Gradient methods for convex problems with stochastic inexact oracle

Pavel Dvurechensky[1], Alexander Gasnikov[2]

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

27.06.14

Traditional Mathematical School „Control, Information and Optimization“
Moscow

Outline

Outline

1. Introduction to convex optimization

2. Concept of (δ, L)-oracle

Outline

1. Introduction to convex optimization
2. Concept of (δ, L)-oracle
3. Stochastic inexact oracle

Outline

1. Introduction to convex optimization
2. Concept of (δ, L)-oracle
3. Stochastic inexact oracle
4. Stochastic Intermediate Gradient Method

Outline

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: $N(\varepsilon) \approx O\left(n^5 \ln\left(\frac{1}{\varepsilon}\right)\right)$.

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: $N(\varepsilon) \approx O\left(n^5 \ln\left(\frac{1}{\varepsilon}\right)\right)$.
2. Medium-size problems, n^3 operations per iteration is ok, interior point methods (based on Newton method): $N(\varepsilon) \approx O\left(n^{7/2} \ln\left(\frac{n}{\varepsilon}\right)\right)$.

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: $N(\varepsilon) \approx O\left(n^5 \ln\left(\frac{1}{\varepsilon}\right)\right)$.
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):$N(\varepsilon) \approx O\left(\frac{n^2}{\varepsilon}\right)$.

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: $N(\varepsilon) \approx O\left(n^5 \ln\left(\frac{1}{\varepsilon}\right)\right)$.
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. $N(\varepsilon) \approx O\left(\frac{n}{\varepsilon^2}\right)$.

Notation

Notation

1. E – finite-dimensional real vector space, E^* – its dual.

Notation

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 $\langle g, x \rangle$.

Notation

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.

Notation

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 $\| \cdot \|$: $d(x) \geq \frac{1}{2}\|x - x_0\|^2$, $\forall x \in Q$.

Notation

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 $\| \cdot \|$: $d(x) \geq \frac{1}{2}\|x - x_0\|^2$, $\forall x \in Q$.
   Examples:
   1. Euclidean distance: $Q = \mathbb{R}^n$, $\| \cdot \| = \| \cdot \|_2$, $d(x) = \frac{1}{2}\|x\|_2^2$, $x_0 = 0$.

Notation

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 $\| \cdot \|$: $d(x) \geq \frac{1}{2}\|x - x_0\|^2$, $\forall x \in Q$. Examples:
    1. Euclidean distance: $Q = \mathbb{R}^n$, $\| \cdot \| = \| \cdot \|_2$, $d(x) = \frac{1}{2}\|x\|_2^2$, $x_0 = 0$.
    2. Entropy $Q = \left\{ x \in \mathbb{R}^n : x_i \geq 0, \sum_{i=1}^n x_i = 1 \right\}$, $\| \cdot \| = \| \cdot \|_1$, $d(x) = \ln n + \sum_{i=1}^n x_i \ln x_i$, $x_0 = \left(\frac{1}{n}, \dots, \frac{1}{n} \right)^T$.

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 $\|\cdot\|$: $d(x) \geq \frac{1}{2}\|x - x_0\|^2$, $\forall x \in Q$. 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: $V(x, z) = d(x) - d(z) - \langle \nabla d(z), x - z \rangle$.

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 $\| \cdot \|$: $d(x) \geq \frac{1}{2}\|x - x_0\|^2$, $\forall x \in Q$.
   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: $V(x, z) = \frac{1}{2}\|x - z\|_2^2$.

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 $\|\cdot\|$: $d(x) \geq \frac{1}{2}\|x - x_0\|^2$, $\forall x \in Q$. 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:

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

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:

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

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: $\|g(x)\|_* \leq M$, $\forall x \in Q$, $\forall g(x) \in \partial f(x)$.
   Then $\|g(x) - g(y)\|_* \leq 2M$

Classes of convex functions: smoothness

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

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

Classes of convex functions: smoothness

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

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

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

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

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

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 $\nu \in [0, 1]$
   $\|g(x) - g(y)\|_* \le L_\nu \|x - y\|^\nu$, $\forall x \in Q$, $\forall g(x) \in \partial f(x)$.

Classes of convex functions: smoothness

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

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

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

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

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

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

Problem formulation: simple case

Consider the problem

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

where

Problem formulation: simple case

Consider the problem

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

where

1. $Q \subset E$ is a closed convex set,

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.

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.

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.

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.

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 $(f(x), g(x))$. $R^2 \overset{\text{def}}{\geq} 2d(x^*)$.

Lower complexity bounds

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

1. Nonsmooth Convex Problem: $f(y_k) - f^* \leq O\left(\frac{MR}{\sqrt{k}}\right)$,
   $N(\varepsilon) = \Omega\left(\frac{M^2R^2}{\varepsilon^2}\right)$.

Lower complexity bounds

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

1. Nonsmooth Convex Problem: $f(y_k) - f^* \leq O\left(\frac{MR}{\sqrt{k}}\right)$,
   $N(\varepsilon) = \Omega\left(\frac{M^2R^2}{\varepsilon^2}\right)$.
2. Nonsmooth Strongly Convex Problem: $f(y_k) - f^* \leq O\left(\frac{M}{\mu k}\right)$,
   $N(\varepsilon) = \Omega\left(\frac{M}{\mu\varepsilon}\right)$.

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: $f(y_k) - f^* \leq O\left(\frac{MR}{\sqrt{k}}\right)$,
   $N(\varepsilon) = \Omega\left(\frac{M^2R^2}{\varepsilon^2}\right)$.
2. Nonsmooth Strongly Convex Problem: $f(y_k) - f^* \leq O\left(\frac{M}{\mu k}\right)$,
   $N(\varepsilon) = \Omega\left(\frac{M}{\mu\varepsilon}\right)$.
3. Smooth Convex Problem: $f(y_k) - f^* \leq O\left(\frac{LR^2}{k^2}\right)$, $N(\varepsilon) = \Omega\left(\sqrt{\frac{LR^2}{\varepsilon}}\right)$.

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

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

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

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

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

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

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

Lower complexity bounds

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

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

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

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

Below we consider mostly the smooth case.

Simple Primal Gradient Method

$f(x)$ is smooth.

Simple Primal Gradient Method

$f(x)$ is smooth.

Method:

1. Choose $x_0 \in Q$.

Simple Primal Gradient Method

$f(x)$ is smooth.

Method:

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

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), $y_k = \arg\min_{i=1,\dots,k} f(x_i) = x_k$.

$f(x)$ 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: $f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2k}$.

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 $\{\alpha_i\}_{i\geq 0}$, $\{\beta_i\}_{i\geq 0}$.

Estimating functions

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

- $d(x) = \frac{1}{2}\|x - x_0\|_2^2$

Estimating functions

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

- $d(x) = \frac{1}{2}\|x - x_0\|_2^2$
- $\Psi_k(x) =$
  $\beta_k d(x) + \sum_{i=0}^{k} \alpha_i \left[f(x_i) + \langle \nabla f(x)(x_i), x - x_i \rangle + \frac{\mu}{2}\|x - x_i\|_2^2 \right]$

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.

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. $A_k = \sum_{i=0}^{k} \alpha_i$.

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

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

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

$$A_k f(y_k) \leq \Psi_k^* \leq$$

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

$$A_k f(y_k) \leq \Psi_k^* \leq \Psi_k(x^*) \leq$$

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

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

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 $x_0 \in Q$.

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. $w_k = \pi_Q \left(x_k - \frac{1}{L} \nabla f(x_k) \right)$.

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

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

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

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

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

Rate of convergence:

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

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

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

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 $x_0 \in Q$.

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. $y_k = \pi_Q\left(x_k - \frac{1}{L}\nabla f(x_k)\right)$.

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

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

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. $z_k = \arg\min_{x \in Q} \Psi_k(x) = \arg\min_{x \in Q}\{\frac{L}{2}\|x - x_0\|_2^2 + \sum_{i=0}^{k} \alpha_i \left[f(x_i) + \langle \nabla f(x)(x_i), x - x_i \rangle + \frac{\mu}{2}\|x - x_i\|_2^2\right]\}$.
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: $f(y_k) - f^* \leq \frac{4L\|x_0 - x^*\|_2^2}{k^2}$.

