Complexity of Stochastic Programming

Anatoli Juditsky

University Grenoble Alpes

MAS2016 August 30, 2016

Problem of interest statement

Uncertain optimization problem

Consider the following model of an uncertain optimization problem:

$$\mathrm{Opt} = \min_{x \in X} \left\{ f(x) := E_P[F(x, \xi)] \right\} \qquad (S)$$

where

- x is the decision variable
- ξ is the random perturbation, $\xi \sim P$, $\xi \in \Xi \subset \mathbb{R}^d$
- X is the feasible set of (S), a subset of $\mathbb{R}^n$
- $F : (X, \Xi) \to \mathbb{R}$: uncertain objective

- We suppose that $F(x, \cdot)$ is measurable and P-integrable for all $x \in X$.
- The distribution P may be known or unknown...

Example

Statistical estimation and learning as stochastic optimization

- We are given an i.i.d. sample $\xi_1, \ldots, \xi_N$ from the unknown distribution P_{θ_*} known to belong to a family $\{P_\theta,\ \theta \in \Theta\}$.
- Finding a (maximum likelihood, contrast, etc) estimation $\widehat{\theta}$ of θ_* amounts to solving

  $$\min_{\theta \in \Theta} \left\{ L(\theta) := E_{P_{\theta_*}}[\ell(\theta, \xi)] \right\}$$

  given the sample $\xi_1, \ldots, \xi_N$.

- Distribution P_{θ_*} is unknown, but a sample from P_{θ_*} is available

- Observation: *limits of performance of algorithms for solving* (S) *are closely related to performance limits of estimation procedures*

Example

2-stage stochastic programs with recourse

- Newsvendor (simple inventory) problem [2] Let
  - $x \geq 0$ be the inventory level – purchased newspaper stock
  - $\xi \sim P$ be the random day demand for the newspaper
  - p be the sale price, and q the purchase price

The newsvendor profit is

$$p \min\{x, \xi\} - qx,$$

so, maximizing the expected profit amounts to solving (S) with

$$F(x, \xi) = qx - p \min\{x, \xi\}.$$

- If the distribution P is known, then "explicit solution" is available to (S):

$$x_* = P^{-1}\left(\frac{p-q}{q}\right).$$

- ... in the setting with unknown P, an i.i.d. sample $\xi_1, \ldots, \xi_N$ from P is available.

Example: 2-stage linear stochastic program with recourse [3, 7, 21]

$$X = \{x : Ax = b,\, x \geq 0\},\ F(x, \xi) = c^T x + Q(x, \xi),$$

where

$$Q(x, \xi) := \min_{y \geq 0} q^T y, \ \text{ subject to } \ Tx + Wy \geq h,$$

with $\xi = [q, h, T, W]$.

- It is typically assumed that "recourse is complete", meaning that the auxiliary problem is feasible for all $x \in X$ and all $\xi \in \Xi$.
- *The case of incomplete recourse is difficult – even deciding if a given 1-stage decision x results in an (a.s.) feasible second-stage problem is usually hard.*

Complexity of Convex Stochastic Programming

We consider only convex programs (S), i.e. such that

- $X \subset \mathbb{R}^n$ is a convex, bounded and closed set
- $f(x)$ is convex

We assume that

- function $F(x, \xi)$ is given explicitly, so that we can compute efficiently its value (and perhaps the derivatives in x) at every given pair $(x, \xi) \in X$
- we can sample from P, that is, generate a sample ξ_1, ξ_2, ... of independent realizations of ξ.

Note that our model is black-box. Thus, our conclusions will not concern the situation where the distribution P is simple and is given in advance. On the other hand, on can show [8] that it is difficult to solve to high accuracy already two-stage programs with easy-to-describe distributions.

Defining complexity

Our objective is to solve (S) to accuracy $\epsilon > 0$ with reliability $1 - \alpha$, i.e.
• being given N realizations $[\xi_i, ..., \xi_N]$, exhibit an approximate solution $\widehat{x}_N$ which satisfies

$$\mathrm{Prob}\{f(\widehat{x}_N) \le f_* + \epsilon\} \le 1 - \alpha$$

where f_* is the optimal value of (S).

• Let $\mathcal{S}$ be a class of stochastic programs and let $\mathcal{M}$ be a method for solving problems from $\mathcal{S}$.
• Let $N(\mathcal{M}, S)$ be the smallest N such that given a sample $\xi^N = [\xi_1, ..., \xi_N]$ of size N, $\mathcal{M}$ is capable of *solving* $S \in \mathcal{S}$ *to accuracy* ϵ *with reliability* $1 - \alpha$.
• We denote

$$N(\mathcal{M}, \mathcal{S}) = \sup_{S \in \mathcal{S}} N(\mathcal{M}, S)$$

the complexity of $\mathcal{M}$ on the class $\mathcal{S}$.

• Finally, we define the complexity of $\mathcal{S}$ as complexity of the "best" method – the value

$$N(\mathcal{S}) = \inf_{\mathcal{M}} N(\mathcal{M}, \mathcal{S}).$$

Lower bound for Lipschitz-continuous $F(\cdot, \xi)$

Given $L, D > 0$ and $0 < \epsilon < \frac{LD}{2}$, consider the pair of stochastic programs

$$\min_{x \in [-D/2, D/2]} \{f_\kappa(x) := E_{P_\kappa}[x\xi] \qquad (S_\kappa)$$

indexed with $\kappa = \pm 1$, with P_κ supported on $\{-L, L\}$, and such that

$$P_1\{-L\} = \frac{1}{2} - \gamma, \;\; P_1\{L\} = \frac{1}{2} + \gamma,$$
$$P_{-1}\{-L\} = \frac{1}{2} + \gamma, \;\; P_{-1}\{L\} = \frac{1}{2} - \gamma$$

with $\gamma = \frac{\epsilon}{LD}$.

