Convergent Subgradient Methods for Nonsmooth Convex Optimization

Yurii Nesterov, CORE/INMA (UCL)

February 4, 2014 (Seminar PremoLab, Moscow)

Joint work with V.Shikhman (CORE)

Outline

History of Developments

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min\limits_{x \in Q} f(x),$

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min_{x \in Q} f(x)$, where

Q is a closed convex set,

History of Developments

Problem: $f_* \stackrel{\text{def}}{=} \min\limits_{x \in Q} f(x)$, where

Q is a closed convex set, f is a closed convex function.

First <u>Primal</u> Subgradient Method (N.Shor, B.Polyak (60's)):

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min\limits_{x \in Q} f(x)$, where

Q is a closed convex set, f is a closed convex function.

First Primal Subgradient Method (N.Shor, B.Polyak (60's)):

$$x_{t+1} = \pi_Q \left(x_t - a_t \nabla f(x_t) \right), \quad t \geq 0,$$

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min\limits_{x \in Q} f(x)$, where

Q is a closed convex set, f is a closed convex function.

First <u>Primal</u> Subgradient Method (N.Shor, B.Polyak (60's)):

$$x_{t+1} = \pi_Q \left(x_t - a_t \nabla f(x_t) \right), \quad t \geq 0, \quad \text{where}$$

- $\pi_Q(x)$ is a Eucleaden projection of x onto Q,

History of Developments

Problem: $f_* \stackrel{\text{def}}{=} \min_{x \in Q} f(x),$ where

Q is a closed convex set, f is a closed convex function.

First <u>Primal</u> Subgradient Method (N.Shor, B.Polyak (60's)):

$$x_{t+1} = \pi_Q \left(x_t - a_t \nabla f(x_t) \right), \quad t \geq 0, \quad \text{where}$$

- $\pi_Q(x)$ is a Eucleaden projection of x onto Q,
- $\nabla f(x_k)$ is arbitrary subgradient of f at x_k,

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min_{x \in Q} f(x)$, where

Q is a closed convex set, f is a closed convex function.

First <u>Primal</u> Subgradient Method (N.Shor, B.Polyak (60's)):

$$x_{t+1} = \pi_Q \left(x_t - a_t \nabla f(x_t) \right), \quad t \geq 0, \quad \text{where}$$

- $\pi_Q(x)$ is a Eucleaden projection of x onto Q,
- $\nabla f(x_k)$ is arbitrary subgradient of f at x_k,
- $a_k > 0$ is a step size parameter.

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min_{x \in Q} f(x)$, where

Q is a closed convex set, f is a closed convex function.

First <u>Primal</u> Subgradient Method (N.Shor, B.Polyak (60's)):

$$x_{t+1} = \pi_Q\left(x_t - a_t \nabla f(x_t)\right), \quad t \geq 0, \quad \text{where}$$

- $\pi_Q(x)$ is a Eucleaden projection of x onto Q,
- $\nabla f(x_k)$ is arbitrary subgradient of f at x_k,
- $a_k > 0$ is a step size parameter.

NB: Euclidean framework is essential!

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min_{x \in Q} f(x)$, where

Q is a closed convex set, f is a closed convex function.

First Primal Subgradient Method (N.Shor, B.Polyak (60's)):

$$x_{t+1} = \pi_Q \left(x_t - a_t \nabla f(x_t) \right), \quad t \geq 0, \quad \text{where}$$

- $\pi_Q(x)$ is a Eucleaden projection of x onto Q,
- $\nabla f(x_k)$ is arbitrary subgradient of f at x_k,
- $a_k > 0$ is a step size parameter.

NB: Euclidean framework is essential! (No monotonicity in f)

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min\limits_{x \in Q} f(x)$, where

Q is a closed convex set, f is a closed convex function.

First Primal Subgradient Method (N.Shor, B.Polyak (60's)):

$$x_{t+1} = \pi_Q\left(x_t - a_t \nabla f(x_t)\right), \quad t \geq 0, \quad \text{where}$$

- $\pi_Q(x)$ is a Eucleaden projection of x onto Q,
- $\nabla f(x_k)$ is arbitrary subgradient of f at x_k,
- $a_k > 0$ is a step size parameter.

NB: Euclidean framework is essential! (No monotonicity in f)

Convergence:

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min_{x \in Q} f(x)$, where

Q is a closed convex set, f is a closed convex function.

First Primal Subgradient Method (N.Shor, B.Polyak (60's)):

$$x_{t+1} = \pi_Q\left(x_t - a_t \nabla f(x_t)\right), \quad t \geq 0, \quad \text{where}$$

- $\pi_Q(x)$ is a Eucleaden projection of x onto Q,
- $\nabla f(x_k)$ is arbitrary subgradient of f at x_k,
- $a_k > 0$ is a step size parameter.

NB: Euclidean framework is essential! (No monotonicity in f)

Convergence: Denote $A_t = \sum_{k=0}^{t} a_k$.

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min\limits_{x \in Q} f(x)$, where

Q is a closed convex set, f is a closed convex function.

First <u>Primal</u> Subgradient Method (N.Shor, B.Polyak (60's)):

$$x_{t+1} = \pi_Q\left(x_t - a_t \nabla f(x_t)\right), \quad t \geq 0, \quad \text{where}$$

- $\pi_Q(x)$ is a Eucleaden projection of x onto Q,
- $\nabla f(x_k)$ is arbitrary subgradient of f at x_k,
- $a_k > 0$ is a step size parameter.

NB: Euclidean framework is essential! (No monotonicity in f)

Convergence: Denote $A_t = \sum\limits_{k=0}^{t} a_k$. Then

$$\frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \leq \frac{1}{A_t}\left[\frac{1}{2}\|x_0 - x_*\|_2^2 + \frac{1}{2}\sum_{k=0}^{t} a_k^2 \|\nabla f(x_k)\|_2^2\right],$$

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min\limits_{x \in Q} f(x)$, where

Q is a closed convex set, f is a closed convex function.

First Primal Subgradient Method (N.Shor, B.Polyak (60's)):

$$x_{t+1} = \pi_Q \left(x_t - a_t \nabla f(x_t) \right), \quad t \geq 0, \quad \text{where}$$

- $\pi_Q(x)$ is a Eucleaden projection of x onto Q,
- $\nabla f(x_k)$ is arbitrary subgradient of f at x_k,
- $a_k > 0$ is a step size parameter.

NB: Euclidean framework is essential! (No monotonicity in f)

Convergence: Denote $A_t = \sum\limits_{k=0}^{t} a_k$. Then

$$\frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \leq \frac{1}{A_t} \left[\frac{1}{2} \| x_0 - x_* \|_2^2 + \frac{1}{2} \sum_{k=0}^{t} a_k^2 \| \nabla f(x_k) \|_2^2 \right],$$

Conditions: $a_t \to 0$,

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min\limits_{x \in Q} f(x)$, where

Q is a closed convex set, f is a closed convex function.

First Primal Subgradient Method (N.Shor, B.Polyak (60's)):

$$x_{t+1} = \pi_Q\left(x_t - a_t \nabla f(x_t)\right), \quad t \geq 0, \quad \text{where}$$

- $\pi_Q(x)$ is a Eucleaden projection of x onto Q,
- $\nabla f(x_k)$ is arbitrary subgradient of f at x_k,
- $a_k > 0$ is a step size parameter.

NB: Euclidean framework is essential! (No monotonicity in f)

Convergence: Denote $A_t = \sum\limits_{k=0}^{t} a_k$. Then

$$\frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \leq \frac{1}{A_t}\left[\frac{1}{2}\|x_0 - x_*\|_2^2 + \frac{1}{2}\sum_{k=0}^{t} a_k^2 \|\nabla f(x_k)\|_2^2\right],$$

Conditions: $a_t \to 0$, $A_t \to \infty$.