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. $z_k = \arg\min_{x\in Q} \Psi_k(x) = \arg\min_{x\in Q}\{\frac{L}{2}\|x - x_0\|_2^2 + \sum_{i=0}^{k} \alpha_i \left[f(x_i) + \langle \nabla f(x)(x_i), x - x_i\rangle + \frac{\mu}{2}\|x - x_i\|_2^2\right]\}$.
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: $f(y_k) - f^* \le \frac{4L\|x_0 - x^*\|_2^2}{k^2}$.
2. Strongly convex case: $f(y_k) - f^* \le \frac{L\|x_0 - x^*\|_2^2}{2} \exp\left(-\frac{k}{2}\sqrt{\frac{\mu}{L}}\right)$.

Generalizations

1. Non-Euclidean setup, e.g. Q – simplex, $d(x)$ – entropy.

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:

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:

$$\hat{x}_i = \frac{exp(-g_i)}{\sum_{i=1}^{n} exp(-g_i)}, \quad i = 1, ..., n.$$

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.

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 $\|x - a\|_2^2 + \lambda\|x\|_1 \to \min$.

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 $O\left(\frac{1}{k}\right)$.

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 $O\left(\exp\left(-k \cdot \mathrm{const}\right)\right)$.

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 $O\left(\exp\left(-k \cdot \mathrm{const}\right)\right)$.
3. Stochastic error, e.g. $\mathbb{E}_\xi f(x, \xi) \to \min_{x \in Q}$.

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 $O\left(\exp\left(-k\cdot\mathrm{const}\right)\right)$.
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
   $\mathbb{E}_{\xi_k}\|\nabla f(x,\xi_k) - \nabla\mathbb{E}_{\xi_k} f(x,\xi_k)\|_*^2 \le \sigma^2$.

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 $O\left(\exp\left(-k\cdot\mathrm{const}\right)\right)$.
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).

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 $O\left(\exp\left(-k \cdot \mathrm{const}\right)\right)$.
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 .

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

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

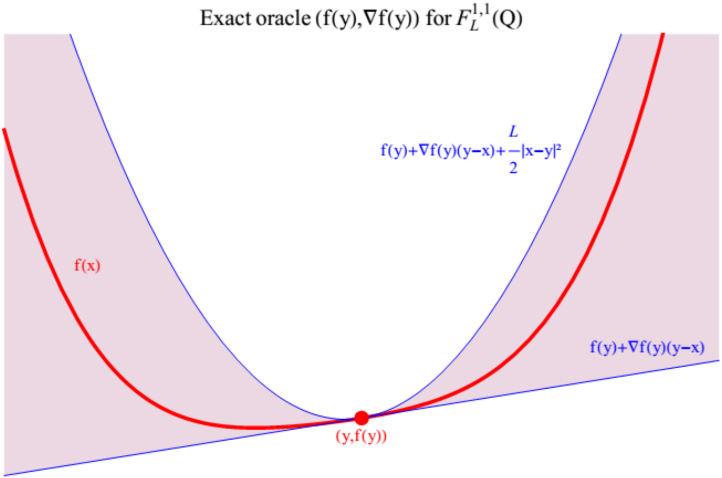
(δ, L)-oracle

[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

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

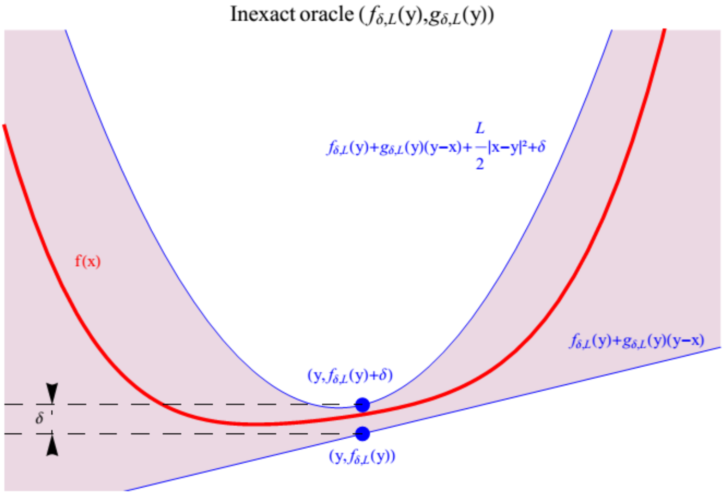
Usual oracle $(f(x), \nabla f(x))$ is replaced by (δ, L)-oracle $(f_{\delta,L}(x), g_{\delta,L}(x))$.

(δ, L)-oracle geometry



Picture from O.Devolder PhD Thesis, 2013

(δ, L)-oracle geometry



Picture from O.Devolder PhD Thesis, 2013

(δ, L)-oracle examples

1. $f(x)$ – convex non-smooth function with Hölder continuous subgradient $\|g(x) - g(y)\|_* \leq L_\nu \|x - y\|^\nu, \quad \forall x, y \in Q$, for some $\nu \in [0, 1]$

(δ, L)-oracle examples

1. $f(x)$ – convex non-smooth function with Hölder continuous subgradient $\|g(x) - g(y)\|_* \leq L_\nu \|x - y\|^\nu, \quad \forall x, y \in Q$, for some $\nu \in [0, 1]$
   Then for any $\delta > 0$ $(f(x), g(x))$ is (δ, L)-oracle, where
   $L = L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}}$.

(δ, L)-oracle examples

1. $f(x)$ – convex non-smooth function with Hölder continuous subgradient $\|g(x) - g(y)\|_* \leq L_\nu \|x - y\|^\nu, \quad \forall x, y \in Q$, for some $\nu \in [0, 1]$
   Then for any $\delta > 0$ $(f(x), g(x))$ is (δ, L)-oracle, where
   $L = L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}}$.
   When $\nu = 0$ we have $L = \frac{L_0^2}{2\delta}$.

(δ, L)-oracle examples

1. $f(x)$ – convex non-smooth function with Hölder continuous subgradient $\|g(x) - g(y)\|_* \leq L_\nu \|x - y\|^\nu, \quad \forall x, y \in Q$, for some $\nu \in [0, 1]$
   Then for any $\delta > 0$ $(f(x), g(x))$ is (δ, L)-oracle, where $L = L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}}$.
   When $\nu = 0$ we have $L = \frac{L_0^2}{2\delta}$.
2. $f(x)$ – smooth with M-Lipschitz continuous gradient.

(δ, L)-oracle examples

1. $f(x)$ – convex non-smooth function with Hölder continuous subgradient $\|g(x) - g(y)\|_* \leq L_\nu \|x - y\|^\nu, \quad \forall x, y \in Q$, for some $\nu \in [0, 1]$
   Then for any $\delta > 0$ $(f(x), g(x))$ is (δ, L)-oracle, where
   $L = L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}}$.
   When $\nu = 0$ we have $L = \frac{L_0^2}{2\delta}$.
2. $f(x)$ – smooth with M-Lipschitz continuous gradient.
   At each point $x \in Q$ we can get
   $\tilde{f}_x : \quad |f(x) - \tilde{f}_x| \leq \Delta_1$ and $\tilde{\nabla} f_x : \quad \|\nabla f(x) - \tilde{\nabla} f_x\| \leq \Delta_2$.

(δ, L)-oracle examples

1. $f(x)$ – convex non-smooth function with Hölder continuous subgradient $\|g(x) - g(y)\|_* \leq L_\nu \|x - y\|^\nu, \quad \forall x, y \in Q$, for some $\nu \in [0, 1]$
   Then for any $\delta > 0$ $(f(x), g(x))$ is (δ, L)-oracle, where
   $L = L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}}$.
   When $\nu = 0$ we have $L = \frac{L_0^2}{2\delta}$.
2. $f(x)$ – smooth with M-Lipschitz continuous gradient.
   At each point $x \in Q$ we can get
   $\tilde{f}_x : \quad |f(x) - \tilde{f}_x| \leq \Delta_1$ and $\tilde{\nabla} f_x : \quad \|\nabla f(x) - \tilde{\nabla} f_x\| \leq \Delta_2$.
   If $\mathrm{diam}\, Q = D < +\infty$ then $(\tilde{f}_x - \Delta_1 - \Delta_2 D, \tilde{\nabla} f_x)$ is a $(2\Delta_1 + 2\Delta_2 D, M)$-oracle.

Convex case: PGM, DGM and FGM revisited

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

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

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

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, $f(y_k) - f^* \leq \frac{2L\|x_0 - x^*\|_2^2}{(k+1)^2} + \frac{1}{3}(k+3)\delta$.