We claim that any algorithm capable of solving (S_1) and (S_{-1}) to accuracy ϵ and reliability $1 - \alpha > 7/8$ requires the sample size N to satisfy

$$N \geq \frac{D^2 L^2}{\epsilon^2} \ln\left(\frac{2}{\alpha}\right).$$

- Of course,

$$f_1(x) = 2\epsilon D^{-1} x, \;\; \text{and} \;\; f_{-1}(x) = -2\epsilon D^{-1} x,$$

thus

$$f_{1,*} = f_1(-D/2) = f_{-1}(D/2) = f_{-1,*} = -\epsilon,$$

while $f_1(x) \geq f_{1,*} + \epsilon = 0$ for $x \geq 0$, and $f_{-1}(x) \geq f_{-1,*} + \epsilon = 0$ for $x \leq 0$.

• Let us consider the problem of testing the hypotheses

$$H_1 : \ \kappa = 1 \ \text{ vs } \ H_1 : \ \kappa = -1.$$

I claim that "by the laws of statistics", one cannot decide upon H_1 and H_2 with the risk (sum of error probabilities) less than β given the sample ξ^N, unless $N \geq \frac{D^2L^2}{\epsilon^2} \ln\left(\frac{4}{\beta}\right)$.

• On the other hand, let $\widehat{x}_N$ be an approximate solution to (S_κ) using an N-sample ξ^N. We can associate with $\widehat{x}_N$ a test $T(\cdot)$ of H_1 vs H_{-1} as follows:

$$T(\xi^N) = -\mathrm{sign}(\widehat{x}_N).$$

Note that, by the above,

$$\alpha \geq \max_{\kappa=\pm 1} \mathrm{Prob}_{\xi_1 \sim P_\kappa}\left\{\mathrm{sign}(\widehat{x}_N) + \kappa \neq 0\right\} = \max_{\kappa=\pm 1} \mathrm{Prob}_{\xi_1 \sim P_\kappa}\left\{T(\xi^N) \neq \kappa\right\} \geq \beta/2.$$

We conclude that

$$\max_{\kappa=\pm 1} \mathrm{Prob}\left\{f_\kappa(\widehat{x}_N) - f_{\kappa,*} \leq \epsilon\right\} \geq 1 - \alpha$$

implies that

$$N \geq \frac{D^2L^2}{\epsilon^2} \ln\left(\frac{4}{\beta}\right) \geq \frac{D^2L^2}{\epsilon^2} \ln\left(\frac{2}{\alpha}\right).$$

Of "laws of statistics..."

Recall that the risk β of any test[1)] is bounded from below by the test affinity $\bar\beta_N$ of distributions P_{-1}^N and P_1^N of ξ^N

$$\beta \geq \bar\beta_N = \sum_{\mu=\{\pm L\}^N} \min\left[P_1\{\mu\}P_{-1}\{\mu\}\right].$$

In its turn, $\bar\beta_N$ can be easily bounded from below using the Hellinger affinity ρ_N of these distributions. Indeed, one has [5]

$$\bar\beta_N \geq 4\rho_N^2 = 4\rho^{2N}$$

where ρ is the Hellinger affinity of P_{-1} and P_1:

$$\rho = \sqrt{P_1\{-L\}P_{-1}\{-L\}} + \sqrt{P_1\{L\}P_{-1}\{L\}} = 2\sqrt{\tfrac{1}{4} - \gamma^2}.$$

We conclude that

$$\bar\beta_N \geq 4(1-4\gamma^2)^N = 4\left(1 - \frac{4\epsilon^2}{L^2D^2}\right)^N,$$

and

$$\frac{N\epsilon^2}{L^2D^2} \geq \ln\left(\frac{4}{\bar\beta_N}\right) \geq \ln\left(\frac{4}{\beta}\right).$$

[1)] *In fact, by the Neyman-Pearson lemma, the smallest risk is attained by the likelihood ratio T_*. In the case in question this test is simply the majority vote:*

$$T(\xi^n) = 2I\{2K \geq N\} - 1.$$

We have proved the following

Theorem 1 *Let $\mathcal{S}_1(D, L)$ be a class of convex stochastic programs such that*

- *$X \subset \mathbb{R}$ is a segment of length $D > 0$*
- *function $F(\cdot, \xi)$ is linear – $F(x, \xi) = \xi x$ and $|\xi| \leq L$.*

Let $\mathcal{M}$ be an algorithm capable of solving all programs from $\mathcal{S}_1(D, L)$ to accuracy ϵ with reliability $1 - \alpha$ using the sample ξ^N.

Then there is a problem $S \in \mathcal{S}_1(D, L)$ such that $\mathcal{M}$ will require a sample of length

$$N \geq \frac{D^2 L^2}{\epsilon^2} \ln\left(\frac{2}{\alpha}\right)$$

to output the solution.

In other words, the complexity $N(\mathcal{S}_1)$ of S_1 is below bounded by $\frac{D^2 L^2}{\epsilon^2} \ln\left(\frac{2}{\alpha}\right)$.

