Stochastic online gradient-free method with inexact oracle and huge-scale optimization

Gasnikov Alexander (PreMoLab MIPT)
Joint work with
Dvurechenski Pavel (Institute Information Transmission Problems RAS)
Lagunovskaya Anastasia (Institute of Applied Mathematic RAS)

Moscow-Berlin Workshop “Advance in optimization and statistics”
Institute Information Transmission Problems, 15 May 2014

Problem formulation

For simplicity, consider at first simple stochastic optimization problem (non online) on a simplex:

$$E_\eta\left[f(x;\eta)\right] \to \min_{x \in S_n(1)},\ S_n(1) = \left\{x \geq 0 :\ \sum_{i=1}^{n} x_i = 1\right\},$$

under the conditions:

1. $E_\eta\left[f(x;\eta)\right]$ – convex function (on x);

2. Subgradient $\nabla_x f(x;\eta)$ satisfied

$$E_\eta\left[\nabla_x f(x;\eta)\right] \equiv \nabla_x E_\eta\left[f(x;\eta)\right];$$

3. $\left\|\nabla_x f(x;\eta)\right\|_\infty \leq M$ – a.s. Sometimes we can substitute this condition by more general conditions:

$$\text{a) } E_\eta\left[\left\|\nabla_x f(x;\eta)\right\|_\infty^2\right] \leq M^2;\quad \text{b) } E_\eta\left[\exp\left(\frac{\left\|\nabla_x f(x;\eta)\right\|_\infty^2}{M^2}\right)\right] \leq \exp(1).$$

Basic algorithm

We use mirror descent method in Nesterov's variant (dual averaging method; Math. Prog., 2009). Let's put $x_i^1 = 1/n$, $i = 1,...,n$, $t = 1,...,N-1$.

Dual averaging method

$$x_i^{t+1} = \frac{\exp\left(-\frac{1}{\beta_{t+1}}\sum_{k=1}^{t}\frac{\partial f\left(x^k;\eta^k\right)}{\partial x_i}\right)}{\sum_{l=1}^{n}\exp\left(-\frac{1}{\beta_{t+1}}\sum_{k=1}^{t}\frac{\partial f\left(x^k;\eta^k\right)}{\partial x_l}\right)}, \quad i = 1,...,n, \quad \beta_t = \frac{M\sqrt{t}}{\sqrt{\ln n}}.$$

Where $\left\{\eta^k\right\}$ – i.i.d. (η^k has the same distribution as η).

Theorem (Nemirovski, Yudin; Nesterov; Juditsky, Lan, Shapiro). *Let the conditions 1, 2, 3.a fulfill, then*

$$E\left[f\left(\frac{1}{N}\sum_{k=1}^{N}x^k;\eta\right)\right] - \min_{x\in S_n(1)} E_\eta\left[f\left(x;\eta\right)\right] \le 2M\sqrt{\frac{\ln n}{N}}.$$

Let the conditions 1, 2, 3 fulfill, then $\Omega \ge 0$

$$P_{x^1,...,x^N}\left\{E_\eta\left[f\left(\frac{1}{N}\sum_{k=1}^{N}x^k;\eta\right)\right] - \min_{x\in S_n(1)} E_\eta\left[f\left(x;\eta\right)\right] \ge \frac{2M}{\sqrt{N}}\left(\sqrt{\ln n} + \sqrt{8\Omega}\right)\right\} \le \exp\left(-\Omega\right).$$

If we change here the condition 3 to 3.b the similar inequality will be true (but with different constants).

Inexact gradient-free oracle

For simplicity, consider at first deterministic (non online) problem (below we generalized the work [Nesterov, 2011] devoted to the gradient-free methods):

$$f(x) \to \min_{x \in S_n(1)},$$

$f(x)$ – convex function in $\mathbb{R}^n$, $\|\nabla f(x)\|_\infty \le M$ – in some large ball in 1-norme, containing $S_n(1)$.

Main drawback: we can calculate only the value of $f(x)$, moreover this value we obtained with uncontrolled (independent from everything) stochastic error $|\tilde{\delta}| \le \delta \le \varepsilon/4$ (with zero bias = mathematical expectation). Below we'll generalize this restrictive assumption about the oracle. We would like to generate (with oracle described above) such a sequence $\{x^k\}$ that with probability $\ge 1-\sigma$:

$$f\left(\frac{1}{N}\sum_{k=1}^{N} x^k\right) - \min_{x \in S_n(1)} f(x) \le \frac{1}{N}\sum_{k=1}^{N} f(x^k) - \min_{x \in S_n(1)} f(x) \le \varepsilon$$

and N is as small as it possible.