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

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

(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 $f(x) \to f_{\delta,L,\mu}(x)$, $\nabla f(x) \to g_{\delta,L,\mu}(x)$.

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, $y_k = \arg\min_{i=1,\dots,k} f(x_i)$,
   $f(y_k) - f^* \leq \frac{L\|x_0 - x^*\|_2^2}{2} \exp\left(-k\frac{\mu}{L}\right) + \delta$.

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

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

(δ, L)-oracle application 1

As we have seen $f(x)$ – convex non-smooth function with Hölder continuous subgradient can be for any $\delta > 0$ equipped with (δ, L)-oracle with $L = L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}}$.

(δ, L)-oracle application 1

As we have seen $f(x)$ – convex non-smooth function with Hölder continuous subgradient can be for any $\delta > 0$ equipped with (δ, L)-oracle with $L = L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}}$.
Then for N iterations of FGM we get

$$f(y_N) - f^* \leq \frac{2R^2}{(N+1)^2} L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}} + \frac{1}{3}(N+1)\delta$$

(δ, L)-oracle application 1

As we have seen $f(x)$ – convex non-smooth function with Hölder continuous subgradient can be for any $\delta > 0$ equipped with (δ, L)-oracle with $L = L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}}$.
Then for N iterations of FGM we get

$$f(y_N) - f^* \leq \frac{2R^2}{(N+1)^2} L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}} + \frac{1}{3}(N+1)\delta$$

Optimizing in δ we get

$$\delta_N = \frac{1-\nu}{1+\nu} \frac{L_\nu R^{1+\nu}}{(N+1)^{\frac{3}{2}(1+\nu)}} 2^\nu, \quad L = \frac{L_\nu (N+1)^{\frac{3(1-\nu)}{2}}}{2^{1-\nu} R^{1-\nu}},$$

(δ, L)-oracle application 1

As we have seen $f(x)$ – convex non-smooth function with Hölder continuous subgradient can be for any $\delta > 0$ equipped with (δ, L)-oracle with $L = L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}}$.

Then for N iterations of FGM we get

$$f(y_N) - f^* \leq \frac{2R^2}{(N+1)^2} L_\nu \left[\frac{L_\nu(1-\nu)}{2\delta(1+\nu)}\right]^{\frac{1-\nu}{1+\nu}} + \frac{1}{3}(N+1)\delta$$

Optimizing in δ we get

$$\delta_N = \frac{1-\nu}{1+\nu} \frac{L_\nu R^{1+\nu}}{(N+1)^{\frac{3}{2}(1+\nu)}} 2^\nu, \quad L = \frac{L_\nu (N+1)^{\frac{3(1-\nu)}{2}}}{2^{1-\nu} R^{1-\nu}},$$

and optimal for every $\nu \in [0,1]$ rate

$$f(y_N) - f^* \leq \frac{2^{1+\nu} L_\nu R^{1+\nu}}{1+\nu} \left(\frac{1}{N+1}\right)^{\frac{1+3\nu}{2}}$$

(δ, L)-oracle application 2

$f(x) = f_1(x) + f_2(x)$, 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.

(δ, L)-oracle application 2

$f(x) = f_1(x) + f_2(x)$, 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 $L = L_1 + \frac{M_2^2}{2\delta}$.

(δ, L)-oracle application 2

$f(x) = f_1(x) + f_2(x)$, 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

$$f(y_N) - f^* \leq \frac{2L_1R^2}{(N+1)^2} + \frac{2M_2R}{\sqrt{N+1}}.$$

(δ, L)-oracle application 3

Theorem [Devolder, Glineur and Nesterov]

Consider a first-order method for the class of L-smooth functions with convergence rate $O\left(\frac{LR^2}{k^p}\right)$ when exact first-order information is used. Assume that the bounds on the performance of this method applied to a problem equipped with an inexact (δ, L)-oracle are given by inequality

$$f(y_k) - f^* \leq \frac{C_1 LR^2}{k^p} + C_2 k^q \delta,$$

where C_1, C_2 are absolute constants. Then the inequality $q \geq p - 1$ must hold.

(δ, L)-oracle application 3

Theorem [Devolder, Glineur and Nesterov]

Consider a first-order method for the class of L-smooth functions with convergence rate $O\left(\frac{LR^2}{k^p}\right)$ when exact first-order information is used. Assume that the bounds on the performance of this method applied to a problem equipped with an inexact (δ, L)-oracle are given by inequality

$$f(y_k) - f^* \leq \frac{C_1 LR^2}{k^p} + C_2 k^q \delta,$$

where C_1, C_2 are absolute constants. Then the inequality $q \geq p - 1$ must hold.

Examples:

PGM $p = 1$ and $q = 0$.

(δ, L)-oracle application 3

Theorem [Devolder, Glineur and Nesterov]

Consider a first-order method for the class of L-smooth functions with convergence rate $O\left(\frac{LR^2}{k^p}\right)$ when exact first-order information is used. Assume that the bounds on the performance of this method applied to a problem equipped with an inexact (δ, L)-oracle are given by inequality

$$f(y_k) - f^* \leq \frac{C_1 LR^2}{k^p} + C_2 k^q \delta,$$

where C_1, C_2 are absolute constants. Then the inequality $q \geq p - 1$ must hold.

Examples:
PGM $p = 1$ and $q = 0$.
FGM $p = 2$ and $q = 1$.

Intermediate gradient method

Devolder, Glineur and Nesterov proposed a method with intermediate rate.
Let $\{\alpha_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ be two sequences of coefficients satisfying

Intermediate gradient method

Devolder, Glineur and Nesterov proposed a method with intermediate rate.
Let $\{\alpha_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ be two sequences of coefficients satisfying

$$0 \leq \alpha_i \leq B_i, \quad \forall i \geq 0, \quad \alpha_k^2 \leq B_k \leq \sum_{i=0}^{k} \alpha_i, \quad \forall k \geq 1$$

Intermediate gradient method

Devolder, Glineur and Nesterov proposed a method with intermediate rate.
Let $\{\alpha_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ be two sequences of coefficients satisfying

$$0 \leq \alpha_i \leq B_i, \quad \forall i \geq 0, \quad \alpha_k^2 \leq B_k \leq \sum_{i=0}^{k} \alpha_i, \quad \forall k \geq 1$$

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

Intermediate gradient method

Devolder, Glineur and Nesterov proposed a method with intermediate rate.
Let $\{\alpha_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ be two sequences of coefficients satisfying

$$0 \leq \alpha_i \leq B_i, \quad \forall i \geq 0, \quad \alpha_k^2 \leq B_k \leq \sum_{i=0}^{k} \alpha_i, \quad \forall k \geq 1$$

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

Method:

1. Choose $x_0 \in Q$.

Intermediate gradient method

Devolder, Glineur and Nesterov proposed a method with intermediate rate.
Let $\{\alpha_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ be two sequences of coefficients satisfying

$$0 \leq \alpha_i \leq B_i, \quad \forall i \geq 0, \quad \alpha_k^2 \leq B_k \leq \sum_{i=0}^{k} \alpha_i, \quad \forall k \geq 1$$

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

Method:

1. Choose $x_0 \in Q$.
2. $w_k = \pi_Q \left(x_k - \frac{1}{L} g_{\delta,L}(x_k) \right)$.

Intermediate gradient method

Devolder, Glineur and Nesterov proposed a method with intermediate rate.
Let $\{\alpha_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ be two sequences of coefficients satisfying

$$0 \leq \alpha_i \leq B_i, \quad \forall i \geq 0, \quad \alpha_k^2 \leq B_k \leq \sum_{i=0}^{k} \alpha_i, \quad \forall k \geq 1$$

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

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

Intermediate gradient method

Devolder, Glineur and Nesterov proposed a method with intermediate rate.
Let $\{\alpha_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ be two sequences of coefficients satisfying

$$0 \leq \alpha_i \leq B_i, \quad \forall i \geq 0, \quad \alpha_k^2 \leq B_k \leq \sum_{i=0}^{k} \alpha_i, \quad \forall k \geq 1$$

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

Method:

1. Choose $x_0 \in Q$.
2. $w_k = \pi_Q \left(x_k - \frac{1}{L} g_{\delta,L}(x_k)\right)$.
3. $z_k = \arg\min_{x\in Q} \Psi_k(x) =$
   $\arg\min_{x\in Q}\{\frac{L}{2}\|x - x_0\|_2^2 + \sum_{i=0}^k \alpha_i \left[f_{\delta,L}(x_i) + \langle g_{\delta,L}(x_i), x - x_i\rangle\right]\}$.
4. $y_k = \frac{A_k - B_k}{A_k} y_{k-1} + \frac{B_k}{A_k} w_k$.

