Accelerate Subgradient Methods

Tianbao Yang

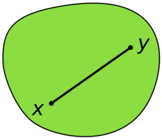
Department of Computer Science
The University of Iowa

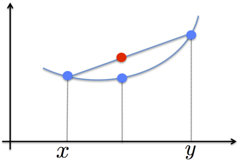
Contributors: students Yi Xu, Yan Yan and colleague Qihang Lin

Outline

Convex Optimization

$$f_* = \min_{\mathbf{w}\in\Omega} f(\mathbf{w})$$





- $\Omega \subseteq \mathbb{R}^d$ is a convex set
- $f(\mathbf{w})$ is a convex function:

Goal: For a sufficiently small $\epsilon > 0$, find a solution $\mathbf{w}$ such that

$$f(\mathbf{w}) - f_* \leq \epsilon$$

Applications

Machine Learning

- Classification and Regression

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}^\top \mathbf{x}_i, y_i) + \lambda R(\mathbf{w})$$

- training examples $(\mathbf{x}_i, y_i), i = 1, \ldots, n$, where $\mathbf{x}_i \in \mathbb{R}^d$, $y_i \in \{1, -1\}$ (classification) or $\mathbf{y}_i \in \mathbb{R}$ (regression)
- $\ell(z, y)$ convex loss function w.r.t z
- $R(\mathbf{w})$ is a regularizer

Applications in Machine Learning

- Classification and Regression

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}^\top \mathbf{x}_i, y_i) + \lambda R(\mathbf{w})$$

- Examples: SVM

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i \mathbf{w}^\top \mathbf{x}_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

- Examples: LASSO

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \lambda\|\mathbf{w}\|_1$$

- Many others (examples given later)

Applications in Machine Learning

- Classification and Regression

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}^\top \mathbf{x}_i, y_i) + \lambda R(\mathbf{w})$$

- Examples: SVM

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i\mathbf{w}^\top \mathbf{x}_i) + \frac{\lambda}{2}\|\mathbf{w}\|_2^2$$

- Examples: LASSO

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} (y_i - \mathbf{w}^\top \mathbf{x}_i)^2 + \lambda\|\mathbf{w}\|_1$$

- Many others (examples given later)

How to solve the Optimization Efficiently?

Concern: Running Time (RT)

Iterative Algorithm

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \Delta \mathbf{w}_t$$

RT $=$ #of Iterations $*$ per-iteration RT

Iteration Complexity

How many iterations $T(\epsilon)$ are needed in order to have $f(\widehat{\mathbf{w}}_T) - f_* \leq \epsilon$

How to solve the Optimization Efficiently?

Concern: Running Time (RT)

Iterative Algorithm

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \Delta\mathbf{w}_t$$

RT $=$ #of Iterations $*$ per-iteration RT

Iteration Complexity

How many iterations $T(\epsilon)$ are needed in order to have $f(\widehat{\mathbf{w}}_T) - f_* \leq \epsilon$

Iteration Complexity of Convex Optimization

The properties of the objective function

- smooth function: function is upper bounded by a quadratic function
- strongly convex: function is lower bounded by a quadratic function

Minimax Iteration Complexity

- smooth and strongly convex: $O(\log(1/\epsilon))$: Accel. Gradient Method
- smooth: $O(1/\sqrt{\epsilon})$: Accel. Gradient Method
- strongly convex: $O(1/\epsilon)$: SubGradient (SG) Method
- non-smooth and non-strongly convex: $O(1/\epsilon^2)$: SG

Iteration Complexity of Convex Optimization

The properties of the objective function

- smooth function: function is upper bounded by a quadratic function
- strongly convex: function is lower bounded by a quadratic function

Minimax Iteration Complexity

- smooth and strongly convex: $O(\log(1/\epsilon))$: Accel. Gradient Method
- smooth: $O(1/\sqrt{\epsilon})$: Accel. Gradient Method
- strongly convex: $O(1/\epsilon)$: SubGradient (SG) Method
- non-smooth and non-strongly convex: $O(1/\epsilon^2)$: SG

Non-smooth and Non-Strongly Convex Optimization

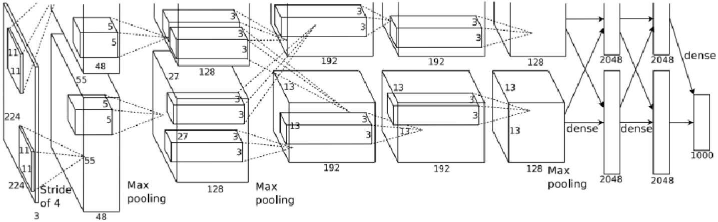
Robust Regression:

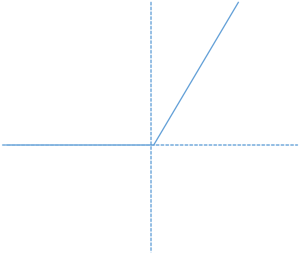
$$\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n}|\mathbf{w}^\top \mathbf{x}_i - y_i|^p, \quad p \in [1,2)$$

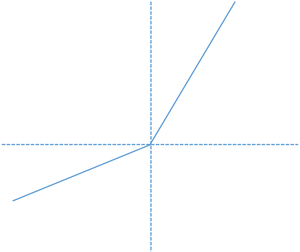
Sparse Classification:

$$\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y_i\mathbf{w}^\top \mathbf{x}_i) + \lambda\|\mathbf{w}\|_1$$

Deep Learning







Subgradient Method

$$f_* = \min_{\mathbf{w}\in\Omega} f(\mathbf{w})$$

SG Method

$$\mathbf{w}_{t+1} = \Pi_\Omega[\mathbf{w}_t - \eta_t \partial f(\mathbf{w}_t)]$$

- $\partial f(\mathbf{w}_t)$ is a subgradient
- η_t step size: $\eta_t \propto 1/\sqrt{t}$
- iteration complexity $O(1/\epsilon^2)$: very slow (e.g., $\epsilon = 10^{-5} \Rightarrow 10^{10}$ iterations)

Accelerate Subgradient Method

$$f_* = \min_{\mathbf{w}\in\Omega} f(\mathbf{w})$$

- Special structure/condition of the objective function
- One example is explicit max-strucutre

$$f(\mathbf{w}) = \max_{\mathbf{u}\in\Omega_2} \mathbf{u}^\top A\mathbf{w} - \phi(\mathbf{w})$$

- Nesterov's Smoothing technique (Nesterov, 2005)

$$F_\mu(\mathbf{w}) = \max_{\mathbf{u}\in\Omega_2} \mathbf{u}^\top A\mathbf{w} - \phi(\mathbf{w}) - \frac{\mu}{2}\|\mathbf{u}\|_2^2$$

- AG for the smoothed problem: $(1/\epsilon)$ for the original problem

Our Methodology for Accelerating SG

$$f_* = \min_{\mathbf{w}\in\Omega} f(\mathbf{w})$$

- Explore local structure around the optimal solution (local error bound)

$$\|\mathbf{w} - \mathbf{w}^*\|_2 \leq c(f(\mathbf{w}) - f_*)^{\theta}, \quad \theta \in (0, 1], \quad \forall \mathbf{w} \in \mathcal{S}_\epsilon$$

- This family is broad enough
- Improved Iteration Complexity $\widetilde{O}\left(\frac{1}{\epsilon^{2(1-\theta)}}\right)$
- With explicit max-structure: $\widetilde{O}\left(\frac{1}{\epsilon^{(1-\theta)}}\right)$

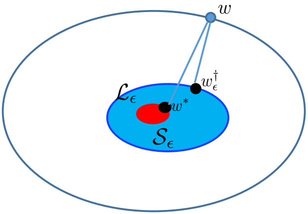
Outline

Some Notations

Notations

- Optimal Set $\Omega_* = \{\mathbf{w} \in \Omega : f(\mathbf{w}) = f_*\}$
- ϵ-level set $\mathcal{L}_\epsilon = \{\mathbf{w} \in \Omega : f(\mathbf{w}) - f_* = \epsilon\}$
- ϵ-sublevel set $\mathcal{S}_\epsilon = \{\mathbf{w} \in \Omega : f(\mathbf{w}) - f_* \leq \epsilon\}$
- $\mathbf{w}^*$: the closest optimal solution to $\mathbf{w}$

$$\mathbf{w}^* = \arg\min_{\mathbf{u}\in\Omega_*} \|\mathbf{u} - \mathbf{w}\|_2$$



- $\mathbf{w}_\epsilon^\dagger$: the closest solution to $\mathbf{w}$ in the ϵ-sublevel set

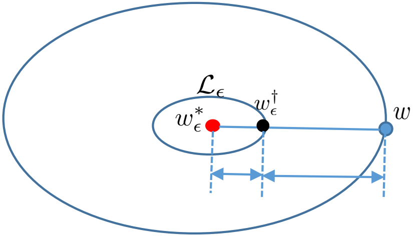
$$\mathbf{w}_\epsilon^\dagger = \arg\min_{\mathbf{u}\in\Omega_*} \|\mathbf{u} - \mathbf{w}\|_2, \quad f(\mathbf{w}) - f_* \leq \epsilon$$

