First- and Zero-order Methods for Large- and Huge-scale Problems

Pavel Dvurechensky[1], Alexander Gasnikov[2]

[1] MIPT/PreMoLab, IITP RAS; [2] MIPT/PreMoLab

07.11.14

MIPT Mathematical Club
Moscow

Outline

1. First-order methods
    - Introduction to convex optimization
    - Concept of (δ, L)-oracle
    - Stochastic inexact oracle
    - Stochastic Intermediate Gradient Method
    - Discussion and directions for further research
2. Random gradient-free methods
    - Problem formulation
    - Smoothing and gradient-free oracle
    - Gradient method modification
    - Fast gradient method modification
    - Discussion and directions for further research
3. Application: web-pages ranking problem

Outline

1. First-order methods
   - Introduction to convex optimization
   - Concept of (δ, L)-oracle
   - Stochastic inexact oracle
   - Stochastic Intermediate Gradient Method
   - Discussion and directions for further research
2. Random gradient-free methods
   - Problem formulation
   - Smoothing and gradient-free oracle
   - Gradient method modification
   - Fast gradient method modification
   - Discussion and directions for further research
3. Application: web-pages ranking problem

Outline

1. First-order methods
   - Introduction to convex optimization
   - Concept of (δ, L)-oracle
   - Stochastic inexact oracle
   - Stochastic Intermediate Gradient Method
   - Discussion and directions for further research
2. Random gradient-free methods
   - Problem formulation
   - Smoothing and gradient-free oracle
   - Gradient method modification
   - Fast gradient method modification
   - Discussion and directions for further research
3. Application: web-pages ranking problem

Introduction

Optimization areas (due to Nemirovski, Yudin, Nesterov), n is the space dimension.

1. Small-size problems, n^4 operations per iteration is ok, ellipsoid methods: $N(\varepsilon) \approx O\left(n^4 \ln\left(\frac{1}{\varepsilon}\right)\right)$.
2. Medium-size problems, n^3 operations per iteration is ok, interior point methods (based on Newton method): $N(\varepsilon) \approx O\left(n^{7/2} \ln\left(\frac{n}{\varepsilon}\right)\right)$.
3. Large-scale problems, n^2 operations per iteration is ok, first order methods (gradient type):$N(\varepsilon) \approx O\left(\frac{n^2}{\varepsilon}\right)$.
4. Huge-scale problems, n or $\ln n$ operations per iteration is ok, coordinate descent schemes, sparsity, randomization. $N(\varepsilon) \approx O\left(\frac{n}{\varepsilon^2}\right)$.

Motivation

We are in the areas of large-scale and huge-scale optimization.
Application areas:

1. Machine Learning and bioinformatics.
2. Modelling of the Internet.
3. BigData.
4. Congestion traffic modelling.

Notation

1. E – finite-dimensional real vector space, E^* – its dual.
2. The value of linear function $g \in E^*$ at $x \in E$ is $\langle g, x \rangle$.
3. $\| \cdot \|$ – some norm on E, $\| \cdot \|_*$ is its dual.
4. $d(x)$ – prox-function, differentiable and strongly convex with the parameter 1 on Q with respect to $\| \cdot \|$: $d(x) \geq \frac{1}{2}\|x - x_0\|^2$, $\forall x \in Q$, $x_0 = \arg\min_{x \in Q} d(x)$.
   Examples:
   1. Euclidean distance: $Q = \mathbb{R}^n$, $\| \cdot \| = \| \cdot \|_2$, $d(x) = \frac{1}{2}\|x\|_2^2$, $x_0 = 0$.
   2. Entropy $Q = \left\{x \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^n x_i = 1\right\}$, $\| \cdot \| = \| \cdot \|_1$, $d(x) = \ln n + \sum_{i=1}^n x_i \ln x_i$, $x_0 = \left(\frac{1}{n}, \dots, \frac{1}{n}\right)^T$.
5. Bregman distance: $V(x, z) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$.
   1. Euclidean distance: $V(x, z) = \frac{1}{2}\|x - z\|_2^2$.
   2. Kullback–Leibler divergence: $V(x, z) = \sum_{i=1}^n x_i \ln \frac{x_i}{z_i}$.

Classes of convex functions: convexity

1. Convex functions:

$$f(y) \geq f(x) + \langle g(x), y - x \rangle, \quad \forall x, y \in Q, \forall g(x) \in \partial f(x).$$

2. Strongly convex functions:

$$f(y) \geq f(x) + \langle g(x), y - x \rangle + \frac{\mu}{2}\|x - y\|^2, \quad \forall x, y \in Q, \forall g(x) \in \partial f(x).$$

3. Uniformly convex functions:

$$f(y) \geq f(x) + \langle g(x), y - x \rangle + \frac{\kappa}{2}\|x - y\|^\rho, \quad \forall x, y \in Q, \forall g(x) \in \partial f(x),$$

where $\rho \geq 2$.

Classes of convex functions: smoothness

1. Bounded subgradient: $\|g(x)\|_* \le M$, $\forall x \in Q$, $\forall g(x) \in \partial f(x)$.
   Then $\|g(x) - g(y)\|_* \le 2M$ and

   $$f(y) \le f(x) + \langle g(x), y - x\rangle + 2M\|x - y\|, \quad \forall x, y \in Q, \forall g(x) \in \partial f(x).$$

2. Lipschitz continuous gradient: $\|\nabla f(x) - \nabla f(y)\|_* \le L\|x - y\|$.
   Then

   $$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|x - y\|^2, \quad \forall x, y \in Q.$$

3. Intermediate level of smoothness: for some $\nu \in [0, 1]$
   $\|g(x) - g(y)\|_* \le L_\nu \|x - y\|^\nu$, $\forall x \in Q$, $\forall g(x) \in \partial f(x)$.
   Then

   $$f(y) \le f(x) + \langle g(x), y - x\rangle + \frac{L_\nu}{1+\nu}\|x - y\|^{1+\nu}, \quad \forall x, y \in Q, \forall g(x) \in$$

Problem formulation: simple case

Consider the problem

$$\min_{x \in Q} f(x),$$

where

1. $Q \subset E$ is a closed convex set,
2. $f(x)$ is either convex or strongly convex, either with bounded subgradient or with Lipschitz continuous gradient.
3. We know all the constants M, L, μ.

The method usually constructs sequences

1. x_k – points where the (sub)gradients are calculated.
2. y_k – approximate solution.
3. $\Psi_k(x)$ – model of the function which approximates $f(x)$ in some sense.

Lower complexity bounds

We assume a black-box first order oracle $(f(x), g(x))$. $R^2 \overset{\text{def}}{\geq} 2d(x^*)$.
The best we can expect from a method in this case.

1. Nonsmooth Convex Problem: $f(y_k) - f^* \geq \Omega\left(\frac{MR}{\sqrt{k}}\right)$,
   $N(\varepsilon) \geq \Omega\left(\frac{M^2R^2}{\varepsilon^2}\right)$.
2. Nonsmooth Strongly Convex Problem: $f(y_k) - f^* \geq \Omega\left(\frac{M}{\mu k}\right)$,
   $N(\varepsilon) \geq \Omega\left(\frac{M}{\mu\varepsilon}\right)$.
3. Smooth Convex Problem: $f(y_k) - f^* \geq \Omega\left(\frac{LR^2}{k^2}\right)$, $N(\varepsilon) \geq \Omega\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$.
4. Smooth Strongly Convex Problem:
   $f(y_k) - f^* \geq \Omega\left(\mu R^2 \exp\left(-k\sqrt{\frac{\mu}{L}}\right)\right)$, $N(\varepsilon) \geq \Omega\left(\sqrt{\frac{L}{\mu}} \ln\left(\frac{\mu R^2}{\varepsilon}\right)\right)$.

Note: In stochastic optimization the best we can expect from a method in this case

1. Nonsmooth or Smooth Convex Problem: $\mathbb{E} f(y_k) - f^* \geq \Omega\left(\frac{1}{\sqrt{k}}\right)$.
2. Nonsmooth or Smooth Strongly Convex Problem:
   $\mathbb{E} f(y_k) - f^* \geq \Omega\left(\frac{1}{k}\right)$.