Intermediate gradient method

Devolder, Glineur and Nesterov proposed a method with intermediate rate.
Let $\{\alpha_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ be two sequences of coefficients satisfying

$$0 \leq \alpha_i \leq B_i, \quad \forall i \geq 0, \quad \alpha_k^2 \leq B_k \leq \sum_{i=0}^{k} \alpha_i, \quad \forall k \geq 1$$

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

Method:

1. Choose $x_0 \in Q$.
2. $w_k = \pi_Q \left(x_k - \frac{1}{L} g_{\delta,L}(x_k)\right)$.
3. $z_k = \arg\min_{x\in Q} \Psi_k(x) =$
   $\arg\min_{x\in Q}\{\frac{L}{2}\|x - x_0\|_2^2 + \sum_{i=0}^{k} \alpha_i \left[f_{\delta,L}(x_i) + \langle g_{\delta,L}(x_i), x - x_i\rangle\right]\}$.
4. $y_k = \frac{A_k - B_k}{A_k} y_{k-1} + \frac{B_k}{A_k} w_k$.
5. $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$, $\tau_k = \frac{\alpha_{k+1}}{B_{k+1}}$.

Intermediate gradient method

Devolder, Glineur and Nesterov proposed a method with intermediate rate.
Let $\{\alpha_i\}_{i\geq 0}$, $\{B_i\}_{i\geq 0}$ be two sequences of coefficients satisfying

$$0 \leq \alpha_i \leq B_i, \quad \forall i \geq 0, \quad \alpha_k^2 \leq B_k \leq \sum_{i=0}^{k} \alpha_i, \quad \forall k \geq 1$$

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

Method:

1. Choose $x_0 \in Q$.
2. $w_k = \pi_Q \left(x_k - \frac{1}{L} g_{\delta,L}(x_k) \right)$.
3. $z_k = \arg\min_{x\in Q} \Psi_k(x) =$
   $\arg\min_{x\in Q} \{ \frac{L}{2} \|x - x_0\|_2^2 + \sum_{i=0}^{k} \alpha_i \left[f_{\delta,L}(x_i) + \langle g_{\delta,L}(x_i), x - x_i \rangle \right] \}$.
4. $y_k = \frac{A_k - B_k}{A_k} y_{k-1} + \frac{B_k}{A_k} w_k$.
5. $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$, $\tau_k = \frac{\alpha_{k+1}}{B_{k+1}}$.

Choose $\alpha_i = \left(\frac{i+p}{p}\right)^{p-1}, \quad B_i = \alpha_i^2$. Then rate of convergence for convex case:

$$f(y_k) - f^* \leq \Theta\left(\frac{L\|x_0 - x^*\|_2^2}{2k^p} \right) + \Theta(k^{p-1}\delta)$$

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

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

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

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

Instead of $(f_{\delta,L}(x), g_{\delta,L}(x))$ ((δ, L)-oracle) we use their stochastic approximations $(F_{\delta,L}(x,\xi), G_{\delta,L}(x,\xi))$.

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 $\Xi \subset \mathbb{R}^d$ 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 $\mathbb{E}_\xi f(x,\xi) \to \min_{x \in Q}$.

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
   $\mathbb{E}_{\xi_k} \|\nabla f(x,\xi_k) - \nabla \mathbb{E}_{\xi_k} f(x,\xi_k)\|_*^2 \leq \sigma^2$.
   Here $\delta = 0$.

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 $A \in \mathbb{R}^{N \times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^N$.

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 $A \in \mathbb{R}^{N\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^N$.
   $\nabla f(x) = A^T Ax - A^T b = \sum_{i=1}^N (x^T a_i - b_i)a_i$ is very difficult to calculate when N is very large.

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 $A \in \mathbb{R}^{N\times n}$, $x \in \mathbb{R}^n$, $b \in \mathbb{R}^N$.
   $\nabla f(x) = A^T Ax - A^T b = \sum_{i=1}^N (x^T a_i - b_i)a_i$ is very difficult to calculate when N is very large.
   The idea is to replace $\nabla f(x)$ by $G_{0,L}(x,\xi) = \frac{N}{M}\sum_{j=1}^M (x^T a_{\xi_j} - b_{\xi_j})a_{\xi_j}$, where $\{\xi_1,\dots,\xi_M\}$ is a subset of rows uniformly chosen from $\{1,\dots,N\}$.
   Here $\delta = 0$.

Stochastic Primal Gradient Method

[Devolder, Glineur and Nesterov, 2011-2013]
Choose $d(x) = \frac{1}{2}\|x - x_0\|_2^2$, $\beta_k > L$, $\gamma_k = \frac{1}{\beta_k}$.

Stochastic Primal Gradient Method

[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 $x_0 \in Q$.

Stochastic Primal Gradient Method

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

Stochastic Primal Gradient Method

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

Stochastic Primal Gradient Method

[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 $\mathbb{E} f(y_N) - f^* \leq \frac{LR^2}{2N} + \frac{3\sigma R}{2\sqrt{N}} + \delta$.

Stochastic Primal Gradient Method

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

Stochastic Primal Gradient Method

[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.

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.

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

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

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

$$A_k f(y_k) \leq \Psi_k^* + E_k \leq$$

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

$$A_k f(y_k) \leq \Psi_k^* + E_k \leq \Psi_k(x^*) + E_k \leq$$

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

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

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 $x_0 \in Q$.

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. $w_k = \pi_Q \left(x_k - \frac{1}{\beta_k} G_{\delta,L}(x_k, \xi_k) \right)$.

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

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

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

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

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

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

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

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:

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

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

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

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 $x_0 \in Q$.

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. $y_k = \pi_Q \left(x_k - \frac{1}{\beta_k} G_{\delta,L}(x_k, \xi_k) \right)$.

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

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. $x_{k+1} = \tau_k z_k + (1 - \tau_k) y_k$, $\tau_k = \frac{\alpha_{k+1}}{A_{k+1}}$.

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

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:

$$\mathbb{E} f(y_k) - f^* \leq \frac{2^{3/2} L R^2}{(k+1)(k+2)} + \frac{2^{9/4}(k+3)^{3/2}\sigma R}{\sqrt{3}(k+1)(k+2)} + \frac{1}{3}(k+3)\delta.$$

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

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

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,

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,

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 $\mathrm{const} \cdot \frac{1}{\sqrt{k}}$.

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:

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)

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

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)

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

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 $p \in [1, 2]$

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,

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),

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

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

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

Auxiliary objects

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

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

Auxiliary objects

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

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

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 $d(x)$, $V(x,z)$.

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.

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. $x_0 = \arg\min_{x\in Q}\{d(x)\}$.

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. $$y_0 = \arg\min_{x\in Q}\{\beta_0 d(x) + \alpha_0\langle G_{\delta,L}(x_0,\xi_0), x - x_0\rangle + h(x)\}$$

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

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 $k = 0, 1, \dots$ 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. $$x_{k+1} = \tau_k z_k + (1-\tau_k)y_k$$

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. $$\hat{x}_{k+1} = \arg\min_{x\in Q}\{\beta_k V(x,z_k) + \alpha_{k+1}\langle G_{\delta,L}(x_{k+1},\xi_{k+1}), x - z_k\rangle + \alpha_{k+1} h(x)\}.$$

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 $k = 0, 1, \dots$ 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. $$w_{k+1} = \tau_k \hat{x}_{k+1} + (1-\tau_k) y_k$$

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 $k = 0, 1, \dots$ 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.

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.

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

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

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

  $$A_k\varphi(y_k) \leq \Psi_k^* + E_k \leq$$

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

$$A_k \varphi(y_k) \leq \Psi_k^* + E_k \leq \Psi_k(x^*) + E_k \leq$$

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

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.