Assumptions

Assumptions

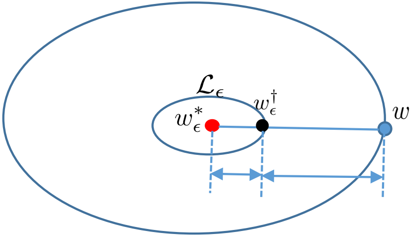
- there exists $\mathbf{w}_0, \epsilon_0$ such that $f(\mathbf{w}_0) - f_* \leq \epsilon_0$
- there exists G such that $\|\partial f(\mathbf{w})\|_2 \leq G$
- Ω_* is a non-empty compact set

An Key Error Inequaltiy



$$\frac{\|w_\epsilon^\dagger - w_\epsilon^*\|_2}{f(w_\epsilon^\dagger) - f(w_\epsilon^*)} \qquad \frac{\|w - w_\epsilon^\dagger\|_2}{f(w) - f(w_\epsilon^\dagger)}$$

An Key Error Inequaltiy



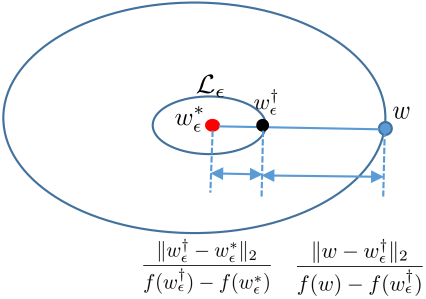
$$\frac{\|w_\epsilon^\dagger - w_\epsilon^*\|_2}{f(w_\epsilon^\dagger) - f(w_\epsilon^*)} \qquad \frac{\|w - w_\epsilon^\dagger\|_2}{f(w) - f(w_\epsilon^\dagger)}$$

$$\frac{\|w_\epsilon^\dagger - w_\epsilon^*\|_2}{f(w_\epsilon^\dagger) - f(w_\epsilon^*)} \geq \frac{\|w - w_\epsilon^\dagger\|_2}{f(w) - f(w_\epsilon^\dagger)}$$

Key Error Inequality (Yang & Lin, 2016)

$$\|w - w_\epsilon^\dagger\|_2 \leq \frac{\|w_\epsilon^\dagger - w_\epsilon^*\|_2}{\epsilon}(f(w) - f(w_\epsilon^\dagger))$$

An Key Error Inequaltiy



$$\frac{\|w_\epsilon^\dagger - w_\epsilon^*\|_2}{f(w_\epsilon^\dagger) - f(w_\epsilon^*)} \geq \frac{\|w - w_\epsilon^\dagger\|_2}{f(w) - f(w_\epsilon^\dagger)}$$

Key Error Inequality (Yang & Lin, 2016)

$$\|w - w_\epsilon^\dagger\|_2 \leq \frac{\|w_\epsilon^\dagger - w_\epsilon^*\|_2}{\epsilon}(f(w) - f(w_\epsilon^\dagger))$$

Outline

SG Method

$\widehat{\mathbf{w}}_T = \mathrm{SG}(\mathbf{w}_0, \eta, T)$

1: **Input**: the number of iterations T, and the initial solution $\mathbf{w}_0 \in \Omega$,
2: Let $\mathbf{w}_1 = \mathbf{w}_0$
3: **for** $t = 1, \ldots, T$ **do**
4: Compute a subgradient $\partial f(\mathbf{w}_t)$
5: Update $\mathbf{w}_{t+1} = \Pi_\Omega[\mathbf{w}_t - \eta \partial f(\mathbf{w}_t)]$
6: **end for**
7: **Output**: $\widehat{\mathbf{w}}_T = \sum_{t=1}^T \frac{\mathbf{w}_t}{T}$

Convergence Guarantee

For any $\mathbf{w} \in \Omega$

$$f(\widehat{\mathbf{w}}_T) - f(\mathbf{w}) \leq \frac{\eta G^2}{2} + \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2}{2\eta T}$$

$$T = \frac{G^2\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\epsilon^2}, \eta = \frac{G^2}{\epsilon} \Rightarrow f(\widehat{\mathbf{w}}_T) - f_* \leq \epsilon$$

RSG Method

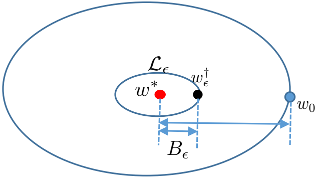
$\mathbf{w}_K = \mathrm{RSG}(\mathbf{w}_0, t, K)$

