathbb{E}_{\xi} f(x, \xi) \)

- **deterministic optimization**: a randomized incremental gradient method for a structured convex problem
stochastic optimization perspective:

- complexity theory: $O\left(\frac{1}{\epsilon^2}\right)$, or $O\left(\frac{1}{\epsilon}\right)$ with strong convexity

deterministic optimization perspective:

- sanity check: should at least beat full gradient methods:
  complexity $O\left(n\frac{L}{\mu} \log \frac{1}{\epsilon}\right)$ or $O\left(n\sqrt{\frac{L}{\mu}} \log \frac{1}{\epsilon}\right)$

- recent progress: SAG and SVRG by exploiting finite average structure
Stochastic average gradient (SAG)

- SAG method (Le Roux, Schmidt, Bach 2012)

\[ x_{k+1} = x_k - \frac{\alpha_k}{n} \sum_{i=1}^{n} g_k^{(i)} \]

where

\[ g_k^{(i)} = \begin{cases} \nabla f_i(x_k) & \text{if } i = i_k \\ g_{k-1}^{(i)} & \text{otherwise} \end{cases} \]

- a randomized variant of incremental aggregated gradient (IAG) of Blatt, Hero, & Gauchman (2007)

- complexity (# component gradient evaluations): \( O(\max\{n, \frac{L}{\mu}\} \log \frac{1}{\varepsilon}) \)
  cf. full gradient method: \( O(n\frac{L}{\mu} \log \frac{1}{\varepsilon}) \), and stochastic gradient: \( O(\frac{1}{\varepsilon}) \)

- need to store most recent gradient of each component, but can be avoided for some structured problems
Stochastic variance reduced gradient (SVRG)

- SVRG (Johnson & Zhang 2013, Mahdavi, Zhang & Jin 2013)

\[ x_{k+1} = x_k - \eta (\nabla f_{i_k}(x_k) - \nabla f_{i_k}(\tilde{x}) + \nabla F(\tilde{x})) \]

and update \( \tilde{x} \) periodically (every few passes)

- still a stochastic gradient method

\[
\mathbb{E}[\nabla f_{i_k}(x_k) - \nabla f_{i_k}(\tilde{x}) + \nabla F(\tilde{x})] \\
= \nabla F(x_k) - \nabla F(\tilde{x}) + \nabla F(\tilde{x}) \\
= \nabla F(x_k)
\]

- expected update direction is the same as \( \mathbb{E} f_{i_k}(x_k) \)
- variance can be diminishing if \( \tilde{x} \) updated periodically

- complexity: \( O \left( (n + \frac{L}{\mu}) \log \frac{1}{\epsilon} \right) \), cf. SAG: \( O \left( \max\{n, \frac{L}{\mu}\} \log \frac{1}{\epsilon} \right) \)
Stochastic variance reduced gradient (SVRG)

- computational cost per iteration:
  - unlike SAG, no need to store gradients for each component
  - need to compute two gradients at each iteration, and also full gradient periodically
  - for many structured problems, two gradients at each iteration can be reduced to only one

- intuition of variance reduction

\[ \nabla f_{i_k}(\tilde{x}) \]
\[ \nabla F(\tilde{x}) - \nabla f_{i_k}(\tilde{x}) \]
\[ \nabla F(x_k) - \nabla f_{i_k}(\tilde{x}) \]
Problem statement and assumptions

\[
\begin{align*}
\text{minimize} \quad & \quad F(x) = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \\
\text{subject to} \quad & \quad x \in \mathbb{R}^d
\end{align*}
\]

assumptions:

- each \( f_i(x) \), for \( i = 1, \ldots, n \), is convex

- each \( f_i(x) \) is smooth with Lipschitz constant \( L \)

\[
\|\nabla f_i(x) - \nabla f_i(y)\| \leq L \|x - y\|
\]

(which implies that \( \nabla F(x) \) also has Lipschitz constant \( L \))

- \( F(x) \) strongly convex: for all \( x, y \in \mathbb{R}^d \),

\[
F(y) \geq F(x) + \nabla F(x)^T (y - x) + \frac{\mu}{2} \|y - x\|^2
\]
SVRG method