Let e $(\tilde{e})$ – random vector with uniform distribution on a unit 1-sphere (ball) in $\mathbb{R}^n$. The main (well known) idea – use dual averaging method for generating $\{x^k\}$, with $f(x)$ change to its smoothed version

$$f^\mu(x) = E_{\tilde{e}} f(x + \mu\tilde{e}), \ \mu = \varepsilon/(2M).$$

One can verify that

$$0 \le f^\mu(x) - f(x) \le M\mu = \frac{\varepsilon}{2}.$$

If we take

$$g^\mu(x;e) = \frac{n}{\mu}\left(f(x+\mu e) - f(x)\right)e.$$

and $\delta = 0$, than conditions 2, 3 is true:

$$E_e g^\mu(x;e) \equiv \nabla f^\mu(x),$$

and with probability 1

$$\left\|g^{\mu}(x;e)\right\|_{\infty} \le Mn.$$

For $\delta > 0$ we can't calculate $f(x+\mu e)$ and $f(x)$ exactly in $g^{\mu}(x;e)$. Nevertheless, we assume the condition 2 is satisfied because of zero bias condition and independence oracle inexactness and choice of random direction e. As for the condition 3 we have to recalculate constant $M := Mn \to 2Mn$:

$$\left\|g^{\mu}(x;e)\right\|_{\infty} \le \left(M + \frac{2\delta}{\mu}\right)n = 2Mn.$$

Let's note that

$$E_e g^{\mu}(x;e) = \frac{n}{\mu} E_e\left(f(x+\mu e) + \tilde{\delta}(x+\mu e)\right)e =$$

$$= \nabla_x E_{\tilde{e}}\left(f(x+\mu\tilde{e}) + \tilde{\delta}(x+\mu\tilde{e})\right) = \nabla f^{\mu}(x) + \nabla_x E_{\tilde{e}}\tilde{\delta}(x+\mu\tilde{e})$$

.

All the results mentioned below remains the same up to a multiplicative constant (in estimation the number of required iterations) if we replace independence and zero bias condition of $\tilde{\delta}$ by

$$\left\|\nabla_x E_{\tilde{e}}\tilde{\delta}(x+\mu\tilde{e})\right\|_{\infty} = \mathrm{O}(\varepsilon/R),$$

where R – the size of the set (in considered case this set is a unit-simplex). If $\tilde{\delta}(x)/\delta$ is $\mathrm{O}(1)$-Lipschitz-continuous then this condition reduces to the $\delta = \mathrm{O}(\varepsilon)$ (for simplicity we leave out all the constants, like R). So the table below became the same. But in general we have $\delta/\mu = \mathrm{O}(\varepsilon)$. From the above we can find $\mu = \mathrm{O}(\varepsilon)$, therefore $\delta = \mathrm{O}(\varepsilon^2)$. So we have to rewrite the table below according to this oracle. Changes are concern only the second column.

Let's return to the independence and zero bias condition of $\tilde{\delta}$. According to the theorem

$$N = \frac{2(2Mn)^2}{(\varepsilon/2)^2}\left(\ln n + 8\ln\left(\sigma^{-1}\right)\right) = \frac{32M^2n^2}{\varepsilon^2}\left(\ln n + 8\ln\left(\sigma^{-1}\right)\right).$$

Hence with probability $\ge 1-\sigma$:

$$\frac{1}{N}\sum_{k=1}^{N} f^{\mu}(x^k) - \min_{x\in S_n(1)} f^{\mu}(x) \le \frac{\varepsilon}{2}.$$

So with probability $\ge 1-\sigma$:

$$f\left(\frac{1}{N}\sum_{k=1}^{N}x^k\right)-\min_{x\in S_n(1)}f(x)\le\frac{1}{N}\sum_{k=1}^{N}f\left(x^k\right)-\min_{x\in S_n(1)}f(x)\le\frac{1}{N}\sum_{k=1}^{N}f^{\mu}\left(x^k\right)-\min_{x\in S_n(1)}f^{\mu}(x)+\frac{\varepsilon}{2}\le\varepsilon.$$