Theorem 1 allows for an immediate n-dimensional extension:

Theorem 2 [15][2)] *Let $\mathcal{S}(D, L)$ be a class of convex stochastic programs such that*

- *$X \subset \mathbb{R}^n$ contains a Euclidean ball of diameter $D > 0$*
- *function $F(\cdot, \xi)$ is Lipschitz-continuous:*

$$|F(x, \xi) - F(x', \xi)| \leq L\|x - x'\|_2, \ \ \forall \xi \in \Xi, \ \forall x, x' \in X.$$

Then the complexity $N(\mathcal{S})$ of the class $\mathcal{S}(D, L)$ of Lipschitz stochastic programs satisfies

$$N(\mathcal{S}) \geq \frac{D^2 L^2}{\epsilon^2} \ln\left(\frac{2}{\alpha}\right).$$

[2)] Usually, the Lipschitz constant L is replaced with "standard deviation" σ of the stochastic subgradient $F'_x(x, \xi)$:

$$\sigma^2 = E_P[\|F'_x(x, \xi) - f'(x)]\|_2^2]$$

Note that $\sigma^2 = L^2(1 - 4\gamma^2)$ in our simple construction.

Observations

- Difficulty of solving stochastic programs depend on the amplitude of the random subgradient F'_x and the size of the problem domain
- One cannot expect solving stochastic programs to "high accuracy" – 1-5% relative accuracy seems to be the attainable limit in many "practical" applications
- One cannot expect finding approximate solution $\widehat{x}$ which is close to the optimal set X_* of (S) – this seems to be a desperate task already in the linear case
- "Regularity" of F does not help – the lower bound holds already for linear functions
- "Higher moments" of ξ do not help – the lower bound holds already for Bernoulli random variables

- ... however, strong convexity of the objective helps...

Lower bound for strongly convex Lipschitz programs

We say that $f : X \to \mathbb{R}$ is strongly convex with parameter $\mu \geq 0$ with respect to the norm $\|\cdot\|_2$ if for any $x, x' \in X$ and $\alpha \in [0,1]$,

$$f(\alpha x + (1-\alpha)x') \leq \alpha f(x) + (1-\alpha)f(x') - \tfrac{1}{2}\mu\alpha(1-\alpha)\|x - x'\|_2^2.$$

Theorem 3 [17, 1] *Let $\mathcal{S}'(L,\mu)$ be a class of convex stochastic programs such that*

- *X contains an Euclidean ball or radius $r = \sqrt{\frac{\epsilon}{\mu}}$*
- *function $F(\cdot,\xi)$ is Lipschitz continuous – for some $L < \infty$ and all $\xi \in \Xi$:*

$$|F(x,\xi) - F(x',\xi)| \leq L\|x - x'\|_2$$

- *f is strongly convex on X with parameter $\mu > 0$.*

Complexity $N(\mathcal{S}')$ of the class $\mathcal{S}'(L,\mu)$ admits the bound

$$N(\mathcal{S}') \geq \kappa(\alpha)\frac{L^2}{\mu\epsilon}.$$

Upper bound by Sample Average Approximation

The classical approach to solving (S) is as follows:

- given a random sample $\xi_1, \ldots, \xi_N$, compute the Sample Average Approximation (SAA) $\widehat{f}_N$ of f

$$\widehat{f}_N(x) = \frac{1}{N} \sum_{i=1}^{N} F(x, \xi_i),$$

  approximate (S) by the problem

$$\min_{x \in X} \widehat{f}_N(x); \tag{S_N}$$

- then use a deterministic algorithm to solve (S_N) and use optimal the solution $\widehat{x}_N$ to (S_N) as an approximate solution to (S).

Standard analysis of (S_N) yields the following

Theorem 3 [after [18]] *Suppose that*

- *$D < \infty$ is the Euclidean diameter of X (that is $\|x - x'\|_2 \le D,\ \forall x, x' \in X$)*
- *$F(\cdot, \xi)$ is Lipschitz-continuous for all $\xi \in \Xi$.*

Then the optimal solution $\widehat{x}_N$ to (S_N) satisfies

$$\mathrm{Prob}\left\{ f(\widehat{x}_N) - f_* \le cLD\sqrt{\frac{n\ln(\frac{nN}{\alpha})}{N}} \right\} \ge 1 - \alpha.$$

Theorem 3 implies an upper complexity bound for the class $(\mathcal{S})$ of Lipschitz-continuus programs by SAA:

$$N(\mathrm{SAA}, \mathcal{S}) = c' \left(\frac{LD}{\epsilon}\right)^2 n \ln\left(\frac{LD}{\alpha\epsilon}\right)$$

which should be compared to the lower bound of Theorem 2:

$$N(\mathcal{S}) \ge \left(\frac{DL}{\epsilon}\right)^2 \ln\left(\frac{2}{\alpha}\right).$$

The extra factor n naturally appears in the standard complexity analysis of the SAA which relies upon the relation

$$\begin{array}{rcl} f(\widehat{x}_N) - f_* & = & [f(\widehat{x}_N) - \widehat{f}_N(\widehat{x}_N)] \\ & & +[\widehat{f}_N(\widehat{x}_N) - \widehat{f}_N(x_*)] \quad [\leq 0] \\ & & +[f(x_*) - \widehat{f}_N(x_*)] \\ & \leq & 2\sup_{x\in X} |f(x) - \widehat{f}_N(x)|. \end{array}$$

Then the uniform convergence argument (see, e.g. [16, 13]) results in the deviation bound involving the metric entropy of the ℓ_2-ball.