**input:** \( \tilde{x}_0, \eta, m \)

**iterate:** for \( s = 1, 2, \ldots \)

\[
\tilde{x} = \tilde{x}_{s-1}
\]

\[
\tilde{v} = \nabla F(\tilde{x})
\]

\[
x_0 = \tilde{x}
\]

**iterate:** for \( k = 1, 2, \ldots, m \)

pick \( i_k \in \{1, \ldots, n\} \) uniformly at random

\[
x_k = x_{k-1} - \eta \left( \nabla f_{i_k}(x_{k-1}) - \nabla f_{i_k}(\tilde{x}) + \tilde{v} \right)
\]

end

set \( \tilde{x}_s = \frac{1}{m} \sum_{k=1}^{m} x_{k-1} \)

end
Convergence analysis of SVRG

- **Theorem:** Suppose $0 < \eta \leq 1/2L$ and $m$ sufficiently large so that

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

then we have geometric convergence in expectation:

$$E F(\tilde{x}_s) - F(x_\star) \leq \rho^s [F(\tilde{x}_0) - F(x_\star)]$$

- **More concretely,** if $\eta = \theta / L$, then

$$\rho = \frac{L/\mu}{\theta (1 - 2\theta)m} + \frac{2\theta}{1 - 2\theta}$$

choosing $\theta = 0.1$ and $m = 50(L/\mu)$ results in $\rho = 1/2$

- **Overall complexity:** $O \left( \left( \frac{L}{\mu} + n \right) \log \left( \frac{1}{\epsilon} \right) \right)$
Proof

• let $g_k = \nabla f_{i_k}(x_{k-1}) - \nabla f_{i_k}(\tilde{x}) + \nabla F(\tilde{x})$, then

$$x_k = x_{k-1} - \eta g_k,$$

and $\mathbb{E}_{i_k}[g_k] = \nabla F(x_{k-1})$

• similar as in classical analysis of stochastic gradient methods

$$\mathbb{E}\|x_k - x_*\|^2 = \mathbb{E}\|x_{k-1} - \eta g_k - x_*\|^2$$

$$= \|x_{k-1} - x_*\|^2 - 2\eta(x_{k-1} - x_*)^T\mathbb{E}[g_k] + \eta^2\mathbb{E}[\|g_k\|^2]$$

$$= \|x_{k-1} - x_*\|^2 - 2\eta(x_{k-1} - x_*)^T\nabla F(x_{k-1}) + \eta^2\mathbb{E}[\|g_k\|^2]$$

$$\leq \|x_{k-1} - x_*\|^2 - 2\eta(F(x_{k-1}) - F(x_*)) + \eta^2\mathbb{E}[\|g_k\|^2]$$

then need to bound $\mathbb{E}[\|g_k\|^2]$ carefully using the finite average structure
• by smoothness of $f_i(x)$,

$$\|\nabla f_i(x) - \nabla f_i(x_*)\|^2 \leq 2L [f_i(x) - f_i(x_*) - \nabla f_i(x_*)^T (x - x_*)]$$

• summing above inequalities over $i = 1, \ldots, n$ and using $\nabla F(x_*) = 0$,

$$\frac{1}{n} \sum_{i=1}^{n} \|\nabla f_i(x) - \nabla f_i(x_*)\|^2 \leq 2L [F(x) - F(x_*)]$$

$$\mathbb{E}\|g_k\|^2 = \mathbb{E}\|\nabla f_{ik}(x_{k-1}) - \nabla f_{ik}(x_*) + \nabla f_{ik}(x_*) - \nabla f_{ik}(\tilde{x}) + \nabla F(\tilde{x})\|^2$$

$$\leq 2\mathbb{E}\|\nabla f_{ik}(x_{k-1}) - \nabla f_{ik}(x_*)\|^2 + 2\mathbb{E}\|\nabla f_{ik}(\tilde{x}) - \nabla f_{ik}(x_*) - \nabla F(\tilde{x})\|^2$$

$$= 2\mathbb{E}\|\nabla f_{ik}(x_{k-1}) - \nabla f_{ik}(x_*)\|^2$$

$$+ 2\mathbb{E}\|\nabla f_{ik}(\tilde{x}) - \nabla f_{ik}(x_*) - \mathbb{E}[\nabla f_{ik}(\tilde{x}) - \nabla f_{ik}(x_*)]\|^2$$