Below we consider mostly the smooth case.

Simple Primal Gradient Method

$f(x)$ is smooth.

Method:

1. Choose $x_0 \in Q$.
2. $x_{k+1} = \arg\min_{x \in Q}\{f(x_k) + \langle \nabla f(x_k), x - x_k \rangle + \frac{L}{2}\|x - x_k\|_2^2\} = \pi_Q\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$.

Output:

$y_k = x_k$, $y_k = \frac{\sum_{i=1}^k x_i}{k}$ (more robust), $y_k = \arg\min_{i=1,\dots,k} f(x_i) = x_k$.

Rate of convergence:

1. Convex case: $f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2k}$.
2. Strongly convex case ($y_k = x_k$): $f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2} \exp\left(-k\frac{\mu}{L}\right)$.

Estimating functions

Assume that $\mu \geq 0$ and we have sequences $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$.

- $d(x) = \frac{1}{2}\|x - x_0\|_2^2$
- $\Psi_k(x) =$
  $\beta_k d(x) + \sum_{i=0}^k \alpha_i \left[f(x_i) + \langle \nabla f(x)(x_i), x - x_i \rangle + \frac{\mu}{2}\|x - x_i\|_2^2 \right] \leq$
  $\beta_k d(x) + f(x) \sum_{i=0}^k \alpha_i$ – model of the objective function.
- $\Psi_k^* = \min_{x \in Q} \Psi_k(x)$ its minimal value. $A_k = \sum_{i=0}^k \alpha_i$.
- If we prove that for all $k \geq 0$ it holds that

$$A_k f(y_k) \leq \Psi_k^*, \quad \Psi_k(x) \leq A_k f(x) + \beta_k d(x), \quad \forall x \in Q.$$

Then

$$A_k f(y_k) \leq \Psi_k^* \leq \Psi_k(x^*) \leq A_k f^* + \beta_k d(x^*),$$

and

$$f(y_k) - f^* \leq \frac{\beta_k}{A_k} d(x^*),$$

which can give us the rate of convergence.

Dual Gradient Method (first by Yu. Nesterov)

$f(x)$ is smooth.
Choose $d(x) = \frac{1}{2}\|x - x_0\|_2^2$, $\alpha_0 = \frac{L}{L-\mu}$, $A_k = \sum_{i=0}^k \alpha_i$, $\alpha_{k+1} = \frac{A_k\mu + L}{L-\mu}$, $\beta_k = L$.

Method:

1. Choose $x_0 \in Q$.
2. $w_k = \pi_Q\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$.
3. $x_{k+1} = \arg\min_{x\in Q} \Psi_k(x) = \arg\min_{x\in Q}\{\frac{L}{2}\|x - x_0\|_2^2 + \sum_{i=0}^k \alpha_i\left[f(x_i) + \langle\nabla f(x_i), x - x_i\rangle + \frac{\mu}{2}\|x - x_i\|_2^2\right]\}$.

Output: $y_k = \frac{\sum_{i=0}^k \alpha_i w_i}{A_k}$.

Rate of convergence:

1. Convex case: $f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2(k+1)}$.
2. Strongly convex case: $f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2}\exp\left(-(k+1)\frac{\mu}{L}\right)$.

Fast Gradient Method (first by Yu. Nesterov)

$f(x)$ is smooth.

Choose $d(x) = \frac{1}{2}\|x - x_0\|_2^2$, $\alpha_0 = 1$, $A_k = \sum_{i=0}^{k} \alpha_i$, $L + \mu A_k = \frac{L\alpha_{k+1}^2}{A_{k+1}}$, $\beta_k = L$.

Method:

1. Choose $x_0 \in Q$.
2. $y_k = \pi_Q\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$.
3. $z_k = \arg\min_{x\in Q} \Psi_k(x) = \arg\min_{x\in Q}\{\frac{L}{2}\|x - x_0\|_2^2 + \sum_{i=0}^{k} \alpha_i \left[f(x_i) + \langle \nabla f(x_i), x - x_i\rangle + \frac{\mu}{2}\|x - x_i\|_2^2\right]\}$.
4. $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$, $\tau_k = \frac{\alpha_{k+1}}{A_{k+1}}$.

Rate of convergence:

1. Convex case: $f(y_k) - f^* \leq \frac{4L\|x_0 - x^*\|_2^2}{k^2}$.
2. Strongly convex case: $f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2} \exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right)$.

Generalizations

1. Non-Euclidean setup, e.g. Q – simplex, $d(x)$ – entropy.
   Auxiliary problem $\min_{x \in Q}\{\langle g, x\rangle + d(x)\}$ can be solved explicitly:
   $$\hat{x}_i = \frac{exp(-g_i)}{\sum_{i=1}^n exp(-g_i)}, \quad i = 1, ..., n.$$
2. Composite optimization: $f(x) \to \varphi(x) := f(x) + h(x)$, where $h(x)$ is a simple convex function: the problem $\min_{x \in Q}\{\langle g, x\rangle + \alpha d(x) + \beta h(x)\}$ is easy solvable.
   Example: LASSO $\|x - a\|_2^2 + \lambda\|x\|_1 \to \min$.
   Non-smooth (but strongly convex) function $\Rightarrow$ lower bound $O\left(\frac{1}{k}\right)$.
   But $\|x - a\|_2^2$ is strongly convex and smooth $\Rightarrow$ we get method with $O(\exp(-k \cdot \mathrm{const}))$.
3. Stochastic error, e.g. $\mathbb{E}_\xi f(x, \xi) \to \min_{x \in Q}$.
   On the step k we can get only
   $\nabla f(x, \xi_k): \quad \mathbb{E}_{\xi_k}\nabla f(x, \xi_k) = \nabla \mathbb{E}_{\xi_k} f(x, \xi_k)$ and
   $\mathbb{E}_{\xi_k}\|\nabla f(x, \xi_k) - \nabla \mathbb{E}_{\xi_k} f(x, \xi_k)\|_*^2 \le \sigma^2$.
4. Deterministic error (will be explained below).
5. Unknown L, μ, R .
6. Primal-dual methods.
7. Saddle-point problems and Variational inequalities.

Outline

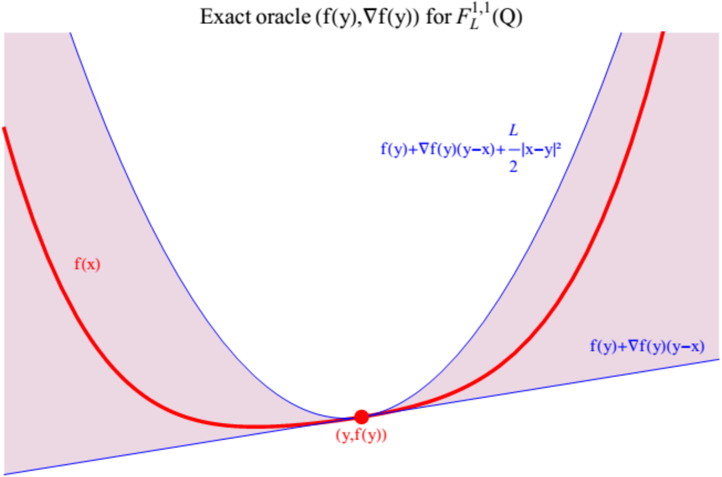
(δ, L)-oracle

[Devolder, Glineur and Nesterov, 2011-2013]
For every $x \in Q$ there are $f_{\delta,L}(x) \in \mathbb{R}$ and $g_{\delta,L}(x) \in E^*$ such that

$$0 \leq f(y) - f_{\delta,L}(x) - \langle g_{\delta,L}(x), y - x \rangle \leq \frac{L}{2}\|x - y\|^2 + \delta, \quad \forall y \in Q.$$

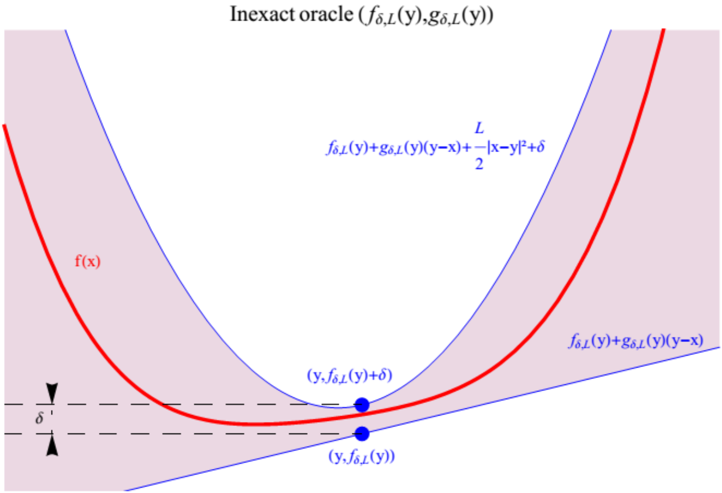
Usual oracle $(f(x), \nabla f(x))$ is replaced by (δ, L)-oracle $(f_{\delta,L}(x), g_{\delta,L}(x))$.