If we put: $x_{\varepsilon}=\frac{1}{N}\sum_{k=1}^{N}x^k$, where $x_i^1=1/n,\ i=1,...,n,\ t=1,...,N-1,$

$$\boxed{x_i^{t+1}=\frac{\exp\left(-\frac{1}{\beta_{t+1}}\sum_{k=1}^{t}g_i^{\mu}\left(x^k;e^k\right)\right)}{\sum_{l=1}^{n}\exp\left(-\frac{1}{\beta_{t+1}}\sum_{k=1}^{t}g_l^{\mu}\left(x^k;e^k\right)\right)},\quad i=1,...,n,\quad \beta_t=\frac{M\sqrt{t}}{\sqrt{\ln n}},}$$

and $g^{\mu}\left(x^k;e^k\right)$ calculate according to the red formula with inexact oracle, then with probability $\ge 1-\sigma$:

$$f(x_{\varepsilon})-\min_{x\in S_n(1)}f(x)\le\varepsilon.$$

Online stochastic inexact gradient-free oracle

1. Instead of green problem formulation we have

$$\frac{1}{N}\sum_{k=1}^{N} E_{\eta^k}\left[f_k\left(x;\eta^k\right)\right] \to \min_{x\in S_n(1)}, \; S_n(1) = \left\{ x \geq 0 : \quad \sum_{i=1}^{n} x_i = 1 \right\}.$$

Instead of condition 3 we have $\left\| \nabla f_k\left(x;\eta^k\right)\right\|_\infty \leq M$ a.s. and instead of lilac condition 2 we have: there exists such $\nabla_x f_k\left(x;\eta^k\right)$ that

$$E_{\eta^k}\left(\nabla_x f_k\left(x;\eta^k\right) - \nabla_x E_{\eta^k}\left[f_k\left(x;\eta^k\right)\right] \middle| \Xi^{k-1} \right) \equiv 0,$$

where $\left\{\eta^k\right\}$ aren't necessarily i.i.d. and Ξ^{k-1} – σ-algebra generated by $\eta^1,...,\eta^{k-1}$. It can be shown (Gasnikov, Nesterov, Spokoiny; Automatic and Remoute Control, 2014) that the main theorem is still being the truth. For example,

$$P_{x^1,...,x^N}\left\{ \frac{1}{N}\sum_{k=1}^{N} E_{\eta^k}\left[f_k\left(x^k;\eta^k\right)\right] - \min_{x\in S_n(1)} \frac{1}{N}\sum_{k=1}^{N} E_{\eta^k}\left[f_k\left(x;\eta^k\right)\right] \geq \frac{2M}{\sqrt{N}}\left(\sqrt{\ln n} + \sqrt{8\Omega}\right)\right\} \leq \exp\left(-\Omega\right).$$

2. Instead of blue problem formulation we have

$$\frac{1}{N}\sum_{k=1}^{N} E_{\xi^k}\left[f_k\left(x;\xi^k\right)\right] \to \min_{x\in S_n(1)}.$$

We have to assume that $\left\{\xi^k\right\}$, $\left\{e^k\right\}$ and oracle inexactness are independent. We put

$$g_k^\mu\left(x;\eta^k = \left(e^k,\xi^k\right)\right) = \frac{n}{\mu}\left(f_k\left(x+\mu e^k;\xi^k\right) - f_k\left(x;\xi^k\right)\right)e^k.$$

That is we have two point (nonlinear) multi armed bandit. We asked gradient-free oracle on k-th iteration the values of function (with $\delta \leq \varepsilon/4$) in two points. With probability $\geq 1-\sigma$ (on random sequence $\left\{x^k\right\}$) after N iteration (see brown formula) we have (pseudo-regret)

$$\frac{1}{N}\sum_{k=1}^{N} E_{\xi^k}\left[f_k\left(x^k;\xi^k\right)\right] - \min_{x\in S_n(1)} \frac{1}{N}\sum_{k=1}^{N} E_{\xi^k}\left[f_k\left(x;\xi^k\right)\right] \le \varepsilon .$$

3. This results can be generalized (R – the size of the set, not necessarily simplex)

Oracle	**Oracle error**	N	**Lower bound (even for deterministic or non online cases with exact oracle)**
Two-point (and more than two-point)	$\delta \le \varepsilon/4$	$N = \mathrm{O}\left(M^2R^2n^2/\varepsilon^2\right)$	=
Two-point (with Lipschitz constant of gradient L)	$\delta = \mathrm{O}\left(\sqrt{\varepsilon/L}\right)(=\mu)$	$N = \mathrm{O}\left(M^2R^2n^2/\varepsilon^2\right)$	=
Two-point (strongly convex case with constant γ)	$\delta = \mathrm{O}\left(\varepsilon\right)$	$\varepsilon = \mathrm{O}\left(M^2n^2 \ln N/(\gamma N)\right)$	= up to a logarithmic factor (we don't know if it is possible in online case to eliminate $\ln N$)
One-point (nonlinear multi armed bandit)	$\delta = \mathrm{O}\left(\varepsilon\right)$	$N = \mathrm{O}\left(M^2R^2n^2/\varepsilon^4\right)$	=
One-point, but f_k is linear on x (multi armed bandit) or f_k is not necessary linear but oracle return partial derivative (derivative on direction).	$\delta = \mathrm{O}\left(M\right)$ It seems too strange! But let's remember that error $\tilde{\delta}$ is independent of anything and has zero bias	$N = \mathrm{O}\left(M^2R^2n/\varepsilon^2\right)$	=