E_k and $\bar{E}_k(x)$

We managed to prove that

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

$$+\sum_{i=0}^{k} \alpha_i (f_{\delta,L}(x_i) - F_{\delta,L}(x_i,\xi_i)) +$$

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

$$\bar{E}_k(x) = \sum_{i=0}^{k} \alpha_i \Big[F_{\delta,L}(x_i,\xi_i) - f_{\delta,L}(x_i) +$$

$$\langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), x - x_i\rangle\Big].$$

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

$$
\begin{aligned}
\varphi(y_k) - \varphi^* \leq \frac{1}{A_k}\Bigg(&\beta_k d(x^*) + \sum_{i=0}^{k} B_i \delta + \\
&+ \sum_{i=0}^{k} \frac{B_i}{\beta_i - L} \|G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i)\|_*^2 + \\
&+ \sum_{i=0}^{k} \alpha_i \langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), x^* - x_i\rangle + \\
&+ \sum_{i=1}^{k} (B_i - \alpha_i)\frac{\alpha_i}{B_i} \langle G_{\delta,L}(x_i,\xi_i) - g_{\delta,L}(x_i), y_{i-1} - z_{i-1}\rangle\Bigg).
\end{aligned}
$$

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.

Additional assumptions

1. $\xi_0, \dots, \xi_k$ are i.i.d random variables.
2. $G_{\delta,L}(x,\xi)$ satisfies the condition
   $\mathbb{E}_\xi \left[\exp\left(\frac{\|G_{\delta,L}(x,\xi) - g_{\delta,L}(x)\|_*^2}{\sigma^2}\right)\right] \leq \exp(1)$.

Additional assumptions

1. $\xi_0, \ldots, \xi_k$ 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 $D = \max_{x,y \in Q} \|x - y\|$.

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.

Auxiliary result 1

Lemma 1 [Juditsky, Lan, Nemirovski, Shapiro]

Let $\xi_0, \ldots, \xi_k$ be a sequence of realizations of the i.i.d. random variables $X_0, \ldots, X_k$ and let $\Delta_i = \Delta_i(\xi_{[i]})$ be a deterministic function of $\xi_{[i]}$ such that for all $i \geq 0$:

$$\mathbb{E}\left[\exp\left(\frac{\Delta_i^2}{\sigma^2}\right)|\xi_{[i-1]}\right] \leq \exp(1)$$

and $c_0, \ldots, c_k$ is a sequence of positive coefficients. Then we have for any $k \geq 0$ and any $\Omega \geq 0$:

$$\mathrm{P}\left(\sum_{i=0}^{k} c_i \Delta_i^2 \geq (1+\Omega)\sum_{i=0}^{k} c_i \sigma^2\right) \leq \exp(-\Omega).$$

Auxiliary result 2

Lemma 2 [Lan, Nemirovski, Shapiro]

Let $\xi_0, \ldots, \xi_k$ be a sequence of realizations of the i.i.d. random variables $X_0, \ldots, X_k$ and let Γ_i and η_i be a deterministic function of $\xi_{[i]}$ such that:

1. $\mathbb{E}\left[\Gamma_i | \xi_{[i-1]}\right] = 0.$
2. $|\Gamma_i| \leq c_i \eta_i$, where c_i is positive deterministic constant.
3. $\mathbb{E}\left[\exp\left(\frac{\eta_i^2}{\sigma^2}\right) | \xi_{[i-1]}\right] \leq \exp(1).$

Then for any $k \geq 0$ and any $\Omega \geq 0$:

$$P\left(\sum_{i=0}^{k} \Gamma_i \geq \sqrt{3\Omega}\sigma \sqrt{\sum_{i=0}^{k} c_i^2}\right) \leq \exp(-\Omega).$$

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

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 $b = 2^{\frac{5-2p}{4}} p^{\frac{1-2p}{2}}$, $R \geq \sqrt{2d(x^*)}$.

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:

$$\mathrm{P}\Bigg(\varphi(y_k) - \varphi^* >$$

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

$$+ 2^{2p-1}\left(\left(\frac{k+p}{p}\right)^{p-1} + 1\right)\delta + \frac{2D\sigma\sqrt{6p\Omega}}{\sqrt{k+p}}\Bigg) \leq$$

$$\leq 3\exp(-3\Omega).$$

SIGM: mean complexity

For SIGM we have ($\sqrt{V(x_0, x^*)} \le R$).

$$\mathbb{E}\varphi(y_k) - \varphi^* \le \frac{c_1 L V(x_0, x^*)}{k^p} + \frac{c_2 \sigma R}{\sqrt{k}} + c_3 k^{p-1} \delta.$$

SIGM: mean complexity

For SIGM we have ($\sqrt{V(x_0, x^*)} \leq R$).

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \frac{c_1 L V(x_0, x^*)}{k^p} + \frac{c_2 \sigma R}{\sqrt{k}} + c_3 k^{p-1}\delta.$$

If we choose $k \geq \max\left\{ \left(\frac{2c_1 L V(x_0,x^*)}{\varepsilon}\right)^{\frac{1}{p}}, \frac{4c_2^2\sigma^2 R^2}{\varepsilon^2} \right\}$,

SIGM: mean complexity

For SIGM we have ($\sqrt{V(x_0, x^*)} \leq R$).

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \frac{c_1 L V(x_0, x^*)}{k^p} + \frac{c_2 \sigma R}{\sqrt{k}} + c_3 k^{p-1}\delta.$$

If we choose $k \geq \max\left\{ \left(\frac{2c_1 LV(x_0,x^*)}{\varepsilon} \right)^{\frac{1}{p}}, \frac{4c_2^2\sigma^2 R^2}{\varepsilon^2} \right\}$, then

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \varepsilon + c_3 k^{p-1}\delta.$$

SIGM: mean complexity

For SIGM we have ($\sqrt{V(x_0, x^*)} \leq R$).

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \frac{c_1 L V(x_0, x^*)}{k^p} + \frac{c_2 \sigma R}{\sqrt{k}} + c_3 k^{p-1} \delta.$$

If we choose $k \geq \max\left\{ \left(\frac{2c_1 L V(x_0, x^*)}{\varepsilon} \right)^{\frac{1}{p}}, \frac{4c_2^2 \sigma^2 R^2}{\varepsilon^2} \right\}$, then

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \varepsilon + c_3 k^{p-1} \delta.$$

The goal is to obtain for μ–strongly convex function

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \varepsilon + \tilde{c}_3 \left(\frac{L}{\mu} \right)^{\frac{p-1}{p}} \delta$$

in

SIGM: mean complexity

For SIGM we have ($\sqrt{V(x_0, x^*)} \leq R$).

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \frac{c_1 L V(x_0, x^*)}{k^p} + \frac{c_2 \sigma R}{\sqrt{k}} + c_3 k^{p-1}\delta.$$

If we choose $k \geq \max\left\{ \left(\frac{2c_1 L V(x_0, x^*)}{\varepsilon}\right)^{\frac{1}{p}}, \frac{4c_2^2\sigma^2 R^2}{\varepsilon^2} \right\}$, then

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \varepsilon + c_3 k^{p-1}\delta.$$

The goal is to obtain for μ–strongly convex function

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \varepsilon + \tilde{c}_3 \left(\frac{L}{\mu}\right)^{\frac{p-1}{p}} \delta$$

in

$$N(\varepsilon) = \tilde{c}_1 \left(\frac{L}{\mu}\right)^{\frac{1}{p}} \ln \frac{\mu R_0^2}{\varepsilon} + \tilde{c}_2 \frac{\sigma^2}{\mu\varepsilon}$$

oracle calls.

SIGM: strongly convex case

Assume that $\varphi(x)$ is strongly convex. Then

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

SIGM: strongly convex case

Assume that $\varphi(x)$ is strongly convex. Then

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

Assume that $V(x,z)$ has quadratic growth with parameter $\varkappa \geq 1$, i.e.:

$$V(x,z) \leq \frac{\varkappa}{2}\|x - z\|^2, \forall x, z \in Q.$$

SIGM: strongly convex case

Assume that $\varphi(x)$ is strongly convex. Then

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

Assume that $V(x,z)$ has quadratic growth with parameter $\varkappa \geq 1$, i.e.:

$$V(x,z) \leq \frac{\varkappa}{2}\|x - z\|^2, \forall x, z \in Q.$$

Then

$$\mathbb{E}V(y_k, x^*) \leq \frac{\varkappa}{\mu}\left(\mathbb{E}\varphi(y_k) - \varphi(x^*)\right).$$

SIGM: strongly convex case

Assume that $\varphi(x)$ is strongly convex. Then

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

Assume that $V(x,z)$ has quadratic growth with parameter $\varkappa \geq 1$, i.e.:

$$V(x,z) \leq \frac{\varkappa}{2}\|x - z\|^2, \forall x, z \in Q.$$

Then

$$\mathbb{E}V(y_k, x^*) \leq \frac{\varkappa}{\mu}\left(\mathbb{E}\varphi(y_k) - \varphi(x^*)\right).$$

Let us change $G_{\delta,L}(x,\xi_j) \to \tilde{G}_{\delta,L}(x,\Xi) = \frac{1}{m}\sum_{j=1}^m G_{\delta,L}(x,\xi_j)$.