Note that the extra factor is not an artefact of the proof [9]:

• *one can indeed construct a family of Lipschitz-continuous functions on an ℓ_2-ball of $\mathbb{R}^n$ such that the set of empirical minimizers (optimal solutions to (S_N)) contains "bad points" $\widehat{x}_N$ – such that $f(\widehat{x}_N) - f_* \geq cL$ unless $N \leq n$.*

Recently, a much better accuracy bounds were obtained using stability argument [6]

Theorem 4 [22, 18] *Suppose that*

- *Euclidean diameter D of X is finite*
- *$F(\cdot, \xi)$ is Lipschitz-continuous and strongly convex with parameter $\mu > 0$ for all $\xi \in \Xi$*

Then the optimal solution $\widehat{x}_N$ to (S_N) satisfies

$$E[f(\widehat{x}_N)] - f_* \le c\frac{L^2}{\mu N}.$$

Furthermore, let $F(\cdot, \xi)$ be "just" convex, and let $\tilde{x}_N$ be the optimal solution to

$$\min_{x \in X} \widehat{f}_N(x) + \lambda \|x - x_0\|_2^2, \tag{S_Z}$$

with $\lambda \asymp \frac{LD}{\sqrt{N}}$ and $x_0 \in X$.
Then the optimal solution $\widetilde{x}_N$ to (S_Z) satisfies

$$E[f(\widetilde{x}_N)] - f_* \le c'\frac{L^2 D^2}{\sqrt{N}}.$$

In other words, when using penalized approximation (S_Z), $N \asymp \frac{D^2 L^2}{\epsilon^2}$ is sufficient to achieve $E[f(\widetilde{x}_N)] - f_* \le \epsilon$.

Solution by Stochastic Approximation

We assume here the stochastic black-box framework [15] of solving (S):

we consider recursive algorithms $\mathcal{M}$ which acquire by parts the information about the problem instance (S):
• at step $t = 0$ the information available to $\mathcal{M}$ is X and the class $\mathcal{S}$ of problems (e.g., stochastic programs with Lipschitz constant $\leq L$)
• at step $t = 0, 1, ...$

- given information available from steps $i = 0, ..., t-1$, $\mathcal{M}$ form a search point $x_t \in X$
- $\mathcal{M}$ requests from the stochastic oracle some local information about (S) at x_t
- $\mathcal{M}$ forms somehow an approximate solution $\bar{x}_t$ at step t.

Here we consider the case where the oracle supplies the values $z_t \in \mathbb{R}$ and $y_t \in \mathbb{R}^n$ such that

$$E_P[z_t] = f(x_t), \quad E_P[y_t] \in \partial f(x_t).$$

As far as (S) is concerned, one can assume that

$$z_t = F(x_t, \xi_t), \quad y_t = F'_x(x_t, \xi_t).$$

(Standard) Stochastic Approximation (SA) algorithm [15]

- Chose somehow $x_0 \in X$, then compute search points

$$x_t = \pi_X\left[x_{t-1} - \gamma_t y_t\right], \quad y_t = F'(x_{t-1}, \xi_t), \quad \gamma_t > 0$$

- form current approximate solution

$$\bar{x}_t = \left[\sum_{i=1}^{t} \gamma_i\right]^{-1} \sum_{i=1}^{t} \gamma_i x_{i-1}.$$

Theorem 5 *Suppose that X has a finite diameter D and $F(\cdot, \xi)$ is Lipschitz-continuous with constant L for all $\xi \in \Xi$.*
Then the SA solution $\bar{x}_N$ with constant stepsizes $\gamma_i \equiv \frac{D}{L\sqrt{N}}$ satisfies after N steps

$$\mathrm{Prob}\left\{ f(\bar{x}_N) - f_* \leq cLD\sqrt{\frac{\ln(\alpha^{-1})}{N}} \right\} \geq 1 - \alpha.$$

I.e., complexity $N(\mathrm{SA}, \mathcal{S})$ of Stochastic Approximation on the class $\mathcal{S}(D, L)$ of Lipschitz stochastic programs satisfies

$$N(\mathrm{SA}, \mathcal{S}) \leq c' \frac{L^2 D^2}{\epsilon^2} \ln(\alpha^{-1}).$$

Heuristic considerations

Suppose we are minimizing a convex $f(x): X \to \mathbb{R}$; we are given $[f(x_i), f'(x_i)]$ at search points $x_1, ..., x_{t-1}$ and we want to decide a new search point x_t.

• The available information about f amounts to the set of affine minorants ϕ_i for f on X:

$$\phi_i(x) = f(x_i) + f'(x_i)^T(x - x_i), \;\; \phi_i(x) \leq f(x), \;\; i = 0, ..., t-1.$$

Their average

$$\bar{\phi}_i(x) = \frac{1}{t}\sum_{i=0}^{t-1}[f(x_i) + f'(x_i)^T(x - x_i)]$$

is also a lower bound for f on x.

• To get a new search point one could try to minimize penalized $\bar{\phi}_i(x)$ on X:

$$x_t = \operatorname*{argmin}_{x \in X}\left\{\sum_{i=0}^{t-1} f'(x_i)^T(x - x_i) + \frac{\beta}{2}\|x - \bar{x}\|_2^2\right\}$$

where $\bar{x} \in X$ is referred to as prox-center and $V(x, \bar{x}) = \frac{\beta}{2}\|x - \bar{x}\|_2^2$ as prox-function.

Dual Stochastic Approximation (DSA) algorithm [15, 14, 10]

- Chose somehow $x_0 \in X$, then compute search points

$$x_t = \pi_X\left[x_0 - \frac{z_t}{\beta_t}\right], \text{ where } z_t = \sum_{i=1}^t y_i \left[= \sum_{i=1}^t F'(x_{i-1}, \xi_i)\right], \ \beta_t \geq \beta_{t-1}$$