(δ, L)-oracle geometry



Picture from O.Devolder PhD Thesis, 2013

(δ, L)-oracle geometry



Picture from O.Devolder PhD Thesis, 2013

Convex case: PGM, DGM and FGM revisited

For $\mu = 0$ the only change we need to do in the schemes is $f(x) \to f_{\delta,L}(x)$, $\nabla f(x) \to g_{\delta,L}(x)$.

[Devolder, Glineur and Nesterov, 2011-2013]:

1. PGM, $y_k = \frac{\sum_{i=1}^{k} x_i}{k}$, $f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2k} + \delta$.
2. DGM, $y_k = \frac{\sum_{i=0}^{k} w_i}{k+1}$, $f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2(k+1)} + \delta$.
3. FGM, $f(y_k) - f^* \leq \frac{2L\|x_0 - x^*\|_2^2}{(k+1)^2} + \frac{1}{3}(k+3)\delta$.

These methods can be generalized to strongly convex case.

(δ, L, μ)-oracle

[Devolder, Glineur and Nesterov, 2011-2013]
For every $x \in Q$ there are $f_{\delta,L,\mu}(x) \in \mathbb{R}$ and $g_{\delta,L,\mu}(x) \in E^*$ such that

$$\frac{\mu}{2}\|x-y\|^2 \leq f(y) - f_{\delta,L,\mu}(x) - \langle g_{\delta,L,\mu}(x), y-x \rangle \leq \frac{L}{2}\|x-y\|^2 + \delta, \forall y \in Q.$$

Usual oracle $(f(x), \nabla f(x))$ is replaced by (δ, L, μ)-oracle $(f_{\delta,L,\mu}(x), g_{\delta,L,\mu}(x))$.

Strongly convex case: PGM, DGM and FGM revisited

The only change we need to do in the schemes is $f(x) \to f_{\delta,L,\mu}(x)$, $\nabla f(x) \to g_{\delta,L,\mu}(x)$.
[Devolder, Glineur and Nesterov, 2011-2013]:

1. PGM, $y_k = \arg\min_{i=1,\dots,k} f(x_i)$,
   $f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2} \exp\left(-k\frac{\mu}{L}\right) + \delta$.
2. DGM, $y_k = \frac{\sum_{i=0}^k \alpha_i w_i}{A_k}$, $f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2} \exp\left(-(k+1)\frac{\mu}{L}\right) + \delta$.
3. FGM, $f(y_k) - f^* \leq L\|x_0 - x^*\|_2^2 \exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right) + \left(1 + \sqrt{\frac{L}{\mu}}\right)\delta$.

(δ, L)-oracle application 1

$f(x) = f_1(x) + f_2(x)$, where $f_1(x)$ is convex, smooth with L_1 - Lipschitz continuous gradient, $f_2(x)$ is convex, non-smooth with M_2 bounded variation of subgradient.
Then $(f_1(x) + f_2(x), \nabla f_1(x) + g_2(x)), \quad g_2(x) \in \partial f_2(x)$ is a (δ, L)-oracle for $f(x)$ with $L = L_1 + \frac{M_2^2}{2\delta}$.
Fixing again number of iterations N and optimizing in δ we obtain

$$f(y_N) - f^* \leq \frac{2L_1 R^2}{(N+1)^2} + \frac{2M_2 R}{\sqrt{N+1}}.$$

Outline

Concept of stochastic inexact oracle

[Devolder, Glineur and Nesterov, 2011-2013]
The function $f(x)$ is equipped with (δ, L)-oracle. For every $x \in Q$ there are $f_{\delta,L}(x) \in \mathbb{R}$ and $g_{\delta,L}(x) \in E^*$ such that

$$0 \le f(y) - f_{\delta,L}(x) - \langle g_{\delta,L}(x), y - x \rangle \le \frac{L}{2}\|x - y\|^2 + \delta, \quad \forall y \in Q.$$

Instead of $(f_{\delta,L}(x), g_{\delta,L}(x))$ ((δ, L)-oracle) we use their stochastic approximations $(F_{\delta,L}(x,\xi), G_{\delta,L}(x,\xi))$.
We associate with x a random variable ξ whose probability distribution is supported $\Xi \subset \mathbb{R}$ and such that

$$\mathbb{E}_\xi F_{\delta,L}(x,\xi) = f_{\delta,L}(x)$$
$$\mathbb{E}_\xi G_{\delta,L}(x,\xi) = g_{\delta,L}(x)$$
$$\mathbb{E}_\xi \| G_{\delta,L}(x,\xi) - g_{\delta,L}(x)\|_*^2 \le \sigma^2.$$

Stochastic inexact oracle: examples

1. Usual stochastic optimization $\mathbb{E}_\xi f(x,\xi) \to \min_{x\in Q}$.
   On the step k we can get only
   $\nabla f(x,\xi_k): \quad \mathbb{E}_{\xi_k}\nabla f(x,\xi_k) = \nabla \mathbb{E}_{\xi_k} f(x,\xi_k)$ and
   $\mathbb{E}_{\xi_k}\|\nabla f(x,\xi_k) - \nabla \mathbb{E}_{\xi_k} f(x,\xi_k)\|_*^2 \le \sigma^2$.
   Here $\delta = 0$.
2. Randomization technique for LASSO.
   $\frac{1}{2}\|Ax-b\|_2^2 + \lambda\|x\|_1 = f(x) + h(x)$, where $A \in \mathbb{R}^{N\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^N$.
   $\nabla f(x) = A^TAx - A^Tb = \sum_{i=1}^N (x^Ta_i - b_i)a_i$ is very difficult to calculate when N is very large.
   The idea is to replace $\nabla f(x)$ by $G_{0,L}(x,\xi) = \frac{N}{M}\sum_{j=1}^M (x^Ta_{\xi_j} - b_{\xi_j})a_{\xi_j}$, where $\{\xi_1,\dots,\xi_M\}$ is a subset of rows uniformly chosen from $\{1,\dots,N\}$.
   Here $\delta = 0$.

[Devolder, Glineur and Nesterov, 2011-2013]

Choose $d(x) = \frac{1}{2}\|x - x_0\|_2^2$, $\beta_k > L$, $\gamma_k = \frac{1}{\beta_k}$.

Method:

1. Choose $x_0 \in Q$.
2. $x_{k+1} = \arg\min_{x \in Q}\{F_{\delta,L}(x_k,\xi_k) + \langle G_{\delta,L}(x_k,\xi_k), x - x_k\rangle + \frac{\beta_k}{2}\|x - x_k\|_2^2\} = \pi_Q\left(x_k - \gamma_k G_{\delta,L}(x_k,\xi_k)\right)$.

Output: $y_k = \frac{\sum_{i=0}^{k-1}\gamma_i x_{i+1}}{\sum_{i=0}^{k-1}\gamma_i}$.

Rate of convergence:

1. If N is fixed in advance: choose $\beta_i = L + \frac{\sigma}{R}\sqrt{N}$ and obtain $\mathbb{E}f(y_N) - f^* \leq \frac{LR^2}{2N} + \frac{3\sigma R}{2\sqrt{N}} + \delta$.
2. Otherwise choose $\beta_i = \frac{(L + \frac{\sigma}{R}\sqrt{i+1})^2}{L + \frac{\sigma}{2R}\sqrt{i+1}}$ and obtain $\mathbb{E}f(y_k) - f^* \leq \Theta\left(\frac{LR^2 \ln k}{k} + \frac{\sigma R \ln k}{\sqrt{k}} + \delta\right)$.