History of Developments

Problem: $f_* \stackrel{\mathrm{def}}{=} \min_{x \in Q} f(x)$, where

Q is a closed convex set, f is a closed convex function.

First <u>Primal</u> Subgradient Method (N.Shor, B.Polyak (60's)):

$$x_{t+1} = \pi_Q \left(x_t - a_t \nabla f(x_t) \right), \quad t \geq 0, \quad \text{where}$$

- $\pi_Q(x)$ is a Eucleaden projection of x onto Q,
- $\nabla f(x_k)$ is arbitrary subgradient of f at x_k,
- $a_k > 0$ is a step size parameter.

NB: Euclidean framework is essential! (No monotonicity in f)

Convergence: Denote $A_t = \sum_{k=0}^{t} a_k$. Then

$$\frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \leq \frac{1}{A_t} \left[\frac{1}{2} \|x_0 - x_*\|_2^2 + \frac{1}{2} \sum_{k=0}^{t} a_k^2 \|\nabla f(x_k)\|_2^2 \right],$$

Conditions: $a_t \to 0$, $A_t \to \infty$. **Optimal:** $a_t = \frac{R}{L\sqrt{t+1}} \Rightarrow O(\frac{L^2 R^2}{\epsilon^2})$.

First Dual Methods

First Dual Methods

1. Mirror Descent Method (Nemirovskii, Yudin, (70's))

First Dual Methods

1. Mirror Descent Method (Nemirovskii, Yudin, (70's))

$$x_{t+1} = \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + d(x) \right\}, \quad t \geq 0,$$

First Dual Methods

1. Mirror Descent Method (Nemirovskii, Yudin, (70's))

$$x_{t+1} = \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + d(x) \right\}, \quad t \geq 0,$$

where $d(x) \geq 0$ is a *prox-function* of Q:

First Dual Methods

1. Mirror Descent Method (Nemirovskii, Yudin, (70's))

$$x_{t+1} = \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + d(x) \right\}, \quad t \geq 0,$$

where $d(x) \geq 0$ is a *prox-function* of Q:

- $d(x_0) = 0$ for some $x_0 \in Q$,

First Dual Methods

1. Mirror Descent Method (Nemirovskii, Yudin, (70's))

$$x_{t+1} = \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + d(x) \right\}, \quad t \geq 0,$$

where $d(x) \geq 0$ is a *prox-function* of Q:

- $d(x_0) = 0$ for some $x_0 \in Q$,
- It is strongly convex:
  $d(y) \geq d(x) + \langle \nabla d(x), y - x \rangle + \frac{1}{2}\|y - x\|^2$, $x, y \in Q$.

First Dual Methods

1. Mirror Descent Method (Nemirovskii, Yudin, (70's))

$$x_{t+1} = \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + d(x) \right\}, \quad t \geq 0,$$

where $d(x) \geq 0$ is a *prox-function* of Q:

- $d(x_0) = 0$ for some $x_0 \in Q$,
- It is strongly convex:
  $d(y) \geq d(x) + \langle \nabla d(x), y - x \rangle + \frac{1}{2}\|y - x\|^2$, $x, y \in Q$.

The norm is arbitrary now!

First Dual Methods

1. Mirror Descent Method (Nemirovskii, Yudin, (70's))

$$x_{t+1} = \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + d(x) \right\}, \quad t \geq 0,$$

where $d(x) \geq 0$ is a *prox-function* of Q:

- $d(x_0) = 0$ for some $x_0 \in Q$,
- It is strongly convex:
  $d(y) \geq d(x) + \langle \nabla d(x), y - x \rangle + \frac{1}{2}\|y - x\|^2$, $x, y \in Q$.

The norm is arbitrary now!

Convergence:

$$\frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \leq \frac{1}{A_t} \left[d(x^*) + \frac{1}{2} \sum_{k=0}^{t} a_k^2 \|\nabla f(x_k)\|_*^2 \right],$$

First Dual Methods

1. Mirror Descent Method (Nemirovskii, Yudin, (70's))

$$x_{t+1} = \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + d(x) \right\}, \quad t \geq 0,$$

where $d(x) \geq 0$ is a *prox-function* of Q:

- $d(x_0) = 0$ for some $x_0 \in Q$,
- It is strongly convex:
  $d(y) \geq d(x) + \langle \nabla d(x), y - x \rangle + \frac{1}{2}\|y - x\|^2$, $x, y \in Q$.

The norm is arbitrary now!

Convergence:

$$\frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \leq \frac{1}{A_t} \left[d(x^*) + \frac{1}{2} \sum_{k=0}^{t} a_k^2 \|\nabla f(x_k)\|_*^2 \right],$$

where $\|s\|_* = \max_{x \in E} \{\langle s, x \rangle : \ \|x\| \leq 1\}$, $s \in E^*$.

Dual averaging (N.2003/2005/2009)

Dual averaging (N.2003/2005/2009)

Small inconsistency:

Dual averaging (N.2003/2005/2009)

Small inconsistency:

In the main objects, $\sum_{k=0}^{t} a_k f(x_k)$ and $\sum_{k=0}^{t} a_k \nabla f(x_k)$,

Dual averaging (N.2003/2005/2009)

Small inconsistency:

In the main objects, $\sum_{k=0}^{t} a_k f(x_k)$ and $\sum_{k=0}^{t} a_k \nabla f(x_k)$, new information enter with *decreasing* weights.

Dual averaging (N.2003/2005/2009)

Small inconsistency:

In the main objects, $\sum\limits_{k=0}^{t} a_k f(x_k)$ and $\sum\limits_{k=0}^{t} a_k \nabla f(x_k)$,
new information enter with *decreasing* weights.

Method:

$$x_{t+1} = \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}, \quad t \geq 0,$$

Dual averaging (N.2003/2005/2009)

Small inconsistency:

In the main objects, $\sum\limits_{k=0}^{t} a_k f(x_k)$ and $\sum\limits_{k=0}^{t} a_k \nabla f(x_k)$,

new information enter with <u>*decreasing*</u> weights.

Method:

$$x_{t+1} = \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}, \quad t \geq 0,$$

where the *scaling coefficients* $\{\gamma_t\}_{t \geq 0}$ are positive.

Dual averaging (N.2003/2005/2009)

Small inconsistency:

In the main objects, $\sum\limits_{k=0}^{t} a_k f(x_k)$ and $\sum\limits_{k=0}^{t} a_k \nabla f(x_k)$,

new information enter with *decreasing* weights.

Method:

$$x_{t+1} = \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}, \quad t \geq 0,$$

where the *scaling coefficients* $\{\gamma_t\}_{t \geq 0}$ are positive.

Convergence:

$$\frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \leq \frac{1}{A_t} \left[\gamma_t d(x^*) + \sum_{k=0}^{t} \frac{a_k^2}{2\gamma_k} \| \nabla f(x_k) \|_*^2 \right].$$

Dual averaging (N.2003/2005/2009)

Small inconsistency:

In the main objects, $\sum\limits_{k=0}^{t} a_k f(x_k)$ and $\sum\limits_{k=0}^{t} a_k \nabla f(x_k)$,

new information enter with <u>*decreasing*</u> weights.

Method:

$$x_{t+1} = \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}, \quad t \geq 0,$$

where the *scaling coefficients* $\{\gamma_t\}_{t \geq 0}$ are positive.

Convergence:

$$\frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \leq \frac{1}{A_t} \left[\gamma_t d(x^*) + \sum_{k=0}^{t} \frac{a_k^2}{2\gamma_k} \| \nabla f(x_k) \|_*^2 \right].$$

Important case: $a_t \equiv 1$, $\gamma_t = \frac{L}{R}\sqrt{t+1}$.