- form current approximate solution

$$\bar{x}_t = \frac{1}{t}\sum_{i=0}^{t-1} x_i.$$

Theorem 6 [14] *Suppose that X has a finite diameter D and $F(\cdot,\xi)$ is Lipschitz-continuous with constant L for all $\xi \in \Xi$. Then the SA solution $\bar{x}_N$ with parameter choice*

$$\beta_0 = \frac{L}{D}, \quad \beta_t = \frac{L^2}{D^2}\sum_{i=0}^{t-1}\beta_i^{-1} \ \left[= \beta_{t-1} + \frac{L^2}{D^2}\frac{1}{\beta_{t-1}}\right]$$

satisfies after N steps

$$\mathrm{Prob}\left\{f(\bar{x}_N) - f_* \leq cLD\sqrt{\frac{\ln(\alpha^{-1})}{N}}\right\} \geq 1 - \alpha.$$

The proof is based on replacing the "Lyapunov function" $\|\cdot\|_2^2$ in the analysis of the classic SA by the "dual function"

$$W_\beta(z) = \min_{x \in X} z^T x + \frac{\beta}{2}\|x - x_0\|_2^2, \ \ x_0 \in X.$$

Note that the minimizer $x_\beta(z)$ satisfies

$$x_\beta(z) = \pi_X\left[x_0 - z/\beta\right].$$

In the simple case of $X = \{x \in \mathbb{R}^n : \ \|x\|_2 \le R\}$, and $x_0 = 0$, one has

$$x_\beta(z) = \begin{cases} -\frac{z}{\beta}, & \|z\|_2 \le \beta R, \\ -\frac{z}{\|z\|_2}R, & \|z\|_2 > \beta R; \end{cases} \qquad W_\beta(z) = \begin{cases} -\frac{z^T z}{2\beta}, & \|z\|_2 \le \beta R, \\ \frac{\beta R^2}{2} - \|z\|_2 R, & \|z\|_2 > \beta R. \end{cases}$$

Observe that

- W_β is concave smooth function on $\mathbb{R}^n$
- $W'_\beta(z) = x_\beta(z)$
- $\|W'_\beta(z) - W'_\beta(z')\|_2 \le \frac{1}{\beta}\|z - z'\|_2$.

Thus

$$\begin{aligned} W_\beta(z') \quad &\ge \quad W_\beta(z) + W'_\beta(z)^T(z' - z) - \frac{\|z' - z\|_2^2}{2\beta} \\ &= \quad W_\beta(z) + x_\beta(z)^T(z' - z) - \frac{\|z' - z\|_2^2}{2\beta} \end{aligned}$$

Now we can write for $\beta_t \geq \beta_{t-1}$,

$$\begin{aligned} W_{\beta_t}(z_t) &\geq W_{\beta_{t-1}}(z_t) \geq W_{\beta_{t-1}}(z_{t-1}) + x_{\beta_{t-1}}(z_{t-1})^T(z_t - z_{t-1}) - \frac{\|z_t - z_{t-1}\|_2^2}{2\beta_{t-1}} \\ &= W_{\beta_{t-1}}(z_{t-1}) + x_{t-1}^T y_t - \frac{\|y_t\|_2^2}{2\beta_{t-1}}, \end{aligned}$$

so that

$$y_t^T x_{t-1} \leq W_{\beta_t}(z_t) - W_{\beta_{t-1}}(z_{t-1}) + \frac{L^2}{2\beta_{t-1}}.$$

Then

$$\sum_{i=1}^t y_i^T x_{i-1} \quad \leq \quad W_{\beta_t}(z_t) - W_{\beta_0}(z_0) + \frac{L^2}{2}\sum_{i=1}^t \beta_{t-1}^{-1} = W_{\beta_t}(z_t) + \frac{L^2}{2}\sum_{i=1}^t \beta_{t-1}^{-1}.$$

Let $x_* \in X$ be a minimizer of f on X, we have

$$\begin{aligned} \sum_{i=1}^t y_i^T(x_{i-1} - x_*) &\leq W_{\beta_t}(z_t) - \left[\sum_{i=1}^t y_i\right]^T x_* + \frac{L^2}{2}\sum_{i=1}^t \beta_{t-1}^{-1} \\ &= \left[W_{\beta_t}(z_t) - z_t^T x_*\right] + \frac{L^2}{2}\sum_{i=1}^t \beta_{t-1}^{-1} \\ &\leq \frac{\beta}{2}\|x_0 - x_*\|_2^2 + \frac{L^2}{2}\sum_{i=1}^t \beta_{t-1}^{-1} \leq \frac{\beta D^2}{2} + \frac{L^2}{2}\sum_{i=1}^t \beta_{t-1}^{-1} \end{aligned}$$

Note that, by convexity of f,

$$\underbrace{f\left(\frac{1}{t}\sum_{i=0}^{t-1} x_i\right)}_{=f(x_t)} - f_* \leq \frac{1}{t}\sum_{i=0}^{t-1}[f(x_i) - f_*] \leq \frac{1}{t}\sum_{i=0}^{t-1} f'(x_i)^T(x_i - x_*).$$

We conclude that

$$\begin{aligned} f(\bar{x}_t) - f_* &\leq \frac{1}{t}\left[\sum_{i=0}^{t-1} y_{i+1}^T(x_i - x_*) - \sum_{i=0}^{t-1} \underbrace{[y_{i+1} - f'(x_i)]}_{:=\zeta_{i+1}}{}^T(x_i - x_*)\right] \\ &\leq \frac{\beta D^2}{2t} + \frac{L^2}{2t}\sum_{i=1}^{t} \beta_{t-1}^{-1} - \frac{1}{t}\sum_{i=0}^{t-1} \zeta_{i+1}^T(x_i - x_*). \end{aligned}$$

However, $\zeta_{i+1}^T(x_i - x_*)$ is a martingale-difference with $|\zeta_{i+1}^T(x_i - x_*)| \leq 2LD$. By the Hoeffding inequality,

$$\mathrm{Prob}\left\{\sum_{i=0}^{t-1} \zeta_{i+1}^T(x_i - x_*) \leq -2LD\sqrt{2t\ln\left(\alpha^{-1}\right)}\right\} \leq \alpha.$$

