November 3, 2010

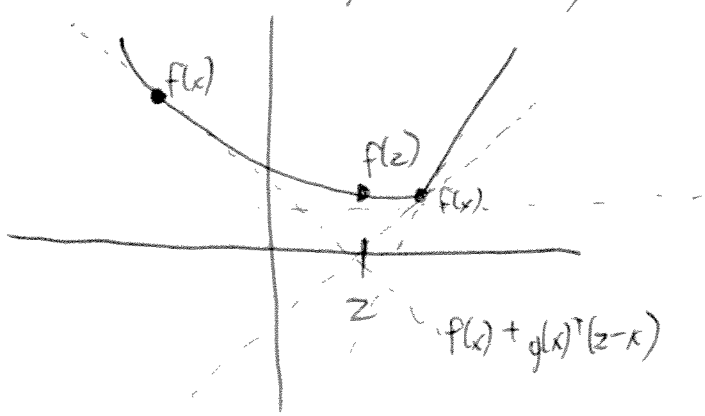
Outline:
1. Iteration Complexity, Assumptions
2. First-Order Complexity Zoo
3. Optimality of a stochastic method
4. Amplification

1. Iteration Complexity and Assumptions

Problem: $\min_x f(x)$, where $f(x)$ is convex (not necessarily differentiable)

A vector $g(x)$ is a sub-gradient of f at x if

$$f(z) \geq f(x) + g(x)^T (z-x), \quad \forall z$$



$\partial f(x)$: set of sub-gradients at x.
$\partial f(x)$ is always non-empty, if f is differentiable at x then $\partial f(x) = \{\nabla f(x)\}$

We are given a first-order oracle

Deterministic Oracle	Stochastic Oracle

On iteration k, algorithm receives:

- objective $f(x^k)$	- noisy objective $F(x^k) = f(x^k) + \omega^k$
- sub-gradient $g(x^k) \in \partial f(x^k)$	- noisy gradient $G(x^k) = g(x^k) + \xi^k$
	where $E[\omega^k] = 0$, $E[\xi^k] = 0$

Deterministic	Stochastic

In terms of ε, how many iterations before:

$\min_K f(x^K) - f(x^*) \leq \varepsilon$	$\min_K E[f(x^K)] - f(x^*) \leq \varepsilon$

For example, we might have $K = O(1/\varepsilon^2)$.

We need some assumptions to get this type of bound, such as $f(x^*) > -\infty$.

A1 (Bounded sub-gradient): There exists an M such that

$M \geq \sup_x \|g(x)\|_2$	$M^2 \geq \sup_x E[\|G(x)\|_2^2]$

only needs to hold on a compact set.

In some cases, we get better rates using quadratic bounds.

E.g. Assume f is twice-differentiable and for all x, $c \leq \text{eigs}(\nabla^2 f(x)) \leq L$, for $c > 0$.

By Taylor expansion:

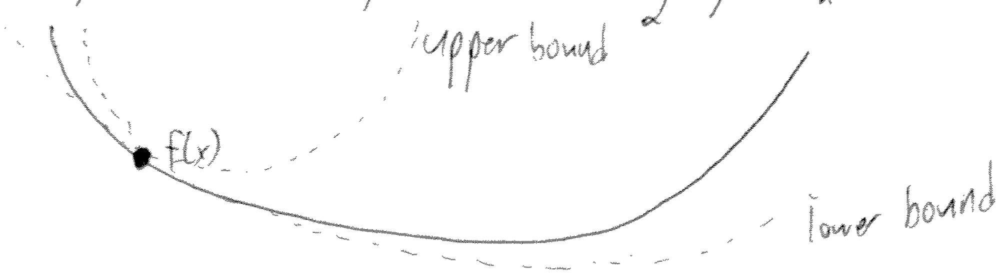
$\forall x, y$: $f(y) = f(x) + (y-x)^T \nabla f(x) + \frac{1}{2}(y-x)^T \nabla^2 f(z)(y-x)$, for some z.

$\cancel{(A)\ f(y) \leq f(x) + (y-x)^T \nabla f(x) + \frac{L}{2}(y-x)^T \nabla^2 f(}$

Then (by spectral decomp.):

(A) $f(y) \leq f(x) + (y-x)^T \nabla f(x) + \frac{L}{2}\|y-x\|_2^2$

(B) $f(y) \geq f(x) + (y-x)^T \nabla f(x) + \frac{c}{2}\|y-x\|_2^2$



We can get (A) and (B) under weaker assumptions:

(A) Gradient of differentiable f is <u>Lipschitz-continuous</u> if

$$|\nabla f(x) - \nabla f(y)| \leq L \|x-y\|.$$

Implies (A). <u>Weak</u> assumption (given differentiability) on a compact set.

(B) A function is <u>strongly convex</u> if $f(x) - \frac{\epsilon}{2}\|x\|_2^2$ is convex.

(strongly convex $\Rightarrow$ strictly convex $\Rightarrow$ convex)

Implies (B) for differentiable functions, but differentiability is <u>not</u> required for strong convexity.