Notes: a) One-point special cases (strictly convexity, Lipschitz gradient) can be considered analogously (though, as far as we know it hasn't been done yet). Crucial role here is played the trade of between choosing regularization parameter $\mu(\varepsilon,\delta,n)$ and step size policy $h_k(\varepsilon,\delta,n)$ [Bubeck, 2011; theorem 4.2], [Bubeck, 2012; theorem 6.4, 6.5]. The main ingredients here are the following: in derivative on direction case: $M \to \mathrm{O}\left(M\sqrt{n}\right)$ and in one-point nonlinear case: $M \to \mathrm{O}\left(\tilde{M}n/\mu\right)$. b) In deterministic non online case if we know additional information about smoothness (strongly convexity) of the problem we can use more precise approximation of the stochastic gradient to improve estimates in the direction $n^2 \to n$ (this reduction isn't completely reachable by the approach mentioned the next sentence). But we have to use oracle more than in two points at one step. If we know in advance that δ is sufficiently small in Lipschitz gradient case we can use (in two-point case) more precise

estimate of the subgradient norm [Nesterov, 2011; formulas (20), (36)]. This allows us ([Nesterov, 2011; Theorems 8, 9]) to improve estimation of required number of iteration. c) The estimation in two point strongly convex case is reached by simple gradient projection method with step-size policy: $h_k = (\gamma k)^{-1}$.

4. In strongly convex case in large probability deviation bound we have to "multiply" estimation not just on $\ln(\sigma^{-1})$ (as it was in other cases) but on $\ln(\ln(N)/\sigma)$. Moreover if we don't have any restrictive conditions on the tails of $\|\nabla f_k(x^k;\xi^k)\|_\infty$ (except existence of second moments), than there exists such $C>0$, that with probability $\geq 1-\sigma$

$$\frac{1}{N}\sum_{k=1}^{N} E_{\xi^k}\left[f_k(x^k;\xi^k)\right] - \min_{x\in S_n(1)} \frac{1}{N}\sum_{k=1}^{N} E_{\xi^k}\left[f_k(x;\xi^k)\right] \leq C \begin{cases} \dfrac{MRn}{\sigma\sqrt{N}} \\ M^2 n^2 \dfrac{\ln N \ln(\ln N)}{\sigma\gamma N} \end{cases}.$$

5. We can play on the prox-structure of the problem to reduce the dependence $N(n)$ (we leave out in the table the constant of strictly convexity of prox-function assuming that this constant ≥ 1). It is not trivial, but with gradient-free oracle well known mirror structure (generated by KL prox-function) is only suboptimal (up to a logarithmic factor). Details can be found in [Bubeck 2011, 2012]. It should be mentioned that one-point linear case (considered in this works) can be naturally generalized for our cases. Except strongly convex case. This case considered under only the Euclidian structure. Unfortunately, this is this well known unavoidable payment (not only in online and stochastic optimization) for acceleration according to strongly convex structure. This is because of the fact that conditional number of prox-function presents in the dependence $N(n)$ in strongly convex case.

6. If we don't have any information about the constant L in non stochastic non online case we can make the same trick as Yu. Nesterov recently described in his universal method. Moreover since $f^{\mu}(x)$ has Lipschitz constant of gradient (even if $f(x)$ is non smooth) $L = \mathrm{O}(M^2/\mu)$ we can determinate M in an adaptive manner. This idea was proposed to us by Yu. Nesterov. As for unknown parameter γ in non online case we can make the following procedure. Start with $\gamma = 1$, if we don't obtain the ε-solution after well known in advance number of iteration $N(\gamma)$ we put $\gamma := \gamma/2$, etc. The number of restarts can be estimates as $\ln(\gamma^{-1})$. This additional factor in the estimation of number of iteration is a payment for adaptability in γ.

7. As we have already mentioned above all the results is true in case that uncontrolled (stochastic) oracle errors $\left\{\delta^k\right\}$ independent of our strategy. But if we consider minimax approach in which $\left\{\delta^k\right\}$ selected in a hostile manner all the estimates have to be significantly relaxed. Hypothesis here is $\delta = \mathrm{O}(\varepsilon) \to \mathrm{O}\left(\varepsilon^2\right)$ and $\to \mathrm{O}\left(\varepsilon^{3/2}\right)$ in Lipschitz gradient case. It would be rather interesting to compare such (minimax) estimates with conception of inexact oracle from the recent works of O. Devolder et al.