Input: the number of stages K and the number of iterations t per-stage, $\mathbf{w}_0 \in \Omega$
Set $\eta_1 = \epsilon_0/(2G^2)$, where ϵ_0 is from our assumption
for $k = 1, \ldots, K$ **do**
    Call subroutine SG to obtain $\mathbf{w}_k = \mathrm{SG}(\mathbf{w}_{k-1}, \eta_k, t)$
    Set $\eta_{k+1} = \eta_k/2$
end for
Output: $\mathbf{w}_K$

A General Convergence of RSG

Distance between the ϵ-level set and the optimal set

$$B_\epsilon = \max_{\mathbf{w}\in\mathcal{L}_\epsilon} \|\mathbf{w} - \mathbf{w}^*\|_2$$



Convergence of RSG

If $t \geq \frac{4G^2B_\epsilon^2}{\epsilon^2}$ and $K = \lceil \log_2(\frac{\epsilon_0}{\epsilon}) \rceil$, then $f(\mathbf{w}_K) - f* \leq 2\epsilon$

- Iteration Complexity of RSG: $O\left(\frac{G^2B_\epsilon^2}{\epsilon^2}\log\left(\frac{\epsilon_0}{\epsilon}\right)\right)$

Comparison with SG

Iteration Complexity of RSG: $O\left(\frac{G^2 B_\epsilon^2}{\epsilon^2}\log\left(\frac{\epsilon_0}{\epsilon}\right)\right)$

Iteration Complexity of SG: $O\left(\frac{G^2\|\mathbf{w}_0 - \mathbf{w}^*\|_2^2}{\epsilon^2}\right)$

- SG: dependence on the distance from the initial solution to the optimal set
- RSG: dependence on the distance from the ϵ-level set to the optimal set
- RSG: log-dependence on the quality of the initial solution (ϵ_0)

Improved Convergence under Local Error Bound Condition

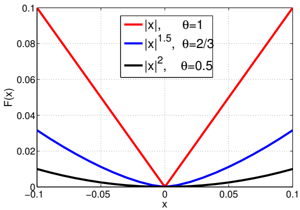
Iteration Complexity: $O\left(\frac{G^2 B_\epsilon^2}{\epsilon^2}\log\left(\frac{\epsilon_0}{\epsilon}\right)\right)$, where $B_\epsilon = \max_{\mathbf{w}\in\mathcal{L}_\epsilon}\|\mathbf{w}-\mathbf{w}^*\|_2$

Local Error Bound:

$$\|\mathbf{w}-\mathbf{w}^*\|_2 \leq c(f(\mathbf{w})-f_*)^\theta, \quad \theta\in(0,1], \quad \forall \mathbf{w}\in\mathcal{S}_\epsilon$$

Implies:

$$B_\epsilon \leq c\epsilon^\theta, \quad \text{and} \quad O\left(\frac{G^2 c^2}{\epsilon^{2(1-\theta)}}\log\left(\frac{\epsilon_0}{\epsilon}\right)\right)$$



Polyhedral Convex Optimization

Linear Convergence: epigraph is a polyhedron: $\theta = 1$

Examples:

- Robust Regression

$$\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} |\mathbf{w}^\top \mathbf{x}_i - y_i|$$

- Sparse Classification:

$$\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} \max(0, 1 - y_i\mathbf{w}^\top \mathbf{x}_i) + \lambda\|\mathbf{w}\|_1$$

Polyhedral Convex Optimization

Linear Convergence: epigraph is a polyhedron: $\theta = 1$

Examples:

- Robust Regression

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} |\mathbf{w}^\top \mathbf{x}_i - y_i|$$

- Sparse Classification:

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} \max(0, 1 - y_i \mathbf{w}^\top \mathbf{x}_i) + \lambda \|\mathbf{w}\|_1$$

Locally Semi-Strongly Convex Function

$$\|\mathbf{w} - \mathbf{w}_*\|_2 \leq c(f(\mathbf{w}) - f_*)^{1/2}, \quad \forall \mathbf{w} \in \mathcal{S}_\epsilon$$

Iteration Complexity: $\widetilde{O}(\frac{c^2 G^2}{\epsilon})$

Examples:

- Robust Regression

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} |\mathbf{w}^\top \mathbf{x}_i - y_i|^p, p \in (1, 2)$$

- ℓ_1 regularized problems: $h(\cdot)$ is strongly convex on any compact set

$$f(\mathbf{w}) = h(A\mathbf{w}) + \lambda\|\mathbf{w}\|_1$$

- Huber Loss:

$$\ell_\delta(z) = \begin{cases} \frac{1}{2}z^2 & \text{if } |z| \leq \delta \\ \delta(|z| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$$