Dual averaging (N.2003/2005/2009)

Small inconsistency:

In the main objects, $\sum\limits_{k=0}^{t} a_k f(x_k)$ and $\sum\limits_{k=0}^{t} a_k \nabla f(x_k)$,
new information enter with *decreasing* weights.

Method:

$$x_{t+1} = \min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} a_k \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}, \quad t \geq 0,$$

where the *scaling coefficients* $\{\gamma_t\}_{t \geq 0}$ are positive.

Convergence:

$$\frac{1}{A_t} \sum_{k=0}^{t} a_k f(x_k) - f_* \leq \frac{1}{A_t} \left[\gamma_t d(x^*) + \sum_{k=0}^{t} \frac{a_k^2}{2\gamma_k} \|\nabla f(x_k)\|_*^2 \right].$$

Important case: $a_t \equiv 1$, $\gamma_t = \frac{L}{R}\sqrt{t+1}$.

Then we get $O\left(\frac{L^2 R^2}{\epsilon^2}\right)$ complexity.

Common drawback

Common drawback

All methods cannot generate a <u>*convergent*</u> *sequence of test points.*

Common drawback

All methods cannot generate a <u>convergent</u> sequence of test points.

Indeed, we can guarantee only $\lim\limits_{t\to\infty} \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k f(x_k) = f_*$.

Common drawback

All methods cannot generate a <u>*convergent*</u> *sequence of test points.*

Indeed, we can guarantee only $\lim\limits_{t\to\infty} \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k f(x_k) = f_*$.

(Allows uncontrollable jumps of objective function.)

Common drawback

All methods cannot generate a convergent sequence of test points.

Indeed, we can guarantee only $\lim\limits_{t\to\infty} \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k f(x_k) = f_*$.

(Allows uncontrollable jumps of objective function.)

Possible treatments:

Common drawback

All methods cannot generate a convergent sequence of test points.

Indeed, we can guarantee only $\lim\limits_{t\to\infty} \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k f(x_k) = f_*$.

(Allows uncontrollable jumps of objective function.)

Possible treatments:

1. Consider the sequence of the record values $f_t^* = \min\limits_{0 \le k \le t} f(x_k)$.

Common drawback

All methods cannot generate a convergent sequence of test points.

Indeed, we can guarantee only $\lim\limits_{t\to\infty} \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k f(x_k) = f_*$.

(Allows uncontrollable jumps of objective function.)

Possible treatments:

1. Consider the sequence of the record values $f_t^* = \min\limits_{0\le k\le t} f(x_k)$.

But: we need to know values $\{f(x_k)\}$

Common drawback

All methods cannot generate a convergent sequence of test points.

Indeed, we can guarantee only $\lim\limits_{t\to\infty} \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k f(x_k) = f_*$.

(Allows uncontrollable jumps of objective function.)

Possible treatments:

1. Consider the sequence of the record values $f_t^* = \min\limits_{0\leq k\leq t} f(x_k)$.

But: we need to know values $\{f(x_k)\}$ (Not always possible!)

Common drawback

All methods cannot generate a <u>*convergent*</u> *sequence of test points.*

Indeed, we can guarantee only $\lim\limits_{t\to\infty} \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k f(x_k) = f_*$.

(Allows uncontrollable jumps of objective function.)

Possible treatments:

1. Consider the sequence of the record values $f_t^* = \min\limits_{0\le k\le t} f(x_k)$.

But: we need to know values $\{f(x_k)\}$ (Not always possible!)

2. Consider the average points $\bar{x}_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k x_k$.

But: the convergent minimizing sequence does not participate in the minimization process.

Common drawback

All methods cannot generate a convergent sequence of test points.

Indeed, we can guarantee only $\lim\limits_{t\to\infty} \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k f(x_k) = f_*$.

(Allows uncontrollable jumps of objective function.)

Possible treatments:

1. Consider the sequence of the record values $f_t^* = \min\limits_{0\le k\le t} f(x_k)$.

But: we need to know values $\{f(x_k)\}$ (Not always possible!)

2. Consider the average points $\bar{x}_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k x_k$.

But: the convergent minimizing sequence does not participate in the minimization process. (Bad for some applications.)

Our goal: development of *convergent* subgradient methods.

How do we prove the rate of convergence?

How do we prove the rate of convergence?

1. Euclidean distance.

How do we prove the rate of convergence?

1. Euclidean distance. Use $\|x_k - x_*\|_2^2$ as a Lyapunov function of the primal process (PGM).

How do we prove the rate of convergence?

1. **Euclidean distance.** Use $\|x_k - x_*\|_2^2$ as a Lyapunov function of the primal process (PGM).
2. **Dual potential.** Define $V(s) = \max\limits_{x \in Q}[\langle s, x \rangle - d(x)$.

How do we prove the rate of convergence?

1. Euclidean distance. Use $\|x_k - x_*\|_2^2$ as a Lyapunov function of the primal process (PGM).

2. Dual potential. Define $V(s) = \max\limits_{x \in Q} [\langle s, x \rangle - d(x)$. Use

$$\Psi(s) = V(s) - \langle s, x_* \rangle$$

as Lyapunov function of the dual process (MDM).

3. Gap functions. Find the upper bounds for the growth of values

$$\max_{x \in Q} \left\{ \sum_{k=0}^{t} a_k \langle \nabla f(x_k), x - x_k \rangle - \gamma_t d(x) \right\}.$$

(Dual averaging.)

How do we prove the rate of convergence?

1. Euclidean distance. Use $\|x_k - x_*\|_2^2$ as a Lyapunov function of the primal process (PGM).

2. Dual potential. Define $V(s) = \max\limits_{x\in Q}[\langle s, x\rangle - d(x)$. Use

$$\Psi(s) = V(s) - \langle s, x_*\rangle$$

as Lyapunov function of the dual process (MDM).

3. Gap functions. Find the upper bounds for the growth of values

$$\max_{x\in Q}\left\{\sum_{k=0}^{t} a_k\langle\nabla f(x_k), x - x_k\rangle - \gamma_t d(x)\right\}.$$

(Dual averaging.)

4. Estimate sequences (Fast GM). Maintain condition

$$A_t f(x_t) \le \sum_{k=0}^{t} a_k[f(y_k) + \langle\nabla f(y_k), x - y_k\rangle] + d(x)$$

for all $x \in Q$

How do we prove the rate of convergence?

1. Euclidean distance. Use $\|x_k - x_*\|_2^2$ as a Lyapunov function of the primal process (PGM).

2. Dual potential. Define $V(s) = \max\limits_{x \in Q}[\langle s, x\rangle - d(x)$. Use

$$\Psi(s) = V(s) - \langle s, x_* \rangle$$

as Lyapunov function of the dual process (MDM).

3. Gap functions. Find the upper bounds for the growth of values

$$\max_{x \in Q} \left\{ \sum_{k=0}^{t} a_k \langle \nabla f(x_k), x - x_k \rangle - \gamma_t d(x) \right\}.$$

(Dual averaging.)

4. Estimate sequences (Fast GM). Maintain condition

$$A_t f(x_t) \le \sum_{k=0}^{t} a_k [f(y_k) + \langle \nabla f(y_k), x - y_k \rangle] + d(x)$$

for all $x \in Q$ (Smooth minimization.)

Relaxed estimate sequences

Relaxed estimate sequences

We are going to maintain the following condition:

Relaxed estimate sequences

We are going to maintain the following condition:

$$A_t f(x_t) \leq \sum_{k=0}^{t} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle] + \gamma_t d(x) + B_t,$$

for all $x \in Q$, where $B_t \geq 0$.