2. <u>First-Order Complexity Zoo</u>

We assume f is convex, there exists x^*, and bounded sub-gradients.

Translation from error on iteration k to number of iterations:

$$O(1/\sqrt{k}) \Rightarrow O(1/\epsilon^2)$$
$$O\left(\frac{\log k}{k}\right) \Rightarrow \tilde{O}(1/\epsilon)$$
$$O(1/k) \Rightarrow O(1/\epsilon)$$
$$O(1/k^2) \Rightarrow O(1/\sqrt{\epsilon})$$
$$1/\exp(O(k)) \Rightarrow O(\log(1/\epsilon))$$

Assumptions/Method	Deterministic	Stochastic
none/sub-gradient	$O(1/\varepsilon^2)$	$O(1/\varepsilon^2)$
Lipschitz/gradient	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$
smoothed to Lipschitz/Nesterov	$O(1/\varepsilon)$	$O(1/\varepsilon^2)$
strongly/sub-gradient	$\tilde{O}(1/\varepsilon)$	$\tilde{O}(1/\varepsilon)$
strongly/epoch averaging	$O(1/\varepsilon)$	$O(1/\varepsilon)$
Lipschitz/Nesterov	$O(1/\sqrt{\varepsilon})$	$O(1/\varepsilon^2)$
Lipschitz +strongly/gradient	$O(\log(1/\varepsilon))$	$O(1/\varepsilon)$
* Lipschitz+strongly+"steps converg"/ Barzilai-Borwein	$o(\log(1/\varepsilon))$	N/A

we prove this one in next section

Notes:

- Lipschitz does not help in stochastic case
- Without Lipschitz, no difference between deterministic and stochastic (so we stochastic)
- Polyak-Ruppert averaging does not give better rates, but can achieve the same rates with a more robust strategy.
- Many of these results are tight, no "first-order" method can do better by more than a constant.
- There is a second-order complexity zoo with faster rates like $O(1/\sqrt{\varepsilon})$ for Lipschitz-Hessian and $O(\log\log(1/\varepsilon))$ for Lipschitz and strongly convex

- We can re-write (sub-)gradient method as:

$$(*) \quad x^{k+1} \leftarrow \underset{x}{\operatorname{argmin}}\ f(x^k) + (x-x^k)^T g(x^k) + \alpha^k \underbrace{\frac{1}{2}\|x-x^k\|_2^2}_{D(x,x^k)}$$

- Mirror Descent: replace $D(x,x^k)$ with another Bregman distance, get similar rates.
  (e.g. if x is a probability and use Kullback-Leibler divergence, get convergence rates for exponentiated gradient variants)

- Composite Objectives / Proximal-Splitting: add extra convex term $r(x)$ to $(*)$ and solve, get the convergence rate of $f(x)$ even if $r(x)$ doesn't satisfy the same assumptions.
  (e.g. if $r(x) = \lambda\|x\|_1$, get convergence rates for iterative soft-thresholding variants)

- No results for MCMC, Kiefer-Wolfowitz, or other biased stochastic methods.

3. Optimality of a stochastic method (following Nemirovsky et al., 2009).

Problem: $\min_x f(x)$, given first-order stochastic oracle.

Assume: strongly convex, Lipschitz gradient, bounded sub-gradient

Algorithm: $\alpha^k \leftarrow \frac{\theta}{k}$, for some $\theta > \frac{1}{2c}$

$$x^{k+1} \leftarrow x^k - \alpha^k G(x^k)$$

This simple algorithm has the 'optimal' expected error of $\epsilon = O(1/k)$

Outline of proof:

1. Express distance from x^{k+1} to x^* in terms of x^k.
2. Use bounded sub-gradient and strong convexity to bound expectation.
3. Use induction to get rate of convergence.
4. Use Lipschitz to bound function value.

Notation: $D^k = \frac{1}{2}\|x^k - x^*\|_2^2$, $d^k = E[D^k]$

1. $D^{k+1} = \frac{1}{2}\|x^k - x^*\|_2^2 = \frac{1}{2}\|(x^k - \alpha^k G(x^k)) - x^*\|_2^2$

$$= \frac{1}{2}\|(x^k - x^*) - \alpha^k G(x^k)\|_2^2$$

$$= \frac{1}{2}\|x^k - x^*\|_2^2 - \alpha^k (x^k - x^*) G(x^k) + \frac{1}{2}(\alpha^k)^2 \|G(x^k)\|_2^2$$

2. Bound expected value

by definition

by definition

by tower property, linearity of expectation, definition of $G(x^k)$, differentiability of f

by bounded sub-gradient

$$d^{k+1} \leq d^k - \alpha^k E[(x^k - x^*)^T \nabla f(x^k)] + \frac{1}{2}(\alpha^k)^2 M^2$$

$$= E[(x^k - x^*)^T(\nabla f(x^k) - \overset{=0}{\nabla f(x^*)})]$$

$$\geq c\, E[\|x^k - x^*\|_2^2] \quad \text{(by strong convexity)}$$

$$= 2cd^k$$

Use this and definition of α^k:

$$d_{k+1} \leq d_k - \frac{\theta}{k}(2cd^k) + \frac{1}{2}\frac{\theta^2}{k^2}M^2$$

3. "By induction": $d_k \leq \frac{\beta}{k}$, for $\beta = \max\{a_1, \frac{1}{2}\frac{\theta^2 M^2}{(2c\theta - 1)}\}$

(implies convergence in parameter values is $O(1/\sqrt{k})$, similar to asymptotic normality arguments).