When choosing $\beta_t = \frac{L^2}{R^2}\sum_{i=0}^{t-1}\beta_i^{-1} \asymp \frac{L}{R}\sqrt{2t}$, we arrive at

$$\mathrm{Prob}\left\{f(\bar{x}_t) - f_* \geq cLD\sqrt{\frac{\ln\left(\alpha^{-1}\right)}{t}}\right\} \leq \alpha.$$

Stochastic approximation for strongly convex objectives

- Suppose that $f(x)$ is strongly convex with parameter $\mu > 0$. Then as soon as

$$f(x) - f_* \leq \delta,$$

one has $\|x - x_*\|_2 \leq \sqrt{\frac{2\delta}{\mu}}$, where $x_* \in X$ is the minimizer of f.[3)]

To minimize a strongly convex function one can proceed in stages: let D be the diameter of X.

- at stage i we are given an approximate solution $\bar{x}^{i-1}$ which satisfies, "with high probability"

  $$\|\bar{x}^{i-1} - x_*\|_2^2 \leq D_{i-1}^2 \leq 2^{-(i-1)} D^2.$$

  We use the SA algorithm tuned for $D = D_i$ until an approximate solution $\bar{x}^i$ satisfying

  $$\|\bar{x}^i - x_*\|_2^2 \leq D_i^2 = \frac{D_{i-1}^2}{2}$$

  is not available.

[3)] It suffices to note that for strongly convex f, $f(x) - f_* \geq \frac{\mu}{2}\|x - x_*\|_2$.

Theorem 7 [17, 11] *Suppose that* X *is a convex, closed and bounded set* $\subset \mathbb{R}^n$, $F(\cdot, \xi)$ *is Lipschitz-continuous on* X *with constant* L *for all* $\xi \in \Xi$, *and such that* f *is strongly convex with parameter* $\mu > 0$.
Then complexity $N(SA, \mathcal{S}')$ *of the stage-wise SA algorithm satisfies*

$$N(SA, \mathcal{S}') \leq c \frac{L^2 \ln(\alpha^{-1})}{\mu \epsilon}.$$

Furthermore, the approximate solution $\bar{x}_N$ *provided by the algorithm satisfies*

$$\|\bar{x}_N - x_*\|_2 \leq \sqrt{\frac{2\epsilon}{\mu}}.$$

Taking into account problem geometry [15]

One can easily see that the statement of Theorem 2 can be rewritten as follows:

Theorem 8 *Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, and let $\mathcal{S}(D, L, \|\cdot\|)$ be a class of convex stochastic programs such that*

- *$X \subset \mathbb{R}^n$ contains a ball of norm $\|\cdot\|$ of diameter $D > 0$*
- *function $F(\cdot, \xi)$ is Lipschitz-continuous:*

$$|F(x, \xi) - F(x', \xi)| \leq L\|x - x'\|, \quad \forall \xi \in \Xi, \ \forall x, x' \in X.$$

Then complexity $N(\mathcal{S})$ of the class $\mathcal{S}(D, L, \|\cdot\|)$ satisfies

$$N(\mathcal{S}) \geq \frac{D^2 L^2}{\epsilon^2} \ln\left(\frac{2}{\alpha}\right).$$

Note that under the premise of Theorem 8, the stochastic subgradient $F'_x(x,\xi)$ satisfies

$$\|F'_x(x,\xi)\|_* \leq L$$

where $\|\cdot\|_*$ is the norm conjugate to $\|\cdot\|$.

Example. Let $\|\cdot\| = \|\cdot\|_1$. Then $\|\cdot\|_* = \|\cdot\|_\infty$, and for $y \in \mathbb{R}^n$,

$$\|y\|_\infty \leq \|y\|_2 \leq \sqrt{n}\|y\|_\infty$$

(these bound is tight).

In other words, the Lipschitz constant of F with respect to $\|\cdot\|_1$ may be $\sqrt{n}$-times smaller than if it were measured using $\|\cdot\|_2$.

Note, that a "natural" choice of the norm $\|\cdot\|$ to use would be the norm $\|\cdot\|_X$ induced by the set X itself – such that the set

$$\overline{X} = \frac{1}{2}(X - X)$$

is the unit ball of $\|\cdot\|$.

There are two questions to be answered:

- is the bound of Theorem 8 tight?
- when applicable, can we efficiently implement an optimization routine which attains the lower bound of Theorem 8?

The general answer is "NO", but

- the answer is "yes" in some important situations [15], e.g., when the norm $\|\cdot\|$ is the ℓ_1-norm and the feasible set is "simple";
- recent research allowed to develop new algorithms of stochastic approximation, which attain the "corrected bounds" [20].