Locally Semi-Strongly Convex Function

$$\|\mathbf{w} - \mathbf{w}_*\|_2 \leq c(f(\mathbf{w}) - f_*)^{1/2}, \quad \forall \mathbf{w} \in \mathcal{S}_\epsilon$$

Iteration Complexity: $\widetilde{O}(\frac{c^2 G^2}{\epsilon})$

Examples:

- Robust Regression

$$\min_{\mathbf{w}} \frac{1}{n} \sum_{i=1}^{n} |\mathbf{w}^\top \mathbf{x}_i - y_i|^p, p \in (1, 2)$$

- ℓ_1 regularized problems: $h(\cdot)$ is strongly convex on any compact set

$$f(\mathbf{w}) = h(A\mathbf{w}) + \lambda\|\mathbf{w}\|_1$$

- Huber Loss:

$$\ell_\delta(z) = \begin{cases} \frac{1}{2}z^2 & \text{if } |z| \leq \delta \\ \delta(|z| - \frac{1}{2}\delta) & \text{otherwise} \end{cases}$$

Why RSG Converges Faster

- Step size is decreasing in a stage-wise manner (sounds familar?)
- Warm-start

Proof of RSG

Proof is very simple given the key inequality

- $\epsilon_k = \frac{\epsilon_0}{2^k}$, thus $\eta_k = \frac{\epsilon_k}{G^2}$
- By induction: assume $f(\mathbf{w}_{k-1}) - f_* \leq \epsilon_{k-1} + \epsilon$
- Apply the convergence result of SG for the k-th stage

$$f(\mathbf{w}_k) - f(\mathbf{w}^\dagger_{k-1,\epsilon}) \leq \frac{\eta_k G^2}{2} + \frac{\|\mathbf{w}_{k-1} - \mathbf{w}^\dagger_{k-1,\epsilon}\|_2^2}{2\eta_k t}$$

  and our key error inequality

$$\|\mathbf{w}_{k-1} - \mathbf{w}^\dagger_{k-1,\epsilon}\|_2 \leq \frac{B_\epsilon}{\epsilon}(f(\mathbf{w}_{k-1}) - f(\mathbf{w}^\dagger_{k-1,\epsilon})) \leq \frac{B_\epsilon}{\epsilon}\epsilon_{k-1} = \frac{B_\epsilon}{\epsilon}2\epsilon_k$$

- Get

$$f(\mathbf{w}_k) - f(\mathbf{w}^\dagger_{k-1,\epsilon}) \leq \frac{\epsilon_k}{2} + \frac{\epsilon_k^2}{2\epsilon_k}\frac{4G^2B_\epsilon^2}{\epsilon^2 t} \leq \epsilon_k$$

Outline

Nesterov's smoothing improves SG

Explicit-max structure

$$f(\mathbf{w}) = \max_{\mathbf{u}\in\Omega_2} \mathbf{u}^\top A\mathbf{w} - \phi(\mathbf{w})$$

Nesterov's Smoothing technique

$$F_\mu(\mathbf{w}) = \max_{\mathbf{u}\in\Omega_2} \mathbf{u}^\top A\mathbf{w} - \phi(\mathbf{w}) - \frac{\mu}{2}\|\mathbf{u}\|_2^2$$

which is a $L_\mu = \frac{\|A\|^2}{\mu}$-smooth function.

Assume: $\max_{\mathbf{u}\in\Omega_2} \|\mathbf{u}\|_2 \leq D$

Nesterov's smoothing improves SG

Explicit-max structure

$$f(\mathbf{w}) = \max_{\mathbf{u}\in\Omega_2} \mathbf{u}^\top A\mathbf{w} - \phi(\mathbf{w})$$

Smoothing parameter

Nesterov's Smoothing technique

$$F_\mu(\mathbf{w}) = \max_{\mathbf{u}\in\Omega_2} \mathbf{u}^\top A\mathbf{w} - \phi(\mathbf{w}) - \frac{\mu}{2}\|\mathbf{u}\|_2^2$$

which is a $L_\mu = \frac{\|A\|^2}{\mu}$-smooth function.