SIGM: strongly convex case

Assume that $\varphi(x)$ is strongly convex. Then

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

Assume that $V(x,z)$ has quadratic growth with parameter $\varkappa \geq 1$, i.e.:

$$V(x,z) \leq \frac{\varkappa}{2}\|x - z\|^2, \forall x, z \in Q.$$

Then

$$\mathbb{E}V(y_k, x^*) \leq \frac{\varkappa}{\mu}\left(\mathbb{E}\varphi(y_k) - \varphi(x^*)\right).$$

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

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \frac{c_1 L V(x_0, x^*)}{k^p} + \frac{c_2 \sigma R}{\sqrt{mk}} + c_3 k^{p-1}\delta.$$

SIGM: strongly convex case

Assume that $\varphi(x)$ is strongly convex. Then

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

Assume that $V(x,z)$ has quadratic growth with parameter $\varkappa \geq 1$, i.e.:

$$V(x,z) \leq \frac{\varkappa}{2}\|x - z\|^2, \forall x, z \in Q.$$

Then

$$\mathbb{E}V(y_k, x^*) \leq \frac{\varkappa}{\mu}\left(\mathbb{E}\varphi(y_k) - \varphi(x^*)\right).$$

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

$$\mathbb{E}\varphi(y_k) - \varphi^* \leq \frac{c_1 L V(x_0, x^*)}{k^p} + \frac{c_2 \sigma R}{\sqrt{mk}} + c_3 k^{p-1}\delta.$$

Let us use restart technique.

SIGM: strongly convex case

Denote $\Delta_k = \varphi(y_{N_k}) - \varphi^*$. Let us perform N_0 iterations with $R^2 = \frac{\varkappa}{\mu}\Delta_0 \stackrel{\text{def}}{=} R_0^2$ and $m = m_0$.

SIGM: strongly convex case

Denote $\Delta_k = \varphi(y_{N_k}) - \varphi^*$. Let us perform N_0 iterations with $R^2 = \frac{\varkappa}{\mu}\Delta_0 \overset{\text{def}}{=} R_0^2$ and $m = m_0$. Then

$$\mathbb{E}\Delta_1 \leq \frac{c_1 L \varkappa \Delta_0}{\mu N_0^p} + \frac{c_2 \sigma \sqrt{\frac{\varkappa}{\mu}\Delta_0}}{\sqrt{m_0 N_0}} + c_3 N_0^{p-1}\delta$$

SIGM: strongly convex case

Denote $\Delta_k = \varphi(y_{N_k}) - \varphi^*$. Let us perform N_0 iterations with $R^2 = \frac{\varkappa}{\mu}\Delta_0 \stackrel{\text{def}}{=} R_0^2$ and $m = m_0$. Then

$$\mathbb{E}\Delta_1 \leq \frac{c_1 L\varkappa\Delta_0}{\mu N_0^p} + \frac{c_2\sigma\sqrt{\frac{\varkappa}{\mu}\Delta_0}}{\sqrt{m_0 N_0}} + c_3 N_0^{p-1}\delta$$

Let us choose $N_0 = \left(\frac{4c_1 L\varkappa}{\mu}\right)^{\frac{1}{p}}$ and $m_0 \geq \max\left\{1, \frac{16c_2^2\sigma^2\varkappa}{\mu N_0\Delta_0}\right\}$.

Denote $\Delta_k = \varphi(y_{N_k}) - \varphi^*$. Let us perform N_0 iterations with $R^2 = \frac{\varkappa}{\mu}\Delta_0 \overset{\text{def}}{=} R_0^2$ and $m = m_0$. Then

$$\mathbb{E}\Delta_1 \leq \frac{c_1 L\varkappa\Delta_0}{\mu N_0^p} + \frac{c_2\sigma\sqrt{\frac{\varkappa}{\mu}\Delta_0}}{\sqrt{m_0 N_0}} + c_3 N_0^{p-1}\delta$$

Let us choose $N_0 = \left(\frac{4c_1 L\varkappa}{\mu}\right)^{\frac{1}{p}}$ and $m_0 \geq \max\left\{1, \frac{16c_2^2\sigma^2\varkappa}{\mu N_0 \Delta_0}\right\}$. Denote $\hat{N} \overset{\text{def}}{=} \left(\frac{4c_1 L\varkappa}{\mu}\right)^{\frac{p-1}{p}}$

SIGM: strongly convex case

Denote $\Delta_k = \varphi(y_{N_k}) - \varphi^*$. Let us perform N_0 iterations with $R^2 = \frac{\varkappa}{\mu}\Delta_0 \stackrel{\text{def}}{=} R_0^2$ and $m = m_0$. Then

$$\mathbb{E}\Delta_1 \leq \frac{c_1 L \varkappa \Delta_0}{\mu N_0^p} + \frac{c_2 \sigma \sqrt{\frac{\varkappa}{\mu}\Delta_0}}{\sqrt{m_0 N_0}} + c_3 N_0^{p-1}\delta$$

Let us choose $N_0 = \left(\frac{4c_1 L\varkappa}{\mu}\right)^{\frac{1}{p}}$ and $m_0 \geq \max\left\{1, \frac{16c_2^2\sigma^2\varkappa}{\mu N_0 \Delta_0}\right\}$. Denote $\hat{N} \stackrel{\text{def}}{=} \left(\frac{4c_1 L\varkappa}{\mu}\right)^{\frac{p-1}{p}}$ Then

$$\frac{c_1 L \varkappa \Delta_0}{\mu N_0^p} = \frac{\Delta_0}{4}, \quad \frac{c_2 \sigma \sqrt{\frac{\varkappa}{\mu}\Delta_0}}{\sqrt{m_0 N_0}} \leq \frac{\Delta_0}{4},$$

SIGM: strongly convex case

Denote $\Delta_k = \varphi(y_{N_k}) - \varphi^*$. Let us perform N_0 iterations with $R^2 = \frac{\varkappa}{\mu}\Delta_0 \stackrel{\text{def}}{=} R_0^2$ and $m = m_0$. Then

$$\mathbb{E}\Delta_1 \leq \frac{c_1 L \varkappa \Delta_0}{\mu N_0^p} + \frac{c_2 \sigma \sqrt{\frac{\varkappa}{\mu}\Delta_0}}{\sqrt{m_0 N_0}} + c_3 N_0^{p-1} \delta$$

Let us choose $N_0 = \left(\frac{4 c_1 L \varkappa}{\mu}\right)^{\frac{1}{p}}$ and $m_0 \geq \max\left\{1, \frac{16 c_2^2 \sigma^2 \varkappa}{\mu N_0 \Delta_0}\right\}$. Denote $\hat{N} \stackrel{\text{def}}{=} \left(\frac{4 c_1 L \varkappa}{\mu}\right)^{\frac{p-1}{p}}$ Then

$$\frac{c_1 L \varkappa \Delta_0}{\mu N_0^p} = \frac{\Delta_0}{4}, \quad \frac{c_2 \sigma \sqrt{\frac{\varkappa}{\mu}\Delta_0}}{\sqrt{m_0 N_0}} \leq \frac{\Delta_0}{4},$$

and

$$\mathbb{E}\Delta_1 \leq \frac{\Delta_0}{2} + c_3 \hat{N} \delta.$$

SIGM: strongly convex case

Also we have

$$\mathbb{E}V(y_{N_0}, x^*) \leq \frac{\varkappa}{\mu}\mathbb{E}\Delta_1 \leq \frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2} + c_3\hat{N}\delta\right).$$

SIGM: strongly convex case

Also we have

$$\mathbb{E}V(y_{N_0}, x^*) \leq \frac{\varkappa}{\mu}\mathbb{E}\Delta_1 \leq \frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2} + c_3\hat{N}\delta\right).$$

We make N_1 iterations with $R^2 = \frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2} + c_3\hat{N}\delta\right) \overset{\text{def}}{=} R_1^2$ and $m = m_1$.