Can be generalized to non-Euclidean setup and composite optimization.

Stochastic estimating functions

- $\Psi_k(x) = \beta_k d(x) + \sum_{i=0}^{k} \alpha_i \left[F_{\delta,L}(x_i,\xi_i) + \langle G_{\delta,L}(x_i,\xi_i), x - x_i\rangle\right]$ – stochastic model of the objective function.
- $\Psi_k^* = \min_{x\in Q} \Psi_k(x)$ its minimal value.
- If we prove that for all $k \geq 0$ it holds that

$$A_k f(y_k) \leq \Psi_k^* + E_k, \quad \Psi_k(x) \leq A_k f(x) + \beta_k d(x) + \bar{E}_k(x), \quad \forall x \in Q.$$

Then

$$A_k f(y_k) \leq \Psi_k^* + E_k \leq \Psi_k(x^*) + E_k \leq A_k f^* + \beta_k d(x^*) + \bar{E}_k(x^*) + E_k,$$

and

$$f(y_k) - f^* \leq \frac{\beta_k}{A_k} d(x^*) + \frac{\bar{E}_k(x^*) + E_k}{A_k},$$

which can give us mean rate of convergence and probability of large deviations.

Stochastic Dual Gradient Method

[Devolder, Glineur and Nesterov, 2011-2013]
Choose $d(x) = \frac{1}{2}\|x - x_0\|_2^2$, $\alpha_0 \in (0,1]$, $A_k = \sum_{i=0}^k \alpha_i$, $\beta_{k+1} \geq \beta_k > L$, $\beta_k \geq \alpha_{k+1}\beta_{k+1}$.

Method:

1. Choose $x_0 \in Q$.
2. $w_k = \pi_Q\left(x_k - \frac{1}{\beta_k} G_{\delta,L}(x_k,\xi_k)\right)$.
3. $x_{k+1} = \arg\min_{x\in Q} \Psi_k(x) = \arg\min_{x\in Q}\{\frac{\beta_k}{2}\|x - x_0\|_2^2 + \sum_{i=0}^k \alpha_i [F_{\delta,L}(x_i,\xi_i) + \langle G_{\delta,L}(x_i,\xi_i), x - x_i\rangle]\}$.

Output: $y_k = \frac{\sum_{i=0}^k \alpha_i w_i}{A_k}$.

Choose $\alpha_i = \frac{1}{\sqrt{2}}$, $\beta_i = L + \frac{2^{1/4}\sigma}{R}\sqrt{i+1}$

Rate of convergence: $\mathbb{E} f(y_k) - f^* \leq \frac{LR^2}{\sqrt{2}(k+1)} + \frac{2^{3/4}\sigma R}{\sqrt{k+1}} + \delta$. Large deviations:
$$P\left\{f(y_k) - f^* > \frac{LR^2}{\sqrt{2}(k+1)} + \left(1 + \frac{\Omega}{2}\right)\frac{2^{3/4}\sigma R}{\sqrt{k+1}} + \frac{\sqrt{3\Omega}\sigma D}{\sqrt{k+1}} + \delta\right\} \leq 2\exp(-\Omega).$$
Can be generalized to non-Euclidean setup and composite optimization.

Stochastic Fast Gradient Method

[Devolder, Glineur and Nesterov, 2011-2013]
Choose $d(x) = \frac{1}{2}\|x - x_0\|_2^2$, $\alpha_0 \in (0,1]$, $A_k = \sum_{i=0}^{k} \alpha_i$, $\beta_{k+1} \geq \beta_k > L$, $\alpha_k^2 \beta_k \leq A_k \beta_{k-1}$.

Method:

1. Choose $x_0 \in Q$.
2. $y_k = \pi_Q\left(x_k - \frac{1}{\beta_k} G_{\delta,L}(x_k,\xi_k)\right)$.
3. $z_k = \arg\min_{x\in Q} \Psi_k(x) = \arg\min_{x\in Q}\{\frac{\beta_k}{2}\|x - x_0\|_2^2 + \sum_{i=0}^{k} \alpha_i \left[F_{\delta,L}(x_i,\xi_i) + \langle G_{\delta,L}(x_i,\xi_i), x - x_i\rangle\right]\}$.
4. $x_{k+1} = \tau_k z_k + (1-\tau_k) y_k$, $\tau_k = \frac{\alpha_{k+1}}{A_{k+1}}$.

Choose $\alpha_i = \frac{i+1}{2\sqrt{2}}$, $\beta_i = L + \frac{\sigma}{2^{1/4}\sqrt{3}R}(i+2)^{3/2}$

Rate of convergence:

$\mathbb{E} f(y_k) - f^* \leq \frac{2^{3/2}LR^2}{(k+1)(k+2)} + \frac{2^{9/4}(k+3)^{3/2}\sigma R}{\sqrt{3}(k+1)(k+2)} + \frac{1}{3}(k+3)\delta$. Large deviations: the same order.

Can be generalized to non-Euclidean setup and composite optimization.

Outline

Problem formulation

The main problem we are going to consider is

$$\min_{x \in Q} \{\varphi(x) := f(x) + h(x)\},$$

where

1. $Q \subset E$ is a closed convex set,
2. $h(x)$ is a simple convex function: the problem $\min_{x \in Q}\{\langle g, x\rangle + \alpha d(x) + \beta h(x)\}$ is easy solvable,
3. $f(x)$ is convex function with stochastic inexact oracle.

Existing results

From the complexity theory: the best convergence rate when $\delta = 0$ is $\mathrm{const} \cdot \frac{1}{\sqrt{k}}$. Some results by Devolder, Glineur and Nesterov, 2011-2013:

1. Stochastic Dual Gradient Method gives the mean rate (and large deviations)

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \Theta\left(\frac{LR^2}{k} + \frac{\sigma R}{\sqrt{k}} + \delta\right).$$

2. Stochastic Fast Gradient Method gives the mean rate (and large deviations)

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \Theta\left(\frac{LR^2}{k^2} + \frac{\sigma R}{\sqrt{k}} + k\delta\right).$$

3. For deterministic case Intermediate Gradient Method gives the rate

$$\varphi(y_k) - \varphi^* \leq \Theta\left(\frac{LR^2}{k^p} + k^{p-1}\delta\right),$$

where we can choose $p \in [1, 2]$.

Stochastic Intermediate Gradient Method

Our goal is the method

1. with mean rate

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \Theta\left(\frac{LR^2}{k^p} + \frac{\sigma R}{\sqrt{k}} + k^{p-1}\delta\right),$$

   where we can choose $p \in [1, 2]$
2. with bounded large deviations from this rate,
3. which is possible to use in non-Euclidean set-up (free choice of the norm),
4. applicable to composite optimization problems.

Auxiliary objects

- Simple gradient mapping

$$y = \arg\min_{x \in Q} \{\beta_k d(x) + \alpha_k \left[F_{\delta,L}(x_k, \xi_k) + \langle G_{\delta,L}(x_k, \xi_k), x - x_k\rangle\right] + h(x)\}.$$

- Minimum of smoothed model of the function

$$z = \arg\min_{x \in Q} \{\beta_k d(x) + \sum_{i=0}^{k} \alpha_i \left[F_{\delta,L}(x_i, \xi_i) + \langle G_{\delta,L}(x_i, \xi_i), x - x_i\rangle\right] + A_k h(x)\}$$

Auxiliary objects

Let $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ be three sequences of coefficients satisfying

$$\alpha_0 \in (0,1], \quad \beta_{i+1} \geq \beta_i > L, \quad \forall i \geq 0,$$
$$0 \leq \alpha_i \leq B_i, \quad \forall i \geq 0,$$
$$\alpha_k^2 \beta_k \leq B_k \beta_{k-1} \leq \left(\sum_{i=0}^{k} \alpha_i \right) \beta_{k-1}, \quad \forall k \geq 1.$$

We define also $A_k = \sum_{i=0}^{k} \alpha_i$ and $\tau_i = \frac{\alpha_{i+1}}{B_{i+1}}$.