Таблица 3

Условия на G	Рекомендуемый метод	Условия, при которых метод рекомендуется	Оценка трудоемкости метода, $M(\nu)$	Потенциальная граница снижения трудоемкости в классе дет. методов, $\leqslant$	Потенциальная граница снижения трудоемкости в классе ранд. методов, $\leqslant$
—	МЦТ	$\nu \leqslant n^{-2}$	$3n]\ln(1/\nu)[$	λ_∞ при $\nu \leqslant n^{-2}$	$\lambda_\infty \ln M(\nu)$ при $\nu \leqslant n^{-2}$
$\alpha_{p,n}(G) \leqslant \alpha$, $1 < p < \infty$, $s = \max(2, p)$, $\bar{\alpha} = \min\{2n; \alpha n^{1/p}\}$	$\overline{ЗС}_p$	$\alpha\nu \geqslant n^{-1/s}$	$c_p \dfrac{\alpha^s}{\nu^s}$	$\lambda_p \alpha^{2s}$ при $\alpha\nu \geqslant n^{-1/s}$	$\lambda_p \alpha^{2s} \ln M(\nu)$ при $\alpha\nu \geqslant n^{-1/s}$
					$\lambda_p \alpha^{2s}$ при $n \geqslant k_p(\alpha\nu)$
	МЦТ	$\alpha\nu < n^{-1/s}$	$3n]\ln(1/\nu)[$	λ_∞ при $\bar{\alpha}\nu < 1/32$ и $\nu \leqslant \bar{\alpha}^{-2}$	$\lambda_\infty \ln M(\nu)$ при $\bar{\alpha}\nu < 1/32$ и $\nu < \bar{\alpha}^{-2}$
				$\lambda_p \ln M(\nu)$ при $\alpha\nu < n^{-1/s}$	$\lambda_p \ln^2 M(\nu)$ при $\alpha\nu < n^{-1/s}$
$\alpha_{1,n}(G) \leqslant \alpha$	$\overline{ЗС}_{1,n}$	$\alpha\nu \geqslant n^{-1/2}$	$\dfrac{c_1 \alpha^2 \ln(n+1)}{\nu^2}$	$\lambda_1 \alpha^4$ при $1/4 > \alpha\nu \geqslant n^{-1/8}$	$\lambda_1 \alpha^4 \ln M(\nu)$ при $1/32 > \alpha\nu \geqslant n^{-1/8}$
				$\lambda_1 \alpha^4 \ln(n+1)$ при $\alpha\nu \geqslant n^{-1/2}$	$\lambda_1 \alpha^4 \ln^2 M(\nu)$ при $\alpha\nu \geqslant n^{-1/2}$
					$\lambda_1 \alpha^4 \ln M(\nu)$ при $n > k_2(\alpha\nu)$
	МЦТ	$\alpha\nu < n^{-1/2}$	$3n]\ln(1/\nu)[$	λ_∞ при $\nu \leqslant n^{-2}$	$\lambda_\infty \ln M(\nu)$ при $\nu \leqslant n^{-2}$
				$\lambda_1 \ln M(\nu)$ при $\alpha\nu < n^{-1/2}$	$\lambda_1 \ln^2 M(\nu)$ при $\alpha\nu < n^{-1/2}$
$\alpha_{\infty,n}(G) \leqslant \alpha$	МЦТ		$3n]\ln(1/\nu)[$	λ_∞ при $\nu \leqslant \alpha^{-2}$ и $\alpha\nu < 1/4$	$\lambda_\infty \ln M(\nu)$ при $\nu \leqslant \alpha^{-2}$ и $\alpha\nu < 1/32$

Page 151 of [15]

Theorem 9 [15] *Let $\mathcal{S}(D, L, \|\cdot\|_p)$ be a class of convex stochastic programs such that*

- *$X \subset \mathbb{R}^n$ contains a ball of norm $\|\cdot\|_p$ of diameter $D > 0$*
- *function $F(\cdot, \xi)$ is Lipschitz-continuous:*

$$|F(x,\xi) - F(x',\xi)| \leq L\|x - x'\|_p, \quad \forall \xi \in \Xi, \ \forall x, x' \in X.$$

Then complexity $N(\mathcal{S})$ of the class $\mathcal{S}(D, L, \|\cdot\|_p)$ satisfies

$$N(\mathcal{S}) \geq c(\alpha) \left(\frac{LD}{\epsilon}\right)^{\min(2,p)} \quad \textit{for } p > 1,$$

and

$$N(\mathcal{S}) \geq c(\alpha) \left(\frac{LD}{\epsilon}\right)^{2} \ln[n] \ \ \textit{for } p = 1.$$

Corresponding upper bounds are provided by the Mirror Descent algorithm

- General Mirror Descent scheme: Nemirovski 1977 [15]
- Modern "Proximal form": Beck & Teboulle 2003 [4]
- Primal-dual versions: Nesterov 2002-2005 [14]

Mirror Descent: the setup

Let $\|\cdot\|$ be a norm on $\mathbb{R}^n$, and let $\omega : X \to \mathbb{R}$ be differentiable on X and strongly convex (with parameter 1) with respect to $\|\cdot\|$:

$$\omega(x') \geq \omega(x) + \nabla\omega(x)^T(x' - x) + \frac{1}{2}\|x' - x\|^2, \forall x, x' \in X.$$

For $x_0 = \operatorname{argmin}_{x \in X} \omega(x)$ we denote

$$V(x, x_0) = \omega(x) - \omega(x_0) - \nabla\omega(x_0)^T(x - x_0)$$

(Bregman divergence [5]).

By construction, $V(\cdot, x_0)$ is strongly convex, $V(x_0, x_0) = 0$, and $x_0 = \operatorname{argmin}_{x \in X} V(x, x_0)$. Note that

$$V(x, x_0) \geq \frac{1}{2}\|x - x_0\|^2.$$

We refer to V as prox-function.
We denote $\Omega_X = [\max_{x,x' \in X} V(x', x)]^{1/2}$ the ω-diameter of X.