Also we have

$$\mathbb{E}V(y_{N_0}, x^*) \leq \frac{\varkappa}{\mu}\mathbb{E}\Delta_1 \leq \frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2} + c_3\hat{N}\delta\right).$$

We make N_1 iterations with $R^2 = \frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2} + c_3\hat{N}\delta\right) \stackrel{\text{def}}{=} R_1^2$ and $m = m_1$.
Then

$$\mathbb{E}\Delta_2 \leq \frac{c_1 L\varkappa}{\mu N_1^p}\left(\frac{\Delta_0}{2} + c_3\hat{N}\delta\right) + \frac{c_2\sigma\sqrt{\frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2} + c_3\hat{N}\delta\right)}}{\sqrt{m_1 N_1}} + c_3 N_1^{p-1}\delta$$

SIGM: strongly convex case

Also we have

$$\mathbb{E}V(y_{N_0}, x^*) \le \frac{\varkappa}{\mu}\mathbb{E}\Delta_1 \le \frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2} + c_3\hat{N}\delta\right).$$

We make N_1 iterations with $R^2 = \frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2} + c_3\hat{N}\delta\right) \stackrel{\text{def}}{=} R_1^2$ and $m = m_1$.
Then

$$\mathbb{E}\Delta_2 \le \frac{c_1 L\varkappa}{\mu N_1^p}\left(\frac{\Delta_0}{2} + c_3\hat{N}\delta\right) + \frac{c_2\sigma\sqrt{\frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2} + c_3\hat{N}\delta\right)}}{\sqrt{m_1 N_1}} + c_3 N_1^{p-1}\delta$$

Let us choose $N_1 = N_0$ and $m_1 \ge \max\left\{1, \frac{16c_2^2\sigma^2\varkappa}{\mu N_0\left(\frac{\Delta_0}{2} + c_3\hat{N}\delta\right)}\right\}$.

SIGM: strongly convex case

Also we have

$$\mathbb{E}V(y_{N_0},x^*) \le \frac{\varkappa}{\mu}\mathbb{E}\Delta_1 \le \frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2}+c_3\hat{N}\delta\right).$$

We make N_1 iterations with $R^2 = \frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2}+c_3\hat{N}\delta\right) \stackrel{\text{def}}{=} R_1^2$ and $m = m_1$.
Then

$$\mathbb{E}\Delta_2 \le \frac{c_1 L\varkappa}{\mu N_1^p}\left(\frac{\Delta_0}{2}+c_3\hat{N}\delta\right) + \frac{c_2\sigma\sqrt{\frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2}+c_3\hat{N}\delta\right)}}{\sqrt{m_1 N_1}} + c_3 N_1^{p-1}\delta$$

Let us choose $N_1 = N_0$ and $m_1 \ge \max\left\{1, \frac{16c_2^2\sigma^2\varkappa}{\mu N_0\left(\frac{\Delta_0}{2}+c_3\hat{N}\delta\right)}\right\}$. Then

$$\frac{c_1 L\varkappa}{\mu N_1^p}\left(\frac{\Delta_0}{2}+c_3\hat{N}\delta\right) = \frac{1}{4}\left(\frac{\Delta_0}{2}+c_3\hat{N}\delta\right), \quad \frac{c_2\sigma\sqrt{\frac{\varkappa}{\mu}\left(\frac{\Delta_0}{2}+c_3\hat{N}\delta\right)}}{\sqrt{m_1 N_1}} \le \frac{1}{4}\left(\frac{\Delta_0}{2}+c_3\hat{N}\delta\right),$$

and

$$\mathbb{E}\Delta_2 \le \frac{1}{2}\left(\frac{\Delta_0}{2}+c_3\hat{N}\delta\right) + c_3\hat{N}\delta.$$

SIGM: strongly convex case

On the k-th iteration we have

$$\mathbb{E}\Delta_k \leq \frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right).$$

SIGM: strongly convex case

On the k-th iteration we have

$$\mathbb{E}\Delta_k \leq \frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right).$$

We continue choosing on the k-th iteration

$$N_k = N_0, \quad m_k = \max\left\{1, \left\lceil \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0 \Delta_0} \right\rceil\right\} \geq \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right)},$$

SIGM: strongly convex case

On the k-th iteration we have

$$\mathbb{E}\Delta_k \leq \frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right).$$

We continue choosing on the k-th iteration

$$N_k = N_0, \quad m_k = \max\left\{1, \left\lceil \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0 \Delta_0} \right\rceil\right\} \geq \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right)},$$

and obtain

$$\mathbb{E}\Delta_{k+1} \leq \frac{1}{2}\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right) + c_3\hat{N}\delta = \frac{\Delta_0}{2^{k+1}} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^{k+1}}\right).$$

SIGM: strongly convex case

On the k-th iteration we have

$$\mathbb{E}\Delta_k \le \frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right).$$

We continue choosing on the k-th iteration

$$N_k = N_0, \quad m_k = \max\left\{1, \left\lceil \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0 \Delta_0} \right\rceil\right\} \ge \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right)},$$

and obtain

$$\mathbb{E}\Delta_{k+1} \le \frac{1}{2}\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right) + c_3\hat{N}\delta = \frac{\Delta_0}{2^{k+1}} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^{k+1}}\right).$$

If we perform $N = \log_2 \frac{\Delta_0}{\varepsilon} = \log_2 \frac{\mu R_0^2}{\varkappa\varepsilon}$ outer iterations we will obtain the error

$$\mathbb{E}\Delta_N \le \varepsilon + \tilde{c}_3\left(\frac{L}{\mu}\right)^{\frac{p-1}{p}}\delta.$$

SIGM: strongly convex case

On the k-th iteration we have

$$\mathbb{E}\Delta_k \le \frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right).$$

We continue choosing on the k-th iteration

$$N_k = N_0, \quad m_k = \max\left\{1, \left\lceil \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0 \Delta_0} \right\rceil\right\} \ge \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right)},$$

and obtain

$$\mathbb{E}\Delta_{k+1} \le \frac{1}{2}\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right) + c_3\hat{N}\delta = \frac{\Delta_0}{2^{k+1}} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^{k+1}}\right).$$

If we perform $N = \log_2\frac{\Delta_0}{\varepsilon} = \log_2\frac{\mu R_0^2}{\varkappa\varepsilon}$ outer iterations we will obtain the error

$$\mathbb{E}\Delta_N \le \varepsilon + \tilde{c}_3\left(\frac{L}{\mu}\right)^{\frac{p-1}{p}}\delta.$$

And the number of oracle calls will be

$$N(\varepsilon) = \sum_{i=0}^{N} N_i m_i$$

SIGM: strongly convex case

On the k-th iteration we have

$$\mathbb{E}\Delta_k \leq \frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right).$$

We continue choosing on the k-th iteration

$$N_k = N_0, \quad m_k = \max\left\{1, \left\lceil \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0 \Delta_0} \right\rceil\right\} \geq \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right)},$$

and obtain

$$\mathbb{E}\Delta_{k+1} \leq \frac{1}{2}\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right) + c_3\hat{N}\delta = \frac{\Delta_0}{2^{k+1}} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^{k+1}}\right).$$

If we perform $N = \log_2 \frac{\Delta_0}{\varepsilon} = \log_2 \frac{\mu R_0^2}{\varkappa\varepsilon}$ outer iterations we will obtain the error

$$\mathbb{E}\Delta_N \leq \varepsilon + \tilde{c}_3\left(\frac{L}{\mu}\right)^{\frac{p-1}{p}}\delta.$$

And the number of oracle calls will be

$$N(\varepsilon) = \sum_{i=0}^{N} N_i m_i \leq N_0 \sum_{i=0}^{N}\left(1 + \frac{16c_2^2\sigma^2\varkappa 2^i}{\mu N_0 \Delta_0}\right)$$

SIGM: strongly convex case

On the k-th iteration we have

$$\mathbb{E}\Delta_k \leq \frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right).$$

We continue choosing on the k-th iteration

$$N_k = N_0, \quad m_k = \max\left\{1, \left\lceil \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0 \Delta_0} \right\rceil\right\} \geq \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right)},$$

and obtain

$$\mathbb{E}\Delta_{k+1} \leq \frac{1}{2}\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right) + c_3\hat{N}\delta = \frac{\Delta_0}{2^{k+1}} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^{k+1}}\right).$$