Assume: $\max_{\mathbf{u}\in\Omega_2} \|\mathbf{u}\|_2 \leq D$

Accelerated Gradient Method for Smooth Function

$\mathbf{w}_T = \mathrm{AG}(\mathbf{w}_0, t, L_\mu)$

Input: the number of iterations t, and the initial solution $\mathbf{w}_0 \in \Omega$
Let $t_0 = t_{-1} = 1$, $\mathbf{w}_{-1} = \mathbf{w}_0$
for $k = 0, \ldots, t$ **do**
  Compute $\mathbf{y}_k = \mathbf{w}_k + \frac{t_{k-1}-1}{t_k}(\mathbf{w}_k - \mathbf{w}_{k-1})$
  Compute $\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}\in\Omega}\{\mathbf{w}^\top \nabla F_\mu(\mathbf{y}_k) + \frac{L_\mu}{2}\|\mathbf{w} - \mathbf{y}_k\|_2^2\}$
  Compute $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
end for
Output: $\mathbf{w}_{t+1}$

Convergence of AG

For any $\mathbf{w} \in \Omega$

$$f(\mathbf{w}_T) - f(\mathbf{w}) \leq \frac{\mu D^2}{2} + \frac{2\|A\|^2\|\mathbf{w}_0 - \mathbf{w}\|_2^2}{\mu T^2}$$

Accelerated Gradient Method for Smooth Function

$\mathbf{w}_T = \mathrm{AG}(\mathbf{w}_0, t, L_\mu)$

Input: the number of iterations t, and the initial solution $\mathbf{w}_0 \in \Omega$
Let $t_0 = t_{-1} = 1$, $\mathbf{w}_{-1} = \mathbf{w}_0$
for $k = 0, \ldots, t$ **do**
  Compute $\mathbf{y}_k = \mathbf{w}_k + \frac{t_{k-1}-1}{t_k}(\mathbf{w}_k - \mathbf{w}_{k-1})$
  Compute $\mathbf{w}_{k+1} = \arg\min_{\mathbf{w}\in\Omega}\{\mathbf{w}^\top \nabla F_\mu(\mathbf{y}_k) + \frac{L_\mu}{2}\|\mathbf{w} - \mathbf{y}_k\|_2^2\}$
  Compute $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
end for
Output: $\mathbf{w}_{t+1}$

Convergence of AG

For any $\mathbf{w} \in \Omega$

$$f(\mathbf{w}_T) - f(\mathbf{w}) \leq \frac{\mu D^2}{2} + \frac{2\|A\|^2\|\mathbf{w}_0 - \mathbf{w}\|_2^2}{\mu T^2}$$

approximation error

Homotopy Smoothing (HOPS)

HOPS (Xu et al., 2016b)

```
Input: the number of stages K and the number of iterations t per-stage,
and the initial solution w_0 ∈ Ω_1
Let μ_1 = ε_0/(2D^2)
for s = 1, ..., K do
    Let w_s = AG(w_{s-1}, t, L_{μ_s})
    Update μ_{s+1} = μ_s/2
end for
Output: w_K
```

Improved Convergence under Local Error Bound Condition

Local Error Bound:

$$\|\mathbf{w} - \mathbf{w}^*\|_2 \leq c(f(\mathbf{w}) - f_*)^{\theta}, \quad \theta \in (0, 1], \quad \forall \mathbf{w} \in \mathcal{S}_{\epsilon}$$

Implies:

$$\text{Iteration Complexity:} \quad O\left(\frac{cD\|A\|}{\epsilon^{(1-\theta)}} \log\left(\frac{\epsilon_0}{\epsilon}\right)\right)$$

Outline

Stochastic Subgradient (SSG) Method

$$\min_{\mathbf{w}\in\Omega} F(\mathbf{w}) \triangleq \mathrm{E}_\xi[f(\mathbf{w};\xi)]$$

SSG Method

$$\mathbf{w}_{t+1} = \Pi_\Omega[\mathbf{w}_t - \eta_t \partial f(\mathbf{w}_t;\xi_t)]$$

- $\eta_t \propto 1/\sqrt{t}$
- More scalable for large-scale problems
- Examples: Empirical Risk Minimization

$$\min_{\mathbf{w}} \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{w}^\top \mathbf{x}_i, y_i) + \lambda R(\mathbf{w})$$

Accelerated Stochastic Subgradient (ASSG) Method

Do the same tricks suffice?

- Step size is decreasing in a stage-wise manner
- Warm-start

Not enough in theory

$$\mathrm{E}[f(\widehat{\mathbf{w}}_T) - f(\mathbf{w})] \leq \frac{\eta G^2}{2} + \frac{\mathrm{E}[\|\mathbf{w}_0 - \mathbf{w}\|_2^2]}{2\eta T}$$

- Expectation bound does not work

Accelerated Stochastic Subgradient (ASSG) Method

Do the same tricks suffice?