Relaxed estimate sequences

We are going to maintain the following condition:

$$A_t f(x_t) \le \sum_{k=0}^{t} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle] + \gamma_t d(x) + B_t,$$

for all $x \in Q$, where $B_t \ge 0$.

With notation $\ell_t(x) = \sum_{k=0}^{t} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle]$,

and $\psi_t^* = \min_{x \in Q} [\ell_t(x) + \gamma_t d(x)]$, this is

$$A_t f(x_t) \le \psi_t^* + B_t.$$

Relaxed estimate sequences

We are going to maintain the following condition:

$$A_t f(x_t) \leq \sum_{k=0}^{t} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle] + \gamma_t d(x) + B_t,$$

for all $x \in Q$, where $B_t \geq 0$.

With notation $\ell_t(x) = \sum_{k=0}^{t} a_k [f(x_k) + \langle \nabla f(x_k), x - x_k \rangle]$,
and $\psi_t^* = \min_{x \in Q} [\ell_t(x) + \gamma_t d(x)]$, this is

$$A_t f(x_t) \leq \psi_t^* + B_t.$$

NB: this condition includes only one sequence $\{x_t\}$.

Consequences

Consequences

Denote $s_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k \nabla f(x_k)$.

Consequences

Denote $s_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k \nabla f(x_k)$.

For arbitrary bounded closed convex set $C \subseteq Q$, denote

$$\xi_C(s) = \max_x \{ \langle s, x \rangle : \; x \in C \}, \; s \in E^*.$$

Consequences

Denote $s_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k \nabla f(x_k)$.

For arbitrary bounded closed convex set $C \subseteq Q$, denote

$$\xi_C(s) = \max_x \{ \langle s, x \rangle : \; x \in C \}, \; s \in E^*.$$

Lemma. For any $t \geq 0$ we have:

$$f(x_t) + f^*(s_t) + \xi_C(-s_t) \leq \frac{1}{A_t}(B_t + \gamma_t D_C),$$

where $D_C = \max\limits_x \{ d(x) : \; x \in C \bigcap Q \}$ and

$$f^*(s) = \max_{x \in E} [\langle s, x \rangle - f(x)].$$

Consequences

Denote $s_t = \frac{1}{A_t} \sum_{k=0}^{t} a_k \nabla f(x_k)$.

For arbitrary bounded closed convex set $C \subseteq Q$, denote

$$\xi_C(s) = \max_x \{\langle s, x \rangle : \ x \in C\}, \ s \in E^*.$$

Lemma. For any $t \geq 0$ we have:

$$f(x_t) + f^*(s_t) + \xi_C(-s_t) \leq \frac{1}{A_t}(B_t + \gamma_t D_C),$$

where $D_C = \max_x \{d(x) : \ x \in C \bigcap Q\}$ and

$$f^*(s) = \max_{x \in E} [\langle s, x \rangle - f(x)].$$

Example 1: $C = \{x_*\}$.

Consequences

Denote $s_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k \nabla f(x_k)$.

For arbitrary bounded closed convex set $C \subseteq Q$, denote

$$\xi_C(s) = \max_x \{ \langle s, x \rangle : \ x \in C \}, \ s \in E^*.$$

Lemma. For any $t \geq 0$ we have:

$$f(x_t) + f^*(s_t) + \xi_C(-s_t) \leq \frac{1}{A_t}(B_t + \gamma_t D_C),$$

where $D_C = \max\limits_x \{ d(x) : \ x \in C \bigcap Q \}$ and

$$f^*(s) = \max_{x \in E} [\langle s, x \rangle - f(x)].$$

Example 1: $C = \{x_*\}$. Then

$$\frac{1}{A_t}(B_t + \gamma_t d(x_*)) \geq f(x_t) + f^*(s_t) - \langle s_t, x_* \rangle$$

Consequences

Denote $s_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k \nabla f(x_k)$.

For arbitrary bounded closed convex set $C \subseteq Q$, denote

$$\xi_C(s) = \max_x \{ \langle s, x \rangle : \; x \in C \}, \; s \in E^*.$$

Lemma. For any $t \geq 0$ we have:

$$f(x_t) + f^*(s_t) + \xi_C(-s_t) \leq \frac{1}{A_t}(B_t + \gamma_t D_C),$$

where $D_C = \max\limits_x \{ d(x) : \; x \in C \bigcap Q \}$ and

$$f^*(s) = \max_{x \in E} [\langle s, x \rangle - f(x)].$$

Example 1: $C = \{x_*\}$. Then

$$\frac{1}{A_t}(B_t + \gamma_t d(x_*)) \geq f(x_t) + f^*(s_t) - \langle s_t, x_* \rangle \geq f(x_t) - f_*.$$

Example 2: separation

Example 2: separation

For arbitrary $R > 0$, denote

$$\|s\|_R^* = \max_{x \in Q} \{ \langle s, x_* - x \rangle : \; \|x - x_*\| \leq R \}, \quad s \in E^*.$$

Example 2: separation

For arbitrary $R > 0$, denote

$$\|s\|_R^* = \max_{x \in Q} \{ \langle s, x_* - x \rangle : \ \|x - x_*\| \leq R \}, \quad s \in E^*.$$

Note that $\|s\|_R^* \geq 0$ for any $s \in E^*$.

Example 2: separation

For arbitrary $R > 0$, denote

$$\|s\|_R^* = \max_{x \in Q} \{ \langle s, x_* - x \rangle : \ \|x - x_*\| \leq R \}, \quad s \in E^*.$$

Note that $\|s\|_R^* \geq 0$ for any $s \in E^*$.

In view of the first-order optimality condition, $\exists g_* \in \partial f(x_*)$:

$$\langle g_*, x - x_* \rangle \geq 0 \text{ for all } x \in Q.$$

Example 2: separation

For arbitrary $R > 0$, denote

$$\|s\|_R^* = \max_{x \in Q} \{ \langle s, x_* - x \rangle : \; \|x - x_*\| \leq R \}, \quad s \in E^*.$$

Note that $\|s\|_R^* \geq 0$ for any $s \in E^*$.

In view of the first-order optimality condition, $\exists g_* \in \partial f(x_*)$:

$$\langle g_*, x - x_* \rangle \geq 0 \text{ for all } x \in Q.$$

Therefore $\|g_*\|_R^* = 0$.

Example 2: separation

For arbitrary $R > 0$, denote

$$\|s\|_R^* = \max_{x \in Q} \{ \langle s, x_* - x \rangle : \; \|x - x_*\| \leq R \}, \quad s \in E^*.$$

Note that $\|s\|_R^* \geq 0$ for any $s \in E^*$.

In view of the first-order optimality condition, $\exists g_* \in \partial f(x_*)$:

$$\langle g_*, x - x_* \rangle \geq 0 \text{ for all } x \in Q.$$

Therefore $\|g_*\|_R^* = 0$.

Thus, $\|s\|_R^*$ measures the quality of hyperplane defined by s, trying to separate the feasible set Q and $\{x \in E : \; f(x) \leq f_*\}$.

Example 2: separation

For arbitrary $R > 0$, denote

$$\|s\|_R^* = \max_{x \in Q} \{ \langle s, x_* - x \rangle : \ \|x - x_*\| \leq R \}, \quad s \in E^*.$$

Note that $\|s\|_R^* \geq 0$ for any $s \in E^*$.

In view of the first-order optimality condition, $\exists g_* \in \partial f(x_*)$:

$$\langle g_*, x - x_* \rangle \geq 0 \text{ for all } x \in Q.$$

Therefore $\|g_*\|_R^* = 0$.