If we perform $N = \log_2 \frac{\Delta_0}{\varepsilon} = \log_2 \frac{\mu R_0^2}{\varkappa\varepsilon}$ outer iterations we will obtain the error

$$\mathbb{E}\Delta_N \leq \varepsilon + \tilde{c}_3\left(\frac{L}{\mu}\right)^{\frac{p-1}{p}}\delta.$$

And the number of oracle calls will be

$$N(\varepsilon) = \sum_{i=0}^{N} N_i m_i \leq N_0 \sum_{i=0}^{N}\left(1 + \frac{16c_2^2\sigma^2\varkappa 2^i}{\mu N_0\Delta_0}\right) \leq N_0 N + \frac{32c_2^2\sigma^2\varkappa 2^N}{\mu\Delta_0} =$$

SIGM: strongly convex case

On the k-th iteration we have

$$\mathbb{E}\Delta_k \leq \frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right).$$

We continue choosing on the k-th iteration

$$N_k = N_0, \quad m_k = \max\left\{1, \left\lceil \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0 \Delta_0} \right\rceil\right\} \geq \frac{16c_2^2\sigma^2\varkappa 2^k}{\mu N_0\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right)},$$

and obtain

$$\mathbb{E}\Delta_{k+1} \leq \frac{1}{2}\left(\frac{\Delta_0}{2^k} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^k}\right)\right) + c_3\hat{N}\delta = \frac{\Delta_0}{2^{k+1}} + 2c_3\hat{N}\delta\left(1 - \frac{1}{2^{k+1}}\right).$$

If we perform $N = \log_2 \frac{\Delta_0}{\varepsilon} = \log_2 \frac{\mu R_0^2}{\varkappa\varepsilon}$ outer iterations we will obtain the error

$$\mathbb{E}\Delta_N \leq \varepsilon + \tilde{c}_3\left(\frac{L}{\mu}\right)^{\frac{p-1}{p}}\delta.$$

And the number of oracle calls will be

$$N(\varepsilon) = \sum_{i=0}^{N} N_i m_i \leq N_0 \sum_{i=0}^{N}\left(1 + \frac{16c_2^2\sigma^2\varkappa 2^i}{\mu N_0\Delta_0}\right) \leq N_0 N + \frac{32c_2^2\sigma^2\varkappa 2^N}{\mu\Delta_0} = \left(\frac{4c_1 L\varkappa}{\mu}\right)^{\frac{1}{p}}\log_2\frac{\mu R_0^2}{\varkappa\varepsilon} + \frac{32c_2^2\sigma^2\varkappa}{\mu\varepsilon}.$$

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.

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.

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.

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.

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.

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.

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.

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.

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.

Conclusion

1. We've considered three gradient methods (PGM, DGM, FGM) and how they perform in the presence of stochastic and deterministic errors.

Conclusion

1. We've considered three gradient methods (PGM, DGM, FGM) and how they perform in the presence of stochastic and deterministic errors.
2. We've considered new SIGM with intermediate rate of convergence in stochastic case.

Conclusion

1. We've considered three gradient methods (PGM, DGM, FGM) and how they perform in the presence of stochastic and deterministic errors.
2. We've considered new SIGM with intermediate rate of convergence in stochastic case.
3. We have seen the following tools which give us the flexibility for solving optimization problems and constructing new methods:
    1. Choice of the norm and prox-function

Conclusion

1. We've considered three gradient methods (PGM, DGM, FGM) and how they perform in the presence of stochastic and deterministic errors.
2. We've considered new SIGM with intermediate rate of convergence in stochastic case.
3. We have seen the following tools which give us the flexibility for solving optimization problems and constructing new methods:
    1. Choice of the norm and prox-function
    2. Composite optimization

Conclusion

1. We've considered three gradient methods (PGM, DGM, FGM) and how they perform in the presence of stochastic and deterministic errors.
2. We've considered new SIGM with intermediate rate of convergence in stochastic case.
3. We have seen the following tools which give us the flexibility for solving optimization problems and constructing new methods:
    1. Choice of the norm and prox-function
    2. Composite optimization
    3. Auxiliary deterministic inexactness

Conclusion

1. We've considered three gradient methods (PGM, DGM, FGM) and how they perform in the presence of stochastic and deterministic errors.
2. We've considered new SIGM with intermediate rate of convergence in stochastic case.
3. We have seen the following tools which give us the flexibility for solving optimization problems and constructing new methods:
    1. Choice of the norm and prox-function
    2. Composite optimization
    3. Auxiliary deterministic inexactness
    4. Artificial randomization

Conclusion

1. We've considered three gradient methods (PGM, DGM, FGM) and how they perform in the presence of stochastic and deterministic errors.
2. We've considered new SIGM with intermediate rate of convergence in stochastic case.
3. We have seen the following tools which give us the flexibility for solving optimization problems and constructing new methods:
    1. Choice of the norm and prox-function
    2. Composite optimization
    3. Auxiliary deterministic inexactness
    4. Artificial randomization
    5. Trade-off between rate of convergence and error accumulation error

Conclusion

1. We've considered three gradient methods (PGM, DGM, FGM) and how they perform in the presence of stochastic and deterministic errors.
2. We've considered new SIGM with intermediate rate of convergence in stochastic case.
3. We have seen the following tools which give us the flexibility for solving optimization problems and constructing new methods:
    1. Choice of the norm and prox-function
    2. Composite optimization
    3. Auxiliary deterministic inexactness
    4. Artificial randomization
    5. Trade-off between rate of convergence and error accumulation error
    6. Restart technique

Conclusion

1. We've considered three gradient methods (PGM, DGM, FGM) and how they perform in the presence of stochastic and deterministic errors.
2. We've considered new SIGM with intermediate rate of convergence in stochastic case.
3. We have seen the following tools which give us the flexibility for solving optimization problems and constructing new methods:
    1. Choice of the norm and prox-function
    2. Composite optimization
    3. Auxiliary deterministic inexactness
    4. Artificial randomization
    5. Trade-off between rate of convergence and error accumulation error
    6. Restart technique
    7. Amplification

Conclusion

1. We've considered three gradient methods (PGM, DGM, FGM) and how they perform in the presence of stochastic and deterministic errors.
2. We've considered new SIGM with intermediate rate of convergence in stochastic case.
3. We have seen the following tools which give us the flexibility for solving optimization problems and constructing new methods:
    1. Choice of the norm and prox-function
    2. Composite optimization
    3. Auxiliary deterministic inexactness
    4. Artificial randomization
    5. Trade-off between rate of convergence and error accumulation error
    6. Restart technique
    7. Amplification
    8. Using the problem structure

References

- O. Devolder, F. Glineur and Yu. Nesterov. *Intermediate Gradient Methods for Smooth Convex Problems with Inexact Oracle*. CORE Discussion Paper 2013/17, available at http://www.uclouvain.be/cps/ucl/doc/core/documents/coredp2013_17web.pdf.
- O. Devolder, F. Glineur and Yu. Nesterov. *First-order Methods of Smooth Convex Optimization with Inexact Oracle*. CORE Discussion Paper 2011/2, available at http://www.optimization-online.org/DB_FILE/2010/12/2865.pdf.
- A. Juditsky, G. Lan, A. Nemirovski, A. Shapiro *Stochastic approximation approach to stochastic programming*. SIAM Journal on Optimization. 2009, 19(4), pp. 1574–1609.
- G. Lan, A. Nemirovski and A. Shapiro. *Validation analysis of mirror descent stochastic approximation method.* Mathematical Programming Serie A, 2012, 134(2), pp. 425–458.
- O. Devolder. *Exactness, Inexactness and Stochasticity in First-Order Methods for Large-Scale Convex Optimization*, PhD thesis (2013).
- S.Ghadimi, G.Lan, *Optimal Stochastic Approximation Algorithms for Strongly Convex Stochastic Composite Optimization I: A Generic Algorithmic Framework*, SIAM J. Optim., 2012, 22(4), pp. 1469–1492.
- S.Ghadimi, G.Lan, *Optimal Stochastic Approximation Algorithms for Strongly Convex Stochastic Composite Optimization II: Shrinking Procedures and Optimal Algorithms*, SIAM J. Optim., 2013, 23(4), pp. 2061–2089.

Thank you for your attention!