The method

Input: The sequences $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$, functions $d(x)$, $V(x,z)$.
Output: The point y_k.

1. $x_0 = \arg\min_{x\in Q}\{d(x)\}$.
2. $$y_0 = \arg\min_{x\in Q}\{\beta_0 d(x) + \alpha_0 \langle G_{\delta,L}(x_0,\xi_0), x - x_0\rangle + h(x)\}$$
3. for $k = 0, 1, \ldots$ repeat
4. $$z_k = \arg\min_{x\in Q}\{\beta_k d(x) + \sum_{i=0}^{k} \alpha_i \langle G_{\delta,L}(x_i,\xi_i), x - x_i\rangle + A_k h(x)\}$$
5. $$x_{k+1} = \tau_k z_k + (1-\tau_k) y_k$$
6. $$\hat{x}_{k+1} = \arg\min_{x\in Q}\{\beta_k V(x,z_k) + \alpha_{k+1}\langle G_{\delta,L}(x_{k+1},\xi_{k+1}), x - z_k\rangle + \alpha_{k+1} h(x)\}.$$
7. $$w_{k+1} = \tau_k \hat{x}_{k+1} + (1-\tau_k) y_k$$
8. $$y_{k+1} = \frac{A_{k+1} - B_{k+1}}{A_{k+1}} y_k + \frac{B_{k+1}}{A_{k+1}} w_{k+1}$$

Estimating functions

- $\Psi_k(x) = \beta_k d(x) + \sum_{i=0}^{k} \alpha_i \left[F_{\delta,L}(x_i,\xi_i) + \langle G_{\delta,L}(x_i,\xi_i), x - x_i\rangle + h(x)\right]$ – model of the objective function.
- $\Psi_k^* = \min_{x\in Q} \Psi_k(x)$ its minimal value.
- If we prove that for all $k \geq 0$ it holds that

$$A_k\varphi(y_k) \leq \Psi_k^* + E_k, \quad \Psi_k(x) \leq A_k\varphi(x) + \beta_k d(x) + \bar{E}_k(x), \quad \forall x \in Q.$$

Then

$$A_k\varphi(y_k) \leq \Psi_k^* + E_k \leq \Psi_k(x^*) + E_k \leq A_k\varphi^* + \beta_k d(x^*) + \bar{E}_k(x^*) + E_k,$$

and

$$\varphi(y_k) - \varphi^* \leq \frac{\beta_k}{A_k} d(x^*) + \frac{\bar{E}_k(x^*) + E_k}{A_k},$$

which can give us mean rate of convergence and probability of large deviations.

General rate of convergence

Theorem 1

Assume that the function f is endowed with stochastic inexact oracle with parameters δ, L, σ. Then the sequence y_k generated by the Stochastic Intermediate Gradient Method, when applied to the composite function φ, satisfies

$$\varphi(y_k) - \varphi^* \leq \frac{1}{A_k}\Bigg(\beta_k d(x^*) + \sum_{i=0}^{k} B_i\delta +$$

$$+ \sum_{i=0}^{k} \frac{B_i}{\beta_i - L}\|G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i)\|_*^2 +$$

$$+ \sum_{i=0}^{k} \alpha_i \langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), x^* - x_i\rangle +$$

$$+ \sum_{i=1}^{k} (B_i - \alpha_i)\frac{\alpha_i}{B_i}\langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), y_{i-1} - z_{i-1}\rangle\Bigg).$$

General mean rate of convergence

Taking expectation we get

Theorem 2

Assume that the function f is endowed with stochastic inexact oracle with parameters δ, L, σ. Then the sequence y_k generated by the Stochastic Intermediate Gradient Method, when applied to the composite function φ, satisfies

$$\mathbb{E}_{\xi_0,\dots,\xi_k}\varphi(y_k) - \varphi^* \leq \frac{\beta_k d(x^*)}{A_k} + \frac{\sum_{i=0}^{k} B_i \delta}{A_k} +$$
$$+ \frac{1}{A_k}\sum_{i=0}^{k} \frac{B_i}{\beta_i - L}\sigma^2.$$

Additional assumptions

1. $\xi_0, \dots, \xi_k$ are i.i.d random variables.
2. $G_{\delta,L}(x,\xi)$ satisfies the condition
$\mathbb{E}_\xi \left[\exp\left(\frac{\|G_{\delta,L}(x,\xi)-g_{\delta,L}(x)\|_*^2}{\sigma^2}\right)\right] \leq \exp(1)$.
3. Set Q is bounded with diameter $D = \max_{x,y\in Q} \|x - y\|$.

$\xi_{[k]} = (\xi_0, \dots, \xi_k)$ – history of the random process after k iterations.

General result for large deviations

Theorem 3

If the assumptions 1, 2, 3 are satisfied, then for all $k \geq 0$ and all $\Omega \geq 0$, the sequence generated by the SIGM satisfies:

$$\mathrm{P}\left(\varphi(y_k) - \varphi^* \geq \frac{\beta_k d(x^*)}{A_k} + \frac{\sum_{i=0}^{k} B_i \delta}{A_k} + \right.$$

$$\left. + \frac{1+\Omega}{A_k} \sum_{i=0}^{k} \frac{B_i}{\beta_i - L} \sigma^2 + \frac{2D\sigma\sqrt{3\Omega}}{A_k} \sqrt{\sum_{i=0}^{k} \alpha_i^2} \right) \leq 3\exp(-\Omega).$$

Choice of the coefficients

Our goal is the rate of $\Theta\left(\frac{LR^2}{k^p} + \frac{\sigma R}{\sqrt{k}} + k^{p-1}\delta\right)$, $p \in [1,2]$.
Let $a = 2^{\frac{2p-1}{2}}$ and $b = 2^{\frac{5-2p}{4}} p^{\frac{1-2p}{2}}$, $R \geq \sqrt{2d(x^*)}$.
Then the sequences

$$\alpha_i = \frac{1}{a}\left(\frac{i+p}{p}\right)^{p-1}, \quad \forall i \geq 0,$$

$$\beta_i = L + \frac{b\sigma}{R}(i+p+1)^{\frac{2p-1}{2}}, \quad \forall i \geq 0,$$

$$B_i = a\alpha_i^2 = \frac{1}{a}\left(\frac{i+p}{p}\right)^{2p-2}, \quad \forall i \geq 0.$$

satisfy all the requirements.

Mean rate of convergence

Theorem 4

If the sequences $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ are chosen from relations above and $p \in [1,2]$ then the sequence generated by the SIGM satisfies:

$$\mathbb{E}_{\xi_0,\ldots,\xi_k}\varphi(y_k) - \varphi^* \leq$$

$$\leq \frac{LR^2 p^p 2^{\frac{2p-3}{2}}}{(k+p)^p} + \frac{\sigma R 2^{\frac{3+2p}{4}}\sqrt{p}(k+p+2)^{p-\frac{1}{2}}}{(k+p)^p} +$$

$$+ 2^{2p-1}\left(\left(\frac{k+p}{p}\right)^{p-1} + 1\right)\delta =$$

$$= \Theta\left(\frac{LR^2}{k^p} + \frac{\sigma R}{\sqrt{k}} + k^{p-1}\delta\right).$$

Large deviations bound

Theorem 5

If the sequences $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ are chosen from relations above and $p \in [1,2]$ then the sequence generated by the SIGM satisfies:

$$\mathrm{P}\Bigg(\varphi(y_k) - \varphi^* > \\ > \frac{LR^2 p^p 2^{\frac{2p-3}{2}}}{(k+p)^p} + \frac{(1+\Omega)\sigma R 2^{\frac{3+2p}{4}}\sqrt{p}(k+p+2)^{p-\frac{1}{2}}}{(k+p)^p} + \\ + 2^{2p-1}\left(\left(\frac{k+p}{p}\right)^{p-1}+1\right)\delta + \frac{2D\sigma\sqrt{6p\Omega}}{\sqrt{k+p}}\Bigg) \leq \\ \leq 3\exp(-\Omega).$$

SIGM: strongly convex case

Let E be Euclidean space and $\|x\|^2 = \langle x, Hx \rangle$ for some $H > 0$.
Assume that $\varphi(x)$ is strongly convex. Then

$$\varphi(x) - \varphi(x^*) \geq \frac{\mu}{2}\|x - x^*\|^2, \quad \forall x \in Q.$$

We assume that prox-function $d(x)$ satisfies $0 = \arg\min_{x \in Q} d(x)$ and $d(0) = 0$ and has quadratic growth with constant V^2: $d(x) \leq \frac{V^2}{2}\|x\|^2$ for all $x \in E$.
Let us change $G_{\delta,L}(x, \xi_j) \to \tilde{G}_{\delta,L}(x, \Xi) = \frac{1}{m}\sum_{j=1}^{m} G_{\delta,L}(x, \xi_j)$.
Then $\sigma^2 \to \frac{\sigma^2}{m}$ and

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \frac{C_1 L d(x^*)}{k^p} + \frac{C_2 \sigma R}{\sqrt{mk}} + C_3 k^{p-1}\delta.$$

Let us use restart technique.