Applications

1. Bilevel programming

We leave out here the details. Restrict ourselves only the main example. Consider the couple of problems:

$$\psi(x,u) \to \min_{u \in Q}, \quad f(x,u(x)) \to \min_{x}.$$

Assume that we can find approximately solution of the first problem $u(x)$. But if it is not the case when we can use automatic differentiation or something like directly using Danskin's formula (or implicit function theorem) we may have a big trouble in computation of $\nabla u(x)$. So the second problem we can solve only by the gradient-free method. The problem is: How to choose the precision of the auxiliary solution $u(x)$ for the total number of arithmetic operations will be as small as it possible? We have to find a "gold middle" in trade of between the numbers of iteration of external (second) problem and the complexity of one iteration (solution of internal problem).

2. Huge-scale optimization

Restrict ourselves by consideration of only one typical example. Suppose that $y(x) \in \mathbb{R}^m$, $x \in \mathbb{R}^n$, $n \ll m$. We have to solve the problem

$$f(x, y(x)) \to \min_{x}.$$

Where the values of the functions $y(x)$ and $f(x, y(x))$ can be efficiently estimated approximately (say, for $\mathrm{O}(m)$ arithmetic calculation). If we will use gradient-free method the total complexity $\sim mn^2$, but if we will use gradient method than complexity typically $\sim m^2 n$.

It should be mentioned that in both of this examples it is worth to use recalculation. That is, if we've already calculated, say $u(x)$, then we have to use this fact in calculation of $u(x+\Delta x)$ (where Δx is small enough). For example, we can use $u(x)$ as a rather good starting point for some iteration process calculation of $u(x+\Delta x)$. As far as we know, strictly mathematical investigation this aspect is still an open problem.

Literature

1. *Dempe S.* Foundations of bilevel programming. Dordrecht: Kluwer Academic Publishers, 2002.
2. *Nesterov Y.* Primal-dual subgradient methods for convex problems // Math. Program. Ser. B. 2009. V. 120(1). P. 261–283.
3. *Juditsky A., Lan G., Nemirovski A., Shapiro A.* Stochastic approximation approach to stochastic programming // SIAM Journal on Optimization. 2009. V. 19. № 4. P. 1574–1609.
4. *Rakhlin A.* Lecture notes on online learning, 2009. http://stat.wharton.upenn.edu/~rakhlin/papers/online_learning.pdf
5. *Hazan E.* The convex optimization approach to regret minimization. In: Optimization for Machine Learning. Eds. S. Sra, S. Nowozin, S. Wright. MIT Press, 2011. P. 287–303.
6. *Bubeck S.* Introduction to online optimization. Princeton University: Lecture Notes, 2011. http://www.princeton.edu/~sbubeck/BubeckLectureNotes.pdf
7. *Nesterov Yu.* Random gradient-free minimization of convex functions. CORE Discussion Paper 2011/1. 2011. http://www.ecore.be/DPs/dp_1297333890.pdf
   Bubeck S., Cesa-Bianchi N. Regret analysis of stochastic and nonstochastic multi-armed bandit problems // Foundation and Trends in Machine Learning. 2012. V. 5. № 1. P. 1–122. http://www.princeton.edu/~sbubeck/SurveyBCB12.pdf
8. *Nesterov Y.E.* Subgradient methods for huge-scale optimization problems // CORE Discussion Paper; 2012/2, 2012. http://dial.academielouvain.be/handle/boreal:107876

9. *Nesterov Y.E.* Efficiency of coordinate descent methods on large scale optimization problem // SIAM Journal on Optimization. 2012. V. 22. № 2. P. 341–362. http://dial.academielouvain.be/handle/boreal:121612
10. *Devolder O., Glineur F., Nesterov Yu.* First order methods of smooth convex optimization with inexact oracle // Math. Progr. Ser. A. Accepted. 2013.
11. *Devolder O., Glineur F., Nesterov Yu.* First order methods with inexact oracle: the smooth strongly convex case. CORE Discussion Paper 2013/16. 2013.
12. *Nesterov Yu.* Universal gradient methods for convex optimization problems. CORE Discussion Paper 2013/63. 2013.
13. *Bubeck S.* Conditional gradient descent and structured sparsity // Princeton Course "The Complexities of Optimization", 2013. http://blogs.princeton.edu/imabandit/2013/04/09/orf523-conditional-gradient-descent-and-structured-sparsity/