- Step size is decreasing in a stage-wise manner
- Warm-start

Not enough in theory

$$\mathrm{E}[f(\widehat{\mathbf{w}}_T) - f(\mathbf{w})] \leq \frac{\eta G^2}{2} + \frac{\mathrm{E}[\|\mathbf{w}_0 - \mathbf{w}\|_2^2]}{2\eta T}$$

- Expectation bound does not work

Accelerated Stochastic Subgradient (ASSG) Method

Use high-probability analysis

$$f(\widehat{\mathbf{w}}_T) - f(\mathbf{w}) \leq \frac{\eta G^2}{2} + \frac{\|\mathbf{w}_0 - \mathbf{w}\|_2^2}{2\eta T} + \frac{\sum_{t=1}^{T} (\text{Var of SSG})_t \|\mathbf{w}_t - \mathbf{w}\|_2}{T}$$

- Another Trick: Domain Shrinking (Xu et al., 2016a)

Accelerated Stochastic Subgradient (ASSG) Method

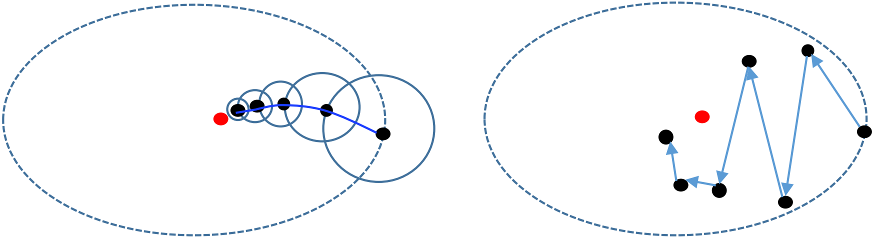
in the k-stage of SSG

$$\mathbf{w}_{t+1}^k = \Pi_{\Omega \cap \mathcal{B}(\mathbf{w}_{k-1}, D_k)}[\mathbf{w}_t^k - \eta_k \partial f(\mathbf{w}_t^k; \xi_t^k)]$$

- $\mathcal{B}(\mathbf{w}, D)$ is a ball centered at **w** with radius D
- D_k is decreasing by half every stage
- In a high probability $1-\delta$, iteration complexity is $\widetilde{O}\left(\frac{c^2 G^2 \log(1/\delta)}{\epsilon^{2(1-\theta)}}\right)$

Domain Shrinking

Mitigates variance in stochastic subgradient



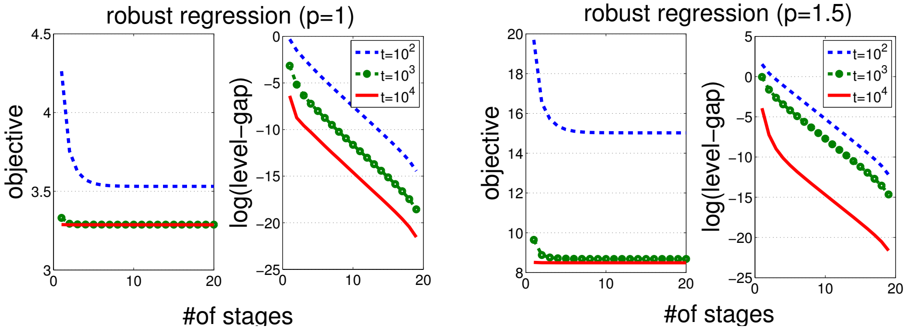
ASSG vs SSG

Outline

RSG: Convergence

robust regression

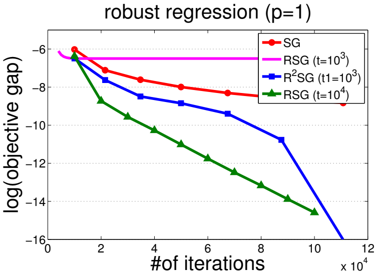
$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} |\mathbf{w}^\top \mathbf{x}_i - y_i|^p, \quad p = 1, \quad p = 1.5$$

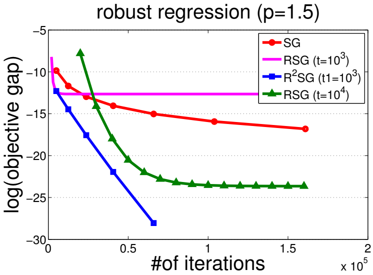


RSG vs SG

robust regression

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n} |\mathbf{w}^\top \mathbf{x}_i - y_i|^p, \quad p = 1, \quad p = 1.5$$