SIGMA

Input: The function $d(x)$, point u_0, number R_0 such that $\|u_0 - x^*\| \leq R_0$, number $p \in [1,2]$. Output: The point u_{k+1}.

1. Set $k = 0$.
2. Define $N_k = \left\lceil \left(\frac{4eC_1LV^2}{\mu}\right)^{\frac{1}{p}} \right\rceil$.
3. for $k = 0, 1, \ldots$ repeat
4. Define
$$m_k = \max\left\{1, \left\lceil \frac{16e^{k+2}C_2^2\sigma^2V^2}{\mu^2R_0^2N_k} \right\rceil \right\},$$
$$R_k^2 = R_0^2e^{-k} + \frac{2^peC_3\delta}{\mu(e-1)}\left(\frac{4eC_1LV^2}{\mu}\right)^{\frac{p-1}{p}}\left(1 - e^{-k}\right).$$
5. Run SIGM with $x_0 = u_k$, prox-function $d\left(\frac{x-u_k}{R_k}\right)$ for N_k steps using oracle $\tilde{G}^k_{\delta,L}(x,\Xi) = \frac{1}{m_k}\sum_{j=1}^{m_k} G_{\delta,L}(x,\xi_j)$ on each step and sequences $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ defined above.
6. Set $u_{k+1} = y_{N_k}$, $k = k+1$.

SIGMA: rate of convergence

Theorem 6

After $k \geq 1$ outer iterations of the SIGMA we have

$$\mathbb{E}\varphi(u_k) - \varphi^* \leq \frac{\mu R_0^2}{2} e^{-k} + \frac{C_3 e 2^{p-1}}{e-1} \left(\frac{4eC_1 LV^2}{\mu} \right)^{\frac{p-1}{p}} \delta,$$

$$\mathbb{E}\|u_k - x^*\|^2 \leq R_0^2 e^{-k} + \frac{C_3 e 2^p}{\mu(e-1)} \left(\frac{4eC_1 LV^2}{\mu} \right)^{\frac{p-1}{p}} \delta.$$

As a consequence if we choose error of the oracle δ satisfying

$$\delta \leq \frac{\varepsilon(e-1)}{2^p C_3 e} \left(\frac{4eC_1 LV^2}{\mu} \right)^{\frac{1-p}{p}}$$

then we need $N = \left\lceil \ln \left(\frac{\mu R_0^2}{\varepsilon} \right) \right\rceil$ outer iterations and no more than

$$\left(1 + \left(\frac{4eC_1 LV^2}{\mu} \right)^{\frac{1}{p}} \right) \left(1 + \ln \left(\frac{\mu R_0^2}{\varepsilon} \right) \right) + \frac{16 e^3 C_2^2 \sigma^2 V^2}{\mu \varepsilon (e-1)}$$

oracle calls to provide $\mathbb{E}\varphi(u_N) - \varphi^* \leq \varepsilon$.

Slightly changing the method we can obtain Large deviations bound.

Outline

1. First-order methods
   - Introduction to convex optimization
   - Concept of (δ, L)-oracle
   - Stochastic inexact oracle
   - Stochastic Intermediate Gradient Method
   - Discussion and directions for further research
2. Random gradient-free methods
   - Problem formulation
   - Smoothing and gradient-free oracle
   - Gradient method modification
   - Fast gradient method modification
   - Discussion and directions for further research
3. Application: web-pages ranking problem

Discussion

We have obtained the method with mean rate of convergence of $\Theta\left(\frac{LR^2}{k^p} + \frac{\sigma R}{\sqrt{k}} + k^{p-1}\delta\right)$, where we can chose $p \in [1,2]$ in advance. It has the following advantages.

1. Large deviation bounds with same asymptotic dependence on k.
2. It can be used for problems from rather general class of problems with stochastic inexact oracle.
   1. Nonsmooth problems.
   2. Auxiliary randomization in initially deterministic problem.
3. Choose $p \in [1,2]$ for optimal trade-off between error accumulation and rate of convergence.
4. Flexibility of optimal choice of the norm and prox-function.
5. Allows to solve composite optimization problems.
6. Can be accelerated in the strongly convex case to have rate $O\left(\mu R_0^2 \exp\left(-\left(\frac{\mu}{L}\right)^{\frac{1}{p}} k\right) + \frac{\sigma^2}{\mu k} + \left(\frac{L}{\mu}\right)^{\frac{p-1}{p}} \delta\right)$.

Directions for further research

1. Numerical experiments.
2. Making these algorithms primal-dual.
3. Adaptive choice of unknown p, L, R, μ, D .
4. Large deviations for heavy tails distributions and large deviations for unbounded sets.
5. Extension to saddle-point problems and variational inequalities: one method working on lower bounds, prox-structure, oracle errors (stochastic and deterministic), composite structure, adaptivity in unknown parameters.
6. Additional linear inequalities which are complex to project on.

Outline

Outline

Notation

1. E – finite-dimensional real vector space,
2. $\|\cdot\|$ – Euclidean norm on E, $\|\cdot\|_*$ is its dual:

$$\|x\| = \sqrt{\langle x, x \rangle}, \quad x \in E, \quad \|g\|_* = \sqrt{\langle g, g \rangle}, \quad g \in E^*.$$

3. $f \in C_L^{1,1}$ if $\|\nabla f(x) - \nabla f(y)\| \le L\|x - y\|$, $x \in E$. This is equivalent to

$$|f(x) - f(y) - \langle \nabla f(y), x - y \rangle| \le \frac{L}{2}\|x - y\|^2, \quad x, y \in E$$

4. $f(x)$ is smooth strongly convex function if for any $x, y \in E$

$$f(x) \ge f(y) + \langle \nabla f(y), x - y \rangle + \frac{\tau}{2}\|x - y\|^2,$$

Problem formulation

The main problem we are going to consider is

$$\min_{x \in E} f(x),$$

where

1. $f(x) \in C_L^{1,1}$ and either,
   1. convex
   2. strongly convex
2. we use only function values measured with error

$$f_\delta(x) = f(x) + \tilde{\delta}(x),$$

   $\tilde{\delta}(x)$ – oracle error satisfying $|\tilde{\delta}(x)| \leq \delta \; \forall x \in E$.
3. Sometimes we additionally assume that $\tilde{\delta}(x) \equiv \tilde{\delta}$ and is a random variable which is independent on everything.
   Our work based on the article by Yu. Nesterov (2011).

Outline

Smoothing the function

Consider smoothing:

$$f_\mu(x) = \mathbb{E}_b f(x+\mu b) = \frac{1}{V_B}\int_{\mathcal{B}} f(x+\mu b)db,$$

where

1. b is a uniformly distributed over unit ball $\mathcal{B} = \{x \in E : \|x\| \leq 1\}$ random vector,
2. V_B is the volume of the unit ball $\mathcal{B}$,
3. $\mu \geq 0$ is the smoothing parameter.

It turns out that

$$\nabla f_\mu(x) = \frac{n}{\mu}\mathbb{E}_s(f(x+\mu s) - f(x))s = \frac{n}{\mu V_S}\int_{\mathcal{S}}(f(x+\mu s) - f(x))s d\sigma(s),$$

where

1. s is a uniformly distributed over unit sphere $\mathcal{S} = \{x \in E : \|x\| = 1\}$ random vector,
2. V_S is the volume of the unit sphere $\mathcal{S}$,
3. $d\sigma(s)$ is unnormalized spherical measure.