Mirror Descent algorithm [14]

The "dual version" of the Mirror Descent algorithm, associated with $\omega(\cdot)$ is as follows:

- Set the prox-center $x_0 = \operatorname{argmin}_{x \in X} \omega(x)$, put $\beta_0 > 0$ and $z_0 = 0$.
- At iteration $t = 1, \ldots$, given $x_{t-1} \in X$, compute

$$y_t = F(x_{t-1}, \xi_t), \quad z_t = \sum_{i=0}^{t} y_i;$$

  and define the new search point x_t:

$$x_t = \operatorname*{argmin}_{x \in X} \left[z_t^T x + \frac{\beta_t}{2} V(x, x_0) \right]$$

  (Bregman projection or prox-transformation of z_t).
- Form the current approximate solution $\bar{x}_t$ according to

$$\bar{x}_t = \frac{1}{t} \sum_{i=0}^{t-1} x_i.$$

Theorem 9 [14] *Suppose that X has a finite ω-diameter Ω_X and $F(\cdot, \xi)$ is Lipschitz-continuous with constant L with respect to the norm $\|\cdot\|$ for all $\xi \in \Xi$. Then the MD solution $\bar{x}_N$ with the choice of parameters*

$$\beta_0 = \frac{L}{\Omega_X}, \quad \beta_t = \frac{L^2}{\Omega_X^2} \sum_{i=0}^{t-1} \beta_i^{-1}$$

satisfies after N steps

$$\mathrm{Prob}\left\{ f(\bar{x}_N) - f_* \leq cL\Omega_X \sqrt{\frac{\ln(\alpha^{-1})}{N}} \right\} \geq 1 - \alpha.$$

As a result, the complexity $N(\mathrm{MD}, \mathcal{S})$ of MD algorith on the class $\mathcal{S}$ of Lipschitz problems admits the bound

$$N(\mathrm{MD}, \mathcal{S}) \leq c' \frac{L^2 \Omega_X^2}{\epsilon^2} \ln(\alpha^{-1}).$$

Observations

[1] The complexity of the class depends on the geometry of the feasible set through its ω-diameter. When Ω_X is "moderate", MD algorithm exhibits dimension-independent convergence.

Let, for instance, $\|\cdot\|$ be the ℓ_p-norm, and let

$$X = \{x \in \mathbb{R} : \ \|x\|_p \leq R\}.$$

In this case, $\Omega_X = O(1)R$ for $1 < p \leq 2$, and $\Omega_X = O(\ln n)R$ for $p = 1$.

For these values of p the complexity bound of MD fits the lower bound of Theorem 8.

On the other hand, when $p > 2$, there is no strongly convex with respect to $\|\cdot\|_p$ function with variation on X independent of n.

[2] In order to implement the MD algorithm, one have to be able to solve efficiently the auxiliary projection problem

$$\min_{x \in X} \left[z^T(x - x_0) + \frac{\beta}{2} V(x, x_0) \right].$$

When [1] and [2] are satisfied we refer to the situation as favorable geometry.

References I

[1] Agarwal, A. et al (2009). Information-theoretic lower bounds on the oracle complexity of convex optimization. In Advances in Neural Information Processing Systems.

[2] Arrow, K. J., Harris, T., & Marschak, J. (1951). Optimal inventory policy. Econometrica: Journal of the Econometric Society, 250-272.

[3] E.M.L. Beale, On minimizing a convex function subject to linear inequalities, Journal of the Royal Statistical Society, Series B, 17 (1955), 173–184.

[4] Beck, A., & Teboulle, M. (2003). Mirror descent and nonlinear projected subgradient methods for convex optimization. Operations Research Letters, 31(3), 167-175.

[5] Bregman, L. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming, USSR Comp. Maths and Math. Phys., Vol. 7(3): 200–217, 1967.

[6] Bousquet, O., & Elisseeff, A. (2002). Stability and generalization. Journal of Machine Learning Research, 2, 499–526.

[7] G. B. Dantzig, Linear programming under uncertainty, Management Science, 1 (1955), 197–206.

[8] Dyer, M., & Stougie, L. (2006). Computational complexity of stochastic programming problems. Mathematical Programming, 106(3), 423–432.

References II

[9] Guigues, V., Juditsky, A., & Nemirovski, A. (2016). Non-asymptotic confidence bounds for the optimal value of a stochastic program. arXiv preprint arXiv:1601.07592.

[10] Juditsky, A., et al (2005). Generalization error bounds for aggregation by mirror descent with averaging. In Advances in neural information processing systems, 603–610.

[11] Juditsky, A., & Nesterov, Y. (2014). Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. Stochastic Systems, 4(1), 44–80.

[12] Kemperman, J. H. B. (1969). On the optimum rate of transmitting information, in Probability and Information Theory, Lecture Notes in Mathematics, 89, 126–169.

[13] Koltchinskii, V. I. (1997). M-estimation, convexity and quantiles. The Annals of Statistics, 435–477.

[14] Nesterov, Y. (2009). Primal-dual subgradient methods for convex problems. Mathematical programming, 120(1), 221–259.

[15] Nemirovski, A. & Yudin, D. B. (1982). Problem complexity and method efficiency in optimization.

[16] Pollard, D. (1984). Convergence of stochastic processes. Springer Ser. in Statistics.

[17] Raginsky, M. & Rakhlin, A. (2011). Information-based complexity, feedback and dynamics in convex programming. IEEE Transactions on Information Theory, 57(10), 7036–7056.

References III

[18] Shalev-Shwartz, S., et al (2009). Stochastic Convex Optimization. In COLT 2009.

[19] Shapiro, A., & Nemirovski, A. (2005). On complexity of stochastic programming problems. In Continuous optimization (pp. 111–146). Springer US.

[20] Srebro, N., Sridharan, K., & Tewari, A. (2011). On the universality of online mirror descent. In Advances in neural information processing systems. 2645–2653.

[21] Walkup, D. W., & Wets, R. J. B. (1967). Stochastic programs with recourse. SIAM Journal on Applied Mathematics, 15(5), 1299–1314.

[22] Zinkevich, M. (2003) Online convex programming and generalized infinitesimal gradient ascent. In Proceedings of the 20th IMCL.