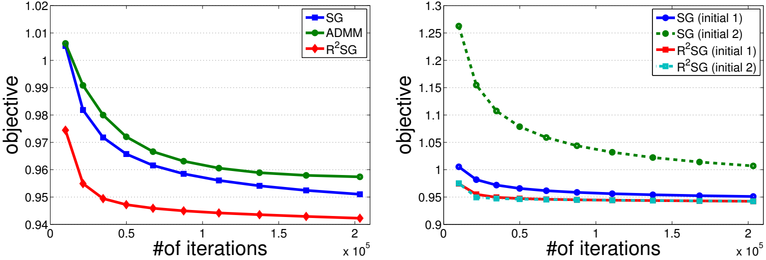




RSG vs SG

Graph-lasso regularized SVM

$$\min_{\mathbf{w}\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^{n}\max(0, 1 - y_i\mathbf{w}^\top\mathbf{x}_i) + \lambda\|F\mathbf{w}\|_1$$



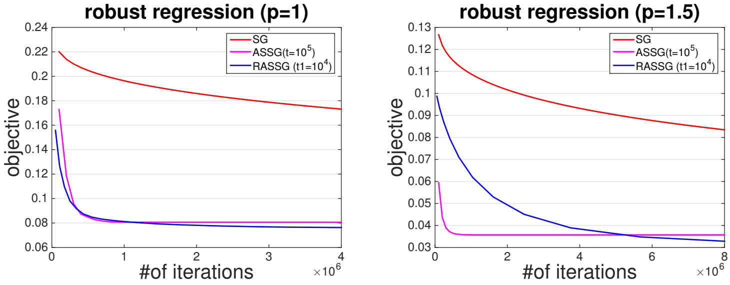
HOPS vs Smoothing

Table: Comparison of different optimization algorithms by the number of iterations and running time for achieving a solution that satisfies $F(x) - F_* \leq \epsilon$.

	Sparse Classification		Matrix Decomposition	
	$\epsilon = 10^{-4}$	$\epsilon = 10^{-5}$	$\epsilon = 10^{-3}$	$\epsilon = 10^{-4}$
PD	526 (4.48s)	1777 (13.91s)	2523 (7.57s)	3441 (9.82s)
APG-D	591 (5.63s)	1122 (10.61s)	1967 (10.30s)	8622 (44.90s)
APG-F	573 (5.12s)	943 (8.51s)	1115 (3.25s)	4151 (11.82s)
HOPS-D	501 (4.39s)	873 (8.35s)	224 (1.54s)	313 (2.16s)
HOPS-F	490 (4.38s)	868 (7.81s)	230 (0.90s)	312 (1.23s)
PD-HOPS	427 (3.41s)	609 (4.87s)	124 (0.48s)	162 (0.66s)

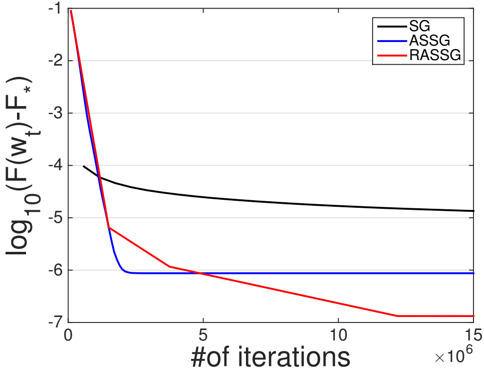
ASSG vs SSG

Million songs data ($n = 463,715$)



ASSG vs SSG

Hinge loss + ℓ_1 regularizer, Covtype data ($n = 581,012$)



Outline

Conclusion

- Developed a key error inequality
- that can accelerate many algorithms
- include stochastic momentum methods, stochastic Nesterov's accelerated gradient methods (Yang et al., 2016)
- Developed a restarted subgradient (RSG) method that has faster convergence than SG
- Developed a homotopy smoothing (HOPS) algorithm with even faster convergence
- Developed an accelerated stochastic subgradient (ASSG) method
- Preliminary Experiments show very promising results

THANK YOU!

QUESTIONS?

References I

Nesterov, Yu. Smooth minimization of non-smooth functions. *Mathematical Programming*, 103(1):127–152, 2005.

Xu, Yi, Lin, Qihang, and Yang, Tianbao. Accelerate stochastic subgradient method by leveraging local error bound. *CoRR*, abs/1607.01027, 2016a.

Xu, Yi, Yan, Yan, Lin, Qihang, and Yang, Tianbao. Homotopy smoothing for non-smooth problems with lower complexity than 1/epsilon. *CoRR*, abs/1607.03815, 2016b.

Yang, Tianbao and Lin, Qihang. Rsg: Beating sgd without smoothness and/or strong convexity. *CoRR*, abs/1512.03107, 2016.

Yang, Tianbao, Lin, Qihang, and Li, Zhe. Unified convergence analysis of stochastic momentum methods for convex and non-convex optimization. *coRR*, 2016.