Thus, $\|s\|_R^*$ measures the quality of hyperplane defined by s, trying to separate the feasible set Q and $\{x \in E : \ f(x) \leq f_*\}$.

Corollary. For any $t \geq 0$ we have:

$$f(x_t) - f_* + \|s_t\|_R^* \leq \frac{1}{A_t}(B_t + \gamma_t G_R),$$

where $G_R = \max_{x \in Q} \{ d(x) : \ \|x - x^*\| \leq R \}$.

Subgradient method with double averaging

Subgradient method with double averaging

1. Compute $x_t^+ = \arg\min_{x \in Q} \{A_t \langle s_t, x \rangle + \gamma_t d(x)\}$.
2. Define $\tau_t = \frac{a_{t+1}}{A_{t+1}}$. Update $x_{t+1} = (1 - \tau_t) x_t + \tau_t x_t^+$.

Subgradient method with double averaging

1. Compute $x_t^+ = \arg\min_{x \in Q} \{A_t \langle s_t, x \rangle + \gamma_t d(x)\}$.
2. Define $\tau_t = \frac{a_{t+1}}{A_{t+1}}$. Update $x_{t+1} = (1 - \tau_t) x_t + \tau_t x_t^+$.

- $\tau_t \equiv 1$, $\gamma_t \equiv 1$ $\Rightarrow$ *mirror descent method* (1975).

Subgradient method with double averaging

1. Compute $x_t^+ = \arg\min\limits_{x \in Q} \{ A_t \langle s_t, x \rangle + \gamma_t d(x) \}$.
2. Define $\tau_t = \frac{a_{t+1}}{A_{t+1}}$. Update $x_{t+1} = (1 - \tau_t) x_t + \tau_t x_t^+$.

- $\tau_t \equiv 1$, $\gamma_t \equiv 1$ $\Rightarrow$ *mirror descent method* (1975).
- $\tau_t \equiv 1$ $\Rightarrow$ *primal-dual averaging scheme* (2003).

Subgradient method with double averaging

1. Compute $x_t^+ = \arg\min\limits_{x \in Q} \{A_t \langle s_t, x \rangle + \gamma_t d(x)\}$.
2. Define $\tau_t = \frac{a_{t+1}}{A_{t+1}}$. Update $x_{t+1} = (1 - \tau_t) x_t + \tau_t x_t^+$.

- $\tau_t \equiv 1$, $\gamma_t \equiv 1 \quad \Rightarrow \quad$ *mirror descent method* (1975).
- $\tau_t \equiv 1 \quad \Rightarrow \quad$ *primal-dual averaging scheme* (2003).
- $x_t^+ = \arg\min\limits_{x \in Q} [\ell_t(x) + \gamma_t d(x)]$.

Subgradient method with double averaging

1. Compute $x_t^+ = \arg\min\limits_{x \in Q} \{ A_t \langle s_t, x \rangle + \gamma_t d(x) \}$.
2. Define $\tau_t = \frac{a_{t+1}}{A_{t+1}}$. Update $x_{t+1} = (1-\tau_t)x_t + \tau_t x_t^+$.

- $\tau_t \equiv 1$, $\gamma_t \equiv 1 \quad \Rightarrow \quad$ *mirror descent method* (1975).
- $\tau_t \equiv 1 \quad \Rightarrow \quad$ *primal-dual averaging scheme* (2003).
- $x_t^+ = \arg\min\limits_{x \in Q} [\ell_t(x) + \gamma_t d(x)]$. It is easy to see that

$$x_t = \frac{1}{A_t} \left[a_0 x_0 + \sum_{k=0}^{t-1} a_{k+1} x_k^+ \right], \quad t \geq 0 \ .$$

Subgradient method with double averaging

1. Compute $x_t^+ = \arg\min_{x \in Q} \{ A_t \langle s_t, x \rangle + \gamma_t d(x) \}$.
2. Define $\tau_t = \frac{a_{t+1}}{A_{t+1}}$. Update $x_{t+1} = (1 - \tau_t) x_t + \tau_t x_t^+$.

- $\tau_t \equiv 1$, $\gamma_t \equiv 1$ $\Rightarrow$ *mirror descent method* (1975).
- $\tau_t \equiv 1$ $\Rightarrow$ *primal-dual averaging scheme* (2003).
- $x_t^+ = \arg\min_{x \in Q} [\ell_t(x) + \gamma_t d(x)]$. It is easy to see that

$$x_t = \frac{1}{A_t} \left[a_0 x_0 + \sum_{k=0}^{t-1} a_{k+1} x_k^+ \right], \quad t \geq 0 \ .$$

- Additional averaging parameters make the primal sequence more stable.

Convergence result

Convergence result

Theorem. Let sequence of parameters $\{\gamma_t\}_{t\geq 0}$ be monotone:

$$\gamma_{t+1} \geq \gamma_t, \quad t \geq 0.$$

Convergence result

Theorem. Let sequence of parameters $\{\gamma_t\}_{t \geq 0}$ be monotone:

$$\gamma_{t+1} \geq \gamma_t, \quad t \geq 0.$$

Then the estimate sequence condition holds with

$$B_t = \frac{1}{2} \sum_{k=0}^{t} \frac{a_k^2}{\gamma_{k-1}} \|\nabla f(x_k)\|_*^2,$$

where $\gamma_{-1} = \gamma_0$.

Convergence result

Theorem. Let sequence of parameters $\{\gamma_t\}_{t\geq 0}$ be monotone:

$$\gamma_{t+1} \geq \gamma_t, \quad t \geq 0.$$

Then the estimate sequence condition holds with

$$B_t = \frac{1}{2} \sum_{k=0}^{t} \frac{a_k^2}{\gamma_{k-1}} \|\nabla f(x_k)\|_*^2,$$

where $\gamma_{-1} = \gamma_0$. Moreover,

$$\frac{1}{\gamma_t} A_t (f(x_t) - f_*) + \frac{1}{2}\|x_t^+ - x_*\|^2 \leq d(x_*) + \frac{1}{\gamma_t} B_t.$$

.

Convergence result

Theorem. Let sequence of parameters $\{\gamma_t\}_{t\geq 0}$ be monotone:

$$\gamma_{t+1} \geq \gamma_t, \quad t \geq 0.$$

Then the estimate sequence condition holds with

$$B_t = \frac{1}{2} \sum_{k=0}^{t} \frac{a_k^2}{\gamma_{k-1}} \| \nabla f(x_k) \|_*^2,$$

where $\gamma_{-1} = \gamma_0$. Moreover,

$$\frac{1}{\gamma_t} A_t (f(x_t) - f_*) + \frac{1}{2} \| x_t^+ - x_* \|^2 \leq d(x_*) + \frac{1}{\gamma_t} B_t.$$

.

- Second statement ensures boundedness of the sequences.

Convergence result

Theorem. Let sequence of parameters $\{\gamma_t\}_{t \geq 0}$ be monotone:

$$\gamma_{t+1} \geq \gamma_t, \quad t \geq 0.$$

Then the estimate sequence condition holds with

$$B_t = \frac{1}{2} \sum_{k=0}^{t} \frac{a_k^2}{\gamma_{k-1}} \|\nabla f(x_k)\|_*^2,$$

where $\gamma_{-1} = \gamma_0$. Moreover,

$$\frac{1}{\gamma_t} A_t (f(x_t) - f_*) + \frac{1}{2} \|x_t^+ - x_*\|^2 \leq d(x_*) + \frac{1}{\gamma_t} B_t.$$

.