4. By Lipschitz: $f(x) \leq f(x^*) + \frac{L}{2}\|x^* - x\|_2^2, \; \forall x$ (since $\nabla f(x^*) = 0$),
So:

$$f(x^k) - f(x^*) \leq \frac{L}{2}\|x^* - x^k\|_2^2$$

take expectation $\downarrow$ $\qquad$ $\downarrow$ by definition

$$E[f(x^k) - f(x^*)] \leq L d^k \leq \frac{L\beta}{k} \quad \square \text{ QED}$$

4. Amplification

- On a given run, the method may do worse than its expectation.
- Can we make sure it doesn't do "too badly"?
- $f(x^k) - f(x^*)$ is a non-negative random variable, use Markov's inequality to bound probability of deviation from expectation.

Recall: $P(X \geq a) \leq \frac{E[X]}{a}$,

Take $a = 2E[X]$ to get:

$$Pr\{f(x^k) - f(x^*) \geq 2E[f(x^k) - f(x^*)]\} \leq 1/2$$

If we run it: twice: $Pr\{\cdots\} \leq 1/4$
thrice: $Pr\{\cdots\} \leq 1/8$
$\vdots$
$\log(1/\delta)$ times: $Pr\{\cdots\} \leq \delta$

- We need $k \log(1/\delta)$ iterations to be within a constant of the bound with probability $1-\delta$.

8

References

A. Agarwal, P. Bartlett, P. Ravikumar, and M. Wainwright. Information-theoretic lower bounds on the oracle complexity of convex optimization. *NIPS*, 1050:3, 2009.

S. Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

D. Bertsekas. Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey. 2010.

D. Bertsekas. *Convex optimization theory*. Athena Scientific, 2009.

D. Bertsekas and J. Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

D. Bertsekas, A. Nedic, and A. Ozdaglar. *Convex analysis and optimization*. Athena Scientific, 2003.

S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge Univ Pr, 2004.

N. Cesa-Bianchi, P. Long, and M. Warmuth. Worst-case quadratic loss bounds for on-line prediction of linear functions by gradient descent. In *IEEE Transactions on Neural Networks*, 1993.

H. Chen, L. Guo, and A. Gao. Convergence and robustness of the Robbins-Monro algorithm truncated at randomly varying bounds. *Stochastic Processes and their Applications*, 27:217–231, 1987.

M. Collins, A. Globerson, T. Koo, X. Carreras, and P. Bartlett. Exponentiated gradient algorithms for conditional random fields and max-margin markov networks. *The Journal of Machine Learning Research*, 9:1775–1822, 2008.

A. Flaxman, A. Kalai, and H. McMahan. Online convex optimization in the bandit setting: gradient descent without a gradient. In *Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 385–394. Society for Industrial and Applied Mathematics, 2005.

E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

E. Hazan and S. Kale. An optimal algorithm for stochastic strongly-convex optimization. Arxiv, 2010.

A. Kalai and S. Vempala. Geometric algorithms for online optimization. In *Journal of Computer and System Sciences*, 2002.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *The Annals of Mathematical Statistics*, 23(3):462–466, 1952.

J. Kivinen and M. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

Y. Narushima, T. Wakamatsu, and H. Yabe. Extended Barzilai-Borwein method for unconstrained minimization problems. *Pacific Journal of Optimizatin*, 2010.

A. Nedic and D. Bertsekas. Convergence rate of incremental subgradient algorithms. *Stochastic Optimization: Algorithms and Applications*, pages 263–304, 2000.

A. Nemirovski. Efficient methods in convex programming. *Lecture notes*, 1994.

A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

Y. Nesterov. *Introductory lectures on convex optimization: A basic course*. Springer Netherlands, 2004.

Y. Nesterov. Accelerating the cubic regularization of Newtons method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

B. Polyak and A. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30:838, 1992.

9

H. Robbins and S. Monro. A stochastic approximation method. *The Annals of Mathematical Statistics*, pages 400–407, 1951.

N. Schraudolph, J. Yu, and S. Günter. A stochastic quasi-Newton method for online convex optimization. 2007.

S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In *Proceedings of the 24th international conference on Machine learning*, page 814, 2007.

J. Spall. *Introduction to stochastic search and optimization: estimation, simulation, and control*. John Wiley and Sons, 2003.

P. Sunehag, J. Trumpf, A. Canberra, and S. Vishwanathan. Variable Metric Stochastic Approximation Theory. *AISTATS*, 2009.

M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. 2003.