Some properties

1. $f_\mu(x) \geq f(x), \quad \forall x \in E.$
2. If $f(x)$ is convex, then $f_\mu(x)$ is also convex.
3. If $f \in C_L^{1,1}$ then $f_\mu \in C_L^{1,1}$.
4. If $f \in C_L^{1,1}$ then $|f_\mu(x) - f(x)| \leq \frac{L\mu^2}{2}, \quad \forall x \in E.$

Random gradient-free oracle

Define random gradient-free oracle

$$g_\mu(x) = \frac{n}{\mu}(f(x + \mu s) - f(x))s,$$

where s is uniformly distributed vector over the unit sphere $\mathcal{S}$.
One can show that

$$\mathbb{E}_s g_\mu(x) = \nabla f_\mu(x).$$

Due to error we can calculate only

$$g_{\mu,\delta}(x) = \frac{n}{\mu}(f_\delta(x + \mu s) - f_\delta(x))s.$$

Some properties

Let $f \in C_L^{1,1}$. Then

1. $\|g_{\mu,\delta}(x)\|_*^2 \leq n^2\mu^2L^2 + 4n^2(\langle \nabla f(x), s\rangle)^2 + \frac{8\delta^2 n^2}{\mu^2} \leq$ $n^2\mu^2L^2 + 4n^2\|\nabla f(x)\|_*^2 + \frac{8\delta^2 n^2}{\mu^2}$.
2. $\mathbb{E}_s\|g_{\mu,\delta}(x)\|_*^2 \leq n^2\mu^2L^2 + 4n\|\nabla f(x)\|_*^2 + \frac{8\delta^2 n^2}{\mu^2}$.

If additionally we assume that $\tilde{\delta}(x) \equiv \tilde{\delta}$ and is a random variable which is independent on s

1. $\mathbb{E}_{s,\tilde{\delta}}\|g_{\mu,\delta}(x)\|_*^2 \leq n^2\mu^2L^2 + 4n\|\nabla f(x)\|_*^2 + \frac{8\delta^2 n^2}{\mu^2}$.

Outline

Problem reformulation

We consider the problem

$$\min_{x \in E} f(x).$$

Assume that we know point x_0 and number R such that $\|x_0 - x^*\| \leq R$, where x^* is the solution of the problem.
Denote $Q = \{x \in E : \|x - x_0\| \leq 2R\}$.
Then we can solve the problem

$$\min_{x \in Q} f(x).$$

The method

Input: The point x_0, number R such that $\|x_0 - x^*\| \leq R$, stepsize $h > 0$.
Output: The point x_k.
Define $Q = \{x \in E : \|x - x_0\| \leq 2R\}$.

1. Generate s_k and corresponding $g_{\mu,\delta}(x_k)$.
2. Calculate $x_{k+1} = \pi_Q(x_k - h g_{\mu,\delta}(x_k))$.

Convergence rate

Denote $\mathcal{U}_k = (s_0, \dots, s_k)$ the history of realizations of the vectors s_k, generated on each iteration of the method, $\phi_0 = f(x_0)$, and $\phi_k = \mathbb{E}_{\mathcal{U}_{k-1}}(f(x_{k-1}))$, $k \geq 1$.

Let $f \in C_L^{1,1}$ and the sequence x_k be generated by the Algorithm above with $h = \frac{1}{8nL}$. Then for any $N \geq 0$, we have

$$\frac{1}{N+1} \sum_{i=0}^{N} (\phi_i - f^*) \leq \frac{8nLR^2}{N+1} + \frac{\mu^2 L(n+8)}{8} + \frac{8\delta nR}{\mu} + \frac{\delta^2 n}{L\mu^2}.$$

If additionally f is strongly convex, then

$$\phi_N - f^* \leq \frac{1}{2} L \left(\delta_\mu + \left(1 - \frac{\tau}{16nL}\right)^N (R^2 - \delta_\mu) \right),$$

where $\delta_\mu = \frac{\mu^2 L(n+8)}{4\tau} + \frac{16n\delta R}{\tau\mu} + \frac{2n\delta^2}{\tau\mu^2 L}$.

Discussion

To achieve desired accuracy ε we need to choose.
In convex case with $|\tilde{\delta}(x)| \leq \delta$

$$N = O\left(\frac{nLR^2}{\varepsilon}\right), \quad \mu = O\left(\sqrt{\frac{\varepsilon}{Ln}}\right), \quad \delta = O\left(\min\left\{\frac{\varepsilon^{\frac{3}{2}}}{L^{\frac{1}{2}} n^{\frac{3}{2}} R}, \frac{\varepsilon}{n}\right\}\right).$$

In convex case with $\tilde{\delta}(x)$ random and independent

$$N = O\left(\frac{nLR^2}{\varepsilon}\right), \quad \mu = O\left(\sqrt{\frac{\varepsilon}{Ln}}\right), \quad \delta = O\left(\frac{\varepsilon}{n}\right).$$

In strongly convex case with $|\tilde{\delta}(x)| \leq \delta$

$$N = O\left(\frac{nL}{\tau} \ln \frac{LR^2}{\varepsilon}\right), \quad \mu = O\left(\sqrt{\frac{\tau\varepsilon}{L^2 n}}\right), \quad \delta = O\left(\min\left\{\left(\frac{\tau\varepsilon}{n}\right)^{\frac{3}{2}} \frac{1}{L^2 R}, \frac{\varepsilon\tau}{nL}\right\}\right).$$

In strongly convex case with $\tilde{\delta}(x)$ random and independent

$$N = O\left(\frac{nL}{\tau} \ln \frac{LR^2}{\varepsilon}\right), \quad \mu = O\left(\sqrt{\frac{\tau\varepsilon}{L^2 n}}\right), \quad \delta = O\left(\frac{\varepsilon\tau}{nL}\right).$$

Outline

Problem formulation and method

We consider the problem

$$\min_{x \in E} f(x),$$

where $f \in C_L^{1,1}$ and is a strongly convex function with parameter $\tau \geq 0$.
We difine $\theta = \frac{1}{64n^2L}$ and $h = \frac{1}{8nL}$ and consider the following method.

Fast Gradient Method Modified

Input: The point x_0, number $\gamma_0 \geq \tau$.
Output: The point x_k.
Set $v_0 = x_0$.

1. Compute $\alpha_k > 0$ satisfying $\frac{\alpha_k^2}{\theta} = (1 - \alpha_k)\gamma_k + \alpha_k \tau \equiv \gamma_{k+1}$.
2. Set $\lambda_k = \frac{\alpha_k}{\gamma_{k+1}}\tau$, $\beta_k = \frac{\alpha_k \gamma_k}{\gamma_k + \alpha_k \tau}$, and $y_k = (1 - \beta_k)x_k + \beta_k v_k$.
3. Generate s_k and corresponding $g_{\mu,\delta}(y_k)$.
4. Calculate $x_{k+1} = y_k - h g_{\mu,\delta}(y_k)$,
   $v_{k+1} = (1 - \lambda_k)v_k + \lambda_k y_k - \frac{\theta}{\alpha_k} g_{\mu,\delta}(y_k)$.

Define $\kappa = \frac{\tau}{L}$. In the case when $\tilde{\delta}(x)$ is random and independent we have for all $k \geq 0$

$$\mathbb{E}_{\mathcal{U}_{k-1}} f(x_k) - f^* \leq \psi_k \left(f(x_0) - f^* + \frac{\gamma_0}{2}\|x_0 - x^*\|^2 \right) +$$
$$+ C_k \left(\frac{5\mu^2 L}{64} + \frac{\delta^2}{4\mu^2 L} \right) + \mu^2 L,$$

where $\psi_k \leq \min\left\{ \left(1 - \frac{\sqrt{\kappa}}{8n}\right)^k, \left(1 + \frac{k}{16n}\sqrt{\frac{\gamma_0}{L}}\right)^{-2} \right\}$, $C_k \leq \min\left\{k, \frac{8n}{\sqrt{\kappa}}\right\}$.