- Second statement ensures boundedness of the sequences.
- Recall: $s_t = \frac{1}{A_t} \sum_{k=0}^{t} a_k \nabla f(x_k)$

Convergence result

Theorem. Let sequence of parameters $\{\gamma_t\}_{t\geq 0}$ be monotone:

$$\gamma_{t+1} \geq \gamma_t, \quad t \geq 0.$$

Then the estimate sequence condition holds with

$$B_t = \frac{1}{2}\sum_{k=0}^{t} \frac{a_k^2}{\gamma_{k-1}} \|\nabla f(x_k)\|_*^2,$$

where $\gamma_{-1} = \gamma_0$. Moreover,

$$\frac{1}{\gamma_t} A_t (f(x_t) - f_*) + \frac{1}{2}\|x_t^+ - x_*\|^2 \leq d(x_*) + \frac{1}{\gamma_t} B_t.$$

.

- Second statement ensures boundedness of the sequences.
- Recall: $s_t = \frac{1}{A_t}\sum_{k=0}^{t} a_k \nabla f(x_k) \quad \Rightarrow \quad$ Convergence for points involved in the model:

$$f(x_t) - f^* \leq \frac{1}{A_t}\left(\gamma_t d(x_*) + \frac{1}{2}\sum_{k=0}^{t} \frac{a_k^2}{\gamma_{k-1}} \|\nabla f(x_k)\|_*^2\right).$$

Subgradient Method with Double Simple Averaging

Subgradient Method with Double Simple Averaging

1. Compute $x_t^+ = \arg\min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}$.
2. Update $x_{t+1} = \frac{t+1}{t+2} x_t + \frac{1}{t+2} x_t^+$.

Subgradient Method with Double Simple Averaging

1. Compute $x_t^+ = \arg\min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}$.
2. Update $x_{t+1} = \frac{t+1}{t+2} x_t + \frac{1}{t+2} x_t^+$.

Here, $s_t = \frac{1}{t+1} \sum_{k=0}^{t} \nabla f(x_k)$ and $x_t = \frac{1}{t+1} \left(x_0 + \sum_{k=0}^{t-1} x_k^+ \right)$.

Subgradient Method with Double Simple Averaging

1. Compute $x_t^+ = \arg\min\limits_{x \in Q} \left\{ \langle \sum\limits_{k=0}^{t} \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}$.
2. Update $x_{t+1} = \frac{t+1}{t+2} x_t + \frac{1}{t+2} x_t^+$.

Here, $s_t = \frac{1}{t+1} \sum\limits_{k=0}^{t} \nabla f(x_k)$ and $x_t = \frac{1}{t+1} \left(x_0 + \sum\limits_{k=0}^{t-1} x_k^+ \right)$.

Theorem. Let $\{\gamma_t\}_{t \geq 0}$ be nondecreasing. For any $t \geq 0$,

$$f(x_t) - f_* + \|s_t\|_R^* \leq \frac{1}{t+1} \left(\gamma_t G_R + \frac{1}{2} \sum_{k=0}^{t} \frac{\|\nabla f(x_k\|_*^2}{\gamma_{k-1}} \right),$$

where $\|s\|_R^* = \max\limits_{x \in Q} \langle s, x - x_* \rangle$, $G_R = \max\limits_{x \in Q} \{ d(x) : \|x - x_*\| \leq R \}$.

Subgradient Method with Double Simple Averaging

1. Compute $x_t^+ = \arg\min_{x\in Q} \left\{ \langle \sum_{k=0}^{t} \nabla f(x_k), x\rangle + \gamma_t d(x) \right\}$.
2. Update $x_{t+1} = \frac{t+1}{t+2} x_t + \frac{1}{t+2} x_t^+$.

Here, $s_t = \frac{1}{t+1} \sum_{k=0}^{t} \nabla f(x_k)$ and $x_t = \frac{1}{t+1} \left(x_0 + \sum_{k=0}^{t-1} x_k^+ \right)$.

Theorem. Let $\{\gamma_t\}_{t\geq 0}$ be nondecreasing. For any $t \geq 0$,

$$f(x_t) - f_* + \|s_t\|_R^* \leq \frac{1}{t+1} \left(\gamma_t G_R + \frac{1}{2} \sum_{k=0}^{t} \frac{\|\nabla f(x_k\|_*^2}{\gamma_{k-1}} \right),$$

where $\|s\|_R^* = \max_{x\in Q} \langle s, x - x_* \rangle$, $G_R = \max_{x\in Q} \{ d(x) : \|x - x_*\| \leq R \}$.

Corollary. Let $\|\nabla f(x)\|_* \leq L$, $\boxed{\gamma_t \to \infty, \text{ and } \frac{\gamma_t}{t+1} \to 0}$.

Subgradient Method with Double Simple Averaging

1. Compute $x_t^+ = \arg\min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}$.
2. Update $x_{t+1} = \frac{t+1}{t+2} x_t + \frac{1}{t+2} x_t^+$.

Here, $s_t = \frac{1}{t+1} \sum_{k=0}^{t} \nabla f(x_k)$ and $x_t = \frac{1}{t+1} \left(x_0 + \sum_{k=0}^{t-1} x_k^+ \right)$.

Theorem. Let $\{\gamma_t\}_{t \geq 0}$ be nondecreasing. For any $t \geq 0$,

$$f(x_t) - f_* + \|s_t\|_R^* \leq \frac{1}{t+1} \left(\gamma_t G_R + \frac{1}{2} \sum_{k=0}^{t} \frac{\|\nabla f(x_k\|_*^2}{\gamma_{k-1}} \right),$$

where $\|s\|_R^* = \max_{x \in Q} \langle s, x - x_* \rangle$, $G_R = \max_{x \in Q} \{ d(x) : \|x - x_*\| \leq R \}$.

Corollary. Let $\|\nabla f(x)\|_* \leq L$, $\boxed{\gamma_t \to \infty, \text{ and } \frac{\gamma_t}{t+1} \to 0}$. Then

$$\lim_{t \to \infty} f(x_t) = f^*, \qquad \lim_{t \to \infty} \|s_t\|_R^* = 0.$$

Subgradient Method with Double Simple Averaging

1. Compute $x_t^+ = \arg\min\limits_{x \in Q} \left\{ \langle \sum\limits_{k=0}^{t} \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}$.
2. Update $x_{t+1} = \frac{t+1}{t+2} x_t + \frac{1}{t+2} x_t^+$.

Here, $s_t = \frac{1}{t+1} \sum\limits_{k=0}^{t} \nabla f(x_k)$ and $x_t = \frac{1}{t+1} \left(x_0 + \sum\limits_{k=0}^{t-1} x_k^+ \right)$.

Theorem. Let $\{\gamma_t\}_{t \geq 0}$ be nondecreasing. For any $t \geq 0$,

$$f(x_t) - f_* + \|s_t\|_R^* \leq \frac{1}{t+1} \left(\gamma_t G_R + \frac{1}{2} \sum_{k=0}^{t} \frac{\|\nabla f(x_k\|_*^2}{\gamma_{k-1}} \right),$$

where $\|s\|_R^* = \max\limits_{x \in Q} \langle s, x - x_* \rangle$, $G_R = \max\limits_{x \in Q} \{ d(x) : \|x - x_*\| \leq R \}$.

Corollary. Let $\|\nabla f(x)\|_* \leq L$, $\boxed{\gamma_t \to \infty, \text{ and } \frac{\gamma_t}{t+1} \to 0}$. Then

$$\lim_{t \to \infty} f(x_t) = f^*, \qquad \lim_{t \to \infty} \|s_t\|_R^* = 0.$$

Optimal choice: $\gamma_t = \frac{L\sqrt{t+1}}{G_R^{1/2}}$

Subgradient Method with Double Simple Averaging

1. Compute $x_t^+ = \arg\min_{x \in Q} \left\{ \langle \sum_{k=0}^{t} \nabla f(x_k), x \rangle + \gamma_t d(x) \right\}$.
2. Update $x_{t+1} = \frac{t+1}{t+2} x_t + \frac{1}{t+2} x_t^+$.