$$\leq 2\mathbb{E}\|\nabla f_{ik}(x_{k-1}) - \nabla f_{ik}(x_*)\|^2 + 2\mathbb{E}\|\nabla f_{ik}(\tilde{x}) - \nabla f_{ik}(x_*)\|^2$$

$$\leq 4L [F(x_{k-1}) - F(x_*) + F(\tilde{x}) - F(x_*)]$$
continue derivation on page 13–29

\[ \mathbb{E}\|x_k - x_*\|^2 \leq \|x_{k-1} - x_*\|^2 - 2\eta(1 - 2L\eta)[F(x_{k-1}) - F(x_*)] + 4L\eta^2[F(\tilde{x}) - F(x_*)] \]

summing over \( k = 1, \ldots, m \), and take expectation w.r.t. whole history

\[ \mathbb{E}\|x_m - x_*\|^2 + 2\eta(1 - 2L\eta) \sum_{k=0}^{m-1} \mathbb{E}[F(x_k) - F(x_*)] \leq \mathbb{E}\|x_0 - x_*\|^2 + 4Lm\eta^2\mathbb{E}[F(x_0) - F(x_*)] \]

\[ \leq \frac{2}{\mu} \mathbb{E}[F(x_0) - F(x_*)] + 4Lm\eta^2\mathbb{E}[F(x_0) - F(x_*)] \]

therefore, for each stage \( s \)

\[ \mathbb{E}[F(\tilde{x}_s) - F(x_*)] \leq \frac{1}{m} \sum_{k=0}^{m-1} \mathbb{E}[F(x_k) - F(x_*)] \leq \frac{1}{2\eta(1 - 2L\eta)m} \left( \frac{2}{\mu} + 4Lm\eta^2 \right) \mathbb{E}[F(x_0) - F(x_*)] \]
Numerical experiments

- binary classification: \((a_1, b_1), \ldots, (a_n, b_n)\) with \(a_i \in \mathbb{R}^d, b_i \in \{+1, -1\}\)

- regularized logistic regression

\[
\text{minimize} \quad \frac{1}{n} \sum_{i=1}^{n} \log(1 + \exp(-b_i a_i^T x)) + \frac{\lambda_2}{2} \|x\|_2^2 + \lambda_1 \|x\|_1
\]

nonsmooth term \(\|x\|_1\) handled by proximal gradient methods

- data sets and characteristics:

<table>
<thead>
<tr>
<th>data sets</th>
<th>(n)</th>
<th>(d)</th>
<th>(\lambda_2)</th>
<th>(\lambda_1)</th>
</tr>
</thead>
<tbody>
<tr>
<td>rcv1</td>
<td>20,242</td>
<td>47,236</td>
<td>(10^{-4})</td>
<td>(10^{-5})</td>
</tr>
<tr>
<td>covertype</td>
<td>581,012</td>
<td>54</td>
<td>(10^{-5})</td>
<td>(10^{-4})</td>
</tr>
<tr>
<td>sido0</td>
<td>12,678</td>
<td>4,932</td>
<td>(10^{-4})</td>
<td>(10^{-4})</td>
</tr>
</tbody>
</table>

(thanks to Lin Xiao for the experiments)
SVRG on rcv1 dataset: varying step size $\eta$ with $m = 2n$
SVRG on rcv1 dataset with $\lambda_2 = 10^{-4}$ and stepsize $\eta = 0.1/L$: varying the period $m$ between full gradient evaluations
SVRG on rcv1 dataset with $\lambda_2 = 10^{-5}$ and stepsize $\eta = 0.1/L$: varying the period $m$ between full gradient evaluations
comparison with related algorithms on rcv1 datasets
comparison with related algorithms on covertype datasets
comparison with related algorithms on sido0 datasets
References

• A. Nemirovski, A. Juditsky, G. Lan and A. Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM Journal on Optimization (2009)


• R. Johnson and T. Zhang, Accelerating stochastic gradient descent using predictive variance reduction NIPS (2013)

• L. Xiao and T. Zhang, A proximal stochastic gradient method with progressive variance reduction, manuscript (2014)