Then for $\tau = 0$ to obtain the accuracy ε we need to choose

$$N = O\left(n\sqrt{\frac{LR^2}{\varepsilon}}\right), \quad \mu = O\left(\sqrt{\frac{\varepsilon}{nL}\sqrt{\frac{\varepsilon}{LR^2}}}\right), \quad \delta = O\left(\frac{\varepsilon}{n}\sqrt{\frac{\varepsilon}{LR^2}}\right)$$

For $\tau > 0$ to obtain the accuracy ε we need to choose

$$N = O\left(n\sqrt{\frac{L}{\tau}}\ln\left(\frac{\tau R^2}{\varepsilon}\right)\right), \quad \mu = O\left(\sqrt{\frac{\varepsilon}{nL}\sqrt{\frac{\tau}{L}}}\right), \quad \delta = O\left(\frac{\varepsilon}{n}\sqrt{\frac{\tau}{L}}\right)$$

Outline

Discussion

1. We have considered two random gradient-free methods with error in the oracle value: gradient-type scheme and fast-gradient-type scheme.
2. We have obtained their mean rate of convergence and bounds on the oracle error ($\tau = 0$):

$$\mathrm{PGM}: \quad N = O\left(\frac{nLR^2}{\varepsilon}\right), \quad \delta = O\left(\frac{\varepsilon}{n}\right).$$

$$\mathrm{FGM}: \quad N = O\left(n\sqrt{\frac{LR^2}{\varepsilon}}\right), \quad \delta = O\left(\frac{\varepsilon}{n}\sqrt{\frac{\varepsilon}{LR^2}}\right).$$

Directions for further research

1. Numerical experiments.
2. Making these algorithms primal-dual
3. Adaptive choice of unknown L, R, τ, μ .
4. Extension to one intermediate method, constrained optimization, prox-structure, other oracle errors (stochastic and deterministic), composite structure, adaptivity in unknown parameters.
5. Extension for other oracles: case $\mu = 0$, $f(x + \mu e_i) - f(x)$, random coordinate descent.
6. Extension to saddle-point problems and variational inequalities.

Outline

Model formulation: random walks

Attention to the whiteboard

Markov chain

$\varphi = (\varphi_1, \varphi_2)^T \in \mathbb{R}^{m_1+m_2}$ - unknown vector of parameters which help to characterize web-sites.
Probability for choosing query i:

$$[\pi_q^0]_i = \frac{f_q(\varphi_1, i)}{\sum_{\tilde{i}\in V_q^1} f_q(\varphi_1, \tilde{i})}$$

Probability of transition from one web-site to another:

$$\frac{g_q(\varphi_2, \tilde{i} \to i)}{\sum_{j:\tilde{i}\to j} g_q(\varphi_2, \tilde{i} \to j)}$$

Finally, probability of moving to i from $\tilde{i}$ equals

$$\alpha \frac{f_q(\varphi_1, i)}{\sum_{\tilde{i}\in V_q^1} f_q(\varphi_1, \tilde{i})} + (1-\alpha)\frac{g_q(\varphi_2, \tilde{i} \to i)}{\sum_{j:\tilde{i}\to j} g_q(\varphi_2, \tilde{i} \to j)}$$

Stationary distribution of Markov chain defines the i-th web-page rank: $[\pi_q]_p$.

$$\pi_q = \alpha\pi_q^0(\varphi) + (1-\alpha)P_q^T(\varphi)\pi_q,$$

Learning problem

We have some pool of experts who rank web-pages for Q queries.
For every query q we have sets of pages $P_q^1, P_q^2, ..., P_q^k$ which are ordered from the most relevant to irrelevant pages.
We choose loss function $h(i,j,x) = \max\{x + b_{ij}, 0\}^2$, where $1 \le i < j \le k$.

To find φ we minimize

$$f(\varphi) = \frac{1}{Q} \sum_q \sum_{1 \le i < j \le k} \sum_{p_1 \in P_q^i, p_2 \in P_q^j} h(i, j, [\pi_q]_{p_2} - [\pi_q]_{p_1})$$

Problem reformulation

$$f(\varphi) = \frac{1}{Q}\sum_q \|(A_q \pi_q^*(\varphi) + b_q)_+\|_2^2 \to \min$$

$$\pi_q^*(\varphi) = \alpha\left[I - (1-\alpha)P_q^T(\varphi)\right]^{-1}\pi_q^0(\varphi).$$

Nemirovski, Nesterov (2012): $\|\tilde{\pi}_q^N(\varphi) - \pi_q^*(\varphi)\|_1 \leq 2(1-\alpha)^{N+1}$ holds for

$$\tilde{\pi}_q^N(\varphi) = \frac{\alpha}{1-(1-\alpha)^{N+1}}\sum_{i=0}^{N}(1-\alpha)^i\left[P_q^T(\varphi)\right]^i\pi_q^0(\varphi)$$

To obtain vector $\tilde{\pi}_q^N(\varphi)$ s.t. $\|\tilde{\pi}_q^N(\varphi) - \pi_q^*(\varphi)\|_1 \leq \Delta$ we need $\frac{s_q(p_q+n_q)}{\alpha}\ln\frac{2}{\Delta}$ a.o.

$$f_\delta(\varphi) = \frac{1}{Q}\sum_q \|(A_q \tilde{\pi}_q^N(\varphi) + b_q)_+\|_2^2$$

satisfies $|f_\delta(\varphi) - f(\varphi)| \leq \Delta\sqrt{2r}(2\sqrt{2r}+2b)$, where $r = \max_q r_q$, $b = \max_q \|b_q\|_2$.

The method

Input: The point φ_0, L – Lipschitz constant for the function $f(\varphi)$, number R such that $\|\varphi_0 - \varphi^*\|_2 \leq R$, accuracy $\varepsilon > 0$, numbers r, b defined above.
Output: The point $\hat{\varphi}_N = \arg\min_{\varphi}\{f(\varphi) : \varphi \in \{\varphi_0, \dots, \varphi_N\}\}$.

1. Define $G = \{\varphi \in \mathbb{R}^m : \|\varphi - \varphi_0\|_2 \leq 2R\}$, $N = 32m\frac{LR^2}{\varepsilon}$, $\delta = \frac{\varepsilon^{\frac{3}{2}}\sqrt{2}}{32mR\sqrt{L(m+8)}}$, $\mu = \sqrt{\frac{2\varepsilon}{L(m+8)}}$;
2. Set $k = 0$;
3. for $k = 0, \dots, N$.
4. Generate random vector s_k uniformly distributed over a unit Euclidean sphere $\mathcal{S}$ in R^m;
5. Set $\hat{N} = \frac{1}{\alpha}\ln\frac{2\sqrt{2r}(2\sqrt{2r}+2b)}{\delta}$;
6. For every q calculate $\tilde{\pi}_q^{\hat{N}}(\varphi_k)$, $\tilde{\pi}_q^{\hat{N}}(\varphi_k + \mu s_k)$ defined in above;
7. Calculate $g_{\mu,\delta}(x_k) = \frac{m}{\mu}(f_\delta(\varphi_k + \mu s_k) - f_\delta(\varphi_k))s_k$;
8. Calculate $\varphi_{k+1} = \Pi_G\left(\varphi_k - \frac{1}{8mL}g_{\mu,\delta}(\varphi_k)\right)$;
9. Set $k = k + 1$;

Complexity

Each iteration of the Algorithm needs approximately $\frac{2Qs(p+n)}{\alpha} \ln \frac{2\sqrt{2r}(2\sqrt{2r}+2b)}{\delta}$ a.o., where $s = \max_q s_q$, $p = \max_q p_q$, $n = \max_q n_q$.
Total number of a.o. for the accuracy ε is given by

$$64m(n+p)sQ\frac{LR^2}{\alpha\varepsilon} \ln \left(4(2r + b\sqrt{2r})\frac{32mR\sqrt{L(m+8)}}{\varepsilon^{\frac{3}{2}}\sqrt{2}} \right).$$

Directions for further research

1. Adaptive choice of unknown L, R, μ .
2. Fast Automatic Differentiation or explicit differentiation application.
3. Numerical experiments.

Thank you for your attention!