Here, $s_t = \frac{1}{t+1} \sum_{k=0}^{t} \nabla f(x_k)$ and $x_t = \frac{1}{t+1} \left(x_0 + \sum_{k=0}^{t-1} x_k^+ \right)$.

Theorem. Let $\{\gamma_t\}_{t \geq 0}$ be nondecreasing. For any $t \geq 0$,

$$f(x_t) - f_* + \|s_t\|_R^* \leq \frac{1}{t+1} \left(\gamma_t G_R + \frac{1}{2} \sum_{k=0}^{t} \frac{\|\nabla f(x_k\|_*^2}{\gamma_{k-1}} \right),$$

where $\|s\|_R^* = \max_{x \in Q} \langle s, x - x_* \rangle$, $G_R = \max_{x \in Q} \{ d(x) : \|x - x_*\| \leq R \}$.

Corollary. Let $\|\nabla f(x)\|_* \leq L$, $\boxed{\gamma_t \to \infty, \text{ and } \frac{\gamma_t}{t+1} \to 0}$. Then

$$\lim_{t \to \infty} f(x_t) = f^*, \quad \lim_{t \to \infty} \|s_t\|_R^* = 0.$$

Optimal choice: $\gamma_t = \frac{L\sqrt{t+1}}{G_R^{1/2}} \Rightarrow f(x_t) - f_* + \|s_t\|_R^* \leq \frac{2LG_R^{1/2}}{\sqrt{t+1}}$.

Optimization problem with known minimax structure

Optimization problem with known minimax structure

Model: $f(x) = \hat{f}(x) + \max\limits_{u \in U} \{ \langle Au, x \rangle - \hat{\phi}(u) \}$,

Optimization problem with known minimax structure

Model: $f(x) = \hat{f}(x) + \max_{u \in U}\{\langle Au, x\rangle - \hat{\phi}(u)\}$, where

- $\hat{f}$ is a closed convex function on Q,
- U is a closed convex set in E_1,
- $\hat{\phi}(\cdot)$ is a closed convex function on U.

Optimization problem with known minimax structure

Model: $f(x) = \hat{f}(x) + \max\limits_{u \in U}\{\langle Au, x\rangle - \hat{\phi}(u)\}$, where

- $\hat{f}$ is a closed convex function on Q,
- U is a closed convex set in E_1,
- $\hat{\phi}(\cdot)$ is a closed convex function on U.

Adjoint problem:
$$f_* = \max_{u \in U}\left\{-\hat{\phi}(u) + \min_{x \in Q}[\langle Au, x\rangle + \hat{f}(x)]\right\}$$
$$= -\min_{u \in U}\left\{\hat{\phi}(u) + \hat{f}^*_Q(-Au)\right\},$$

Optimization problem with known minimax structure

Model: $f(x) = \hat{f}(x) + \max\limits_{u \in U}\{\langle Au, x\rangle - \hat{\phi}(u)\}$, where

- $\hat{f}$ is a closed convex function on Q,
- U is a closed convex set in E_1,
- $\hat{\phi}(\cdot)$ is a closed convex function on U.

Adjoint problem:
$$f_* = \max_{u \in U} \left\{ -\hat{\phi}(u) + \min_{x \in Q} [\langle Au, x\rangle + \hat{f}(x)] \right\}$$
$$= -\min_{u \in U} \left\{ \hat{\phi}(u) + \hat{f}^*_Q(-Au) \right\},$$

where $\hat{f}^*_Q(s) \stackrel{\mathrm{def}}{=} \max\limits_{x \in Q}[\langle s, x\rangle - \hat{f}(x)]$.

Optimization problem with known minimax structure

Model: $f(x) = \hat{f}(x) + \max\limits_{u \in U}\{\langle Au, x\rangle - \hat{\phi}(u)\}$, where

- $\hat{f}$ is a closed convex function on Q,
- U is a closed convex set in E_1,
- $\hat{\phi}(\cdot)$ is a closed convex function on U.

Adjoint problem:
$$f_* = \max_{u \in U} \left\{ -\hat{\phi}(u) + \min_{x \in Q}[\langle Au, x\rangle + \hat{f}(x)] \right\}$$
$$= -\min_{u \in U} \left\{ \hat{\phi}(u) + \hat{f}^*_Q(-Au) \right\},$$

where $\hat{f}^*_Q(s) \stackrel{\rm def}{=} \max\limits_{x \in Q}[\langle s, x\rangle - \hat{f}(x)]$.

PD-problem: $\min\limits_{x \in Q,\, u \in U}\{\Phi(x,u) \stackrel{\rm def}{=} f(x) + \hat{\phi}(u) + \hat{f}^*_Q(-Au)\} = 0.$

Optimization problem with known minimax structure

Model: $f(x) = \hat{f}(x) + \max\limits_{u \in U}\{\langle Au, x\rangle - \hat{\phi}(u)\}$, where

- $\hat{f}$ is a closed convex function on Q,
- U is a closed convex set in E_1,
- $\hat{\phi}(\cdot)$ is a closed convex function on U.

Adjoint problem:
$$f_* = \max_{u \in U}\left\{-\hat{\phi}(u) + \min_{x \in Q}[\langle Au, x\rangle + \hat{f}(x)]\right\} = -\min_{u \in U}\left\{\hat{\phi}(u) + \hat{f}^*_Q(-Au)\right\},$$

where $\hat{f}^*_Q(s) \stackrel{\text{def}}{=} \max\limits_{x \in Q}[\langle s, x\rangle - \hat{f}(x)]$.

PD-problem: $\min\limits_{x \in Q,\, u \in U}\{\Phi(x,u) \stackrel{\text{def}}{=} f(x) + \hat{\phi}(u) + \hat{f}^*_Q(-Au)\} = 0.$

Denote by $u(x) = \arg\max\limits_{u \in U}\{\langle Au, x\rangle - \hat{\phi}(u)\}$.

Optimization problem with known minimax structure

Model: $f(x) = \hat{f}(x) + \max\limits_{u \in U}\{\langle Au, x\rangle - \hat{\phi}(u)\}$, where

- $\hat{f}$ is a closed convex function on Q,
- U is a closed convex set in E_1,
- $\hat{\phi}(\cdot)$ is a closed convex function on U.

Adjoint problem:
$$f_* = \max_{u \in U}\left\{-\hat{\phi}(u) + \min_{x \in Q}[\langle Au, x\rangle + \hat{f}(x)]\right\} = -\min_{u \in U}\left\{\hat{\phi}(u) + \hat{f}_Q^*(-Au)\right\},$$

where $\hat{f}_Q^*(s) \stackrel{\mathrm{def}}{=} \max\limits_{x \in Q}[\langle s, x\rangle - \hat{f}(x)]$.

PD-problem: $\min\limits_{x \in Q,\, u \in U}\{\Phi(x,u) \stackrel{\mathrm{def}}{=} f(x) + \hat{\phi}(u) + \hat{f}_Q^*(-Au)\} = 0.$

Denote by $u(x) = \arg\max\limits_{u \in U}\{\langle Au, x\rangle - \hat{\phi}(u)\}$. Then

$$\nabla f(x) \stackrel{\mathrm{def}}{=} \nabla \hat{f}(x) + Au(x) \in \partial f(x).$$

Interpretation

Assume $d(x) \leq D$ for all $x \in Q$. Denote $u_k = u(x_k)$.

Interpretation

Assume $d(x) \leq D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Then

$$f(x_k) + \langle \nabla f(x_k), x - x_k \rangle$$

Interpretation

Assume $d(x) \leq D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Then

$$f(x_k) + \langle \nabla f(x_k), x - x_k \rangle$$
$$= \hat{f}(x_k) + \langle Au_k, x_k \rangle - \hat{\phi}(u_k) + \langle \nabla \hat{f}(x_k) + Au_k, x - x_k \rangle$$

Interpretation

Assume $d(x) \le D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Then

$$\begin{aligned} & f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \\ = & \hat{f}(x_k) + \langle A u_k, x_k \rangle - \hat{\phi}(u_k) + \langle \nabla \hat{f}(x_k) + A u_k, x - x_k \rangle \\ \le & \hat{f}(x) + \langle A u_k, x \rangle - \hat{\phi}(u_k). \end{aligned}$$

Interpretation

Assume $d(x) \leq D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Then

$$\begin{aligned} & f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \\ = & \hat{f}(x_k) + \langle Au_k, x_k \rangle - \hat{\phi}(u_k) + \langle \nabla \hat{f}(x_k) + Au_k, x - x_k \rangle \\ \leq & \hat{f}(x) + \langle Au_k, x \rangle - \hat{\phi}(u_k). \end{aligned}$$

Denote $\bar{u}_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k u_k \in U$.

Interpretation

Assume $d(x) \leq D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Then

$$\begin{aligned} & f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \\ = & \hat{f}(x_k) + \langle A u_k, x_k \rangle - \hat{\phi}(u_k) + \langle \nabla \hat{f}(x_k) + A u_k, x - x_k \rangle \\ \leq & \hat{f}(x) + \langle A u_k, x \rangle - \hat{\phi}(u_k). \end{aligned}$$

Denote $\bar{u}_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k u_k \in U$. Then

$$\ell_t(x) \leq A_t \hat{f}(x) + A_t \langle A \bar{u}_t, x \rangle - \sum_{k=0}^{t} a_k \hat{\phi}(u_k)$$

Interpretation

Assume $d(x) \le D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Then

$$\begin{aligned} & f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \\ = {} & \hat{f}(x_k) + \langle A u_k, x_k \rangle - \hat{\phi}(u_k) + \langle \nabla \hat{f}(x_k) + A u_k, x - x_k \rangle \\ \le {} & \hat{f}(x) + \langle A u_k, x \rangle - \hat{\phi}(u_k). \end{aligned}$$

Denote $\bar{u}_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k u_k \in U$. Then

$$\begin{aligned} \ell_t(x) & \le A_t \hat{f}(x) + A_t \langle A \bar{u}_t, x \rangle - \sum_{k=0}^{t} a_k \hat{\phi}(u_k) \\ & \le \; A_t[\hat{f}(x) + \langle A_t \bar{u}_t, x \rangle - \hat{\phi}(\bar{u}_t)]. \end{aligned}$$

Interpretation

Assume $d(x) \leq D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Then

$$
\begin{aligned}
& f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \\
= & \hat{f}(x_k) + \langle Au_k, x_k \rangle - \hat{\phi}(u_k) + \langle \nabla \hat{f}(x_k) + Au_k, x - x_k \rangle \\
\leq & \hat{f}(x) + \langle Au_k, x \rangle - \hat{\phi}(u_k).
\end{aligned}
$$

Denote $\bar{u}_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k u_k \in U$. Then

$$
\begin{aligned}
\ell_t(x) & \leq A_t \hat{f}(x) + A_t \langle A\bar{u}_t, x \rangle - \sum_{k=0}^{t} a_k \hat{\phi}(u_k) \\
& \leq A_t [\hat{f}(x) + \langle A_t \bar{u}_t, x \rangle - \hat{\phi}(\bar{u}_t)].
\end{aligned}
$$

Therefore, $\psi_t^* = \min\limits_{x \in Q} \{\ell_t(x) + \gamma_t d(x)\} \leq \min\limits_{x \in Q} \ell_t(x) + \gamma_t D$

Interpretation

Assume $d(x) \leq D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Then

$$\begin{aligned} & f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \\ = & \hat{f}(x_k) + \langle Au_k, x_k \rangle - \hat{\phi}(u_k) + \langle \nabla \hat{f}(x_k) + Au_k, x - x_k \rangle \\ \leq & \hat{f}(x) + \langle Au_k, x \rangle - \hat{\phi}(u_k). \end{aligned}$$

Denote $\bar{u}_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k u_k \in U$. Then

$$\begin{aligned} \ell_t(x) & \leq A_t \hat{f}(x) + A_t \langle A\bar{u}_t, x \rangle - \sum_{k=0}^{t} a_k \hat{\phi}(u_k) \\ & \leq A_t [\hat{f}(x) + \langle A_t \bar{u}_t, x \rangle - \hat{\phi}(\bar{u}_t)]. \end{aligned}$$

Therefore, $\psi_t^* = \min\limits_{x \in Q} \{ \ell_t(x) + \gamma_t d(x) \} \leq \min\limits_{x \in Q} \ell_t(x) + \gamma_t D$

$$\leq A_t \min_{x \in Q} [\hat{f}(x) + \langle A\bar{u}_t, x \rangle - \hat{\phi}(\bar{u}_t)] + \gamma_t D$$

Interpretation

Assume $d(x) \leq D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Then

$$\begin{aligned} & f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \\ = & \hat{f}(x_k) + \langle A u_k, x_k \rangle - \hat{\phi}(u_k) + \langle \nabla \hat{f}(x_k) + A u_k, x - x_k \rangle \\ \leq & \hat{f}(x) + \langle A u_k, x \rangle - \hat{\phi}(u_k). \end{aligned}$$

Denote $\bar{u}_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k u_k \in U$. Then

$$\begin{aligned} \ell_t(x) & \leq A_t \hat{f}(x) + A_t \langle A \bar{u}_t, x \rangle - \sum_{k=0}^{t} a_k \hat{\phi}(u_k) \\ & \leq A_t [\hat{f}(x) + \langle A_t \bar{u}_t, x \rangle - \hat{\phi}(\bar{u}_t)]. \end{aligned}$$

Therefore, $\psi_t^* = \min\limits_{x \in Q} \{\ell_t(x) + \gamma_t d(x)\} \leq \min\limits_{x \in Q} \ell_t(x) + \gamma_t D$

$$\begin{aligned} & \leq A_t \min_{x \in Q} [\hat{f}(x) + \langle A \bar{u}_t, x \rangle - \hat{\phi}(\bar{u}_t)] + \gamma_t D \\ & = -A_t [\hat{\phi}(\bar{u}_t) + \hat{f}_Q^*(-A \bar{u}_t)] + \gamma_t D. \end{aligned}$$

Interpretation

Assume $d(x) \leq D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Then

$$\begin{aligned} & f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \\ = & \hat{f}(x_k) + \langle A u_k, x_k \rangle - \hat{\phi}(u_k) + \langle \nabla \hat{f}(x_k) + A u_k, x - x_k \rangle \\ \leq & \hat{f}(x) + \langle A u_k, x \rangle - \hat{\phi}(u_k). \end{aligned}$$

Denote $\bar{u}_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k u_k \in U$. Then

$$\begin{aligned} \ell_t(x) & \leq A_t \hat{f}(x) + A_t \langle A \bar{u}_t, x \rangle - \sum_{k=0}^{t} a_k \hat{\phi}(u_k) \\ & \leq A_t [\hat{f}(x) + \langle A_t \bar{u}_t, x \rangle - \hat{\phi}(\bar{u}_t)]. \end{aligned}$$

Therefore, $\psi_t^* = \min\limits_{x \in Q} \{\ell_t(x) + \gamma_t d(x)\} \leq \min\limits_{x \in Q} \ell_t(x) + \gamma_t D$

$$\begin{aligned} & \leq A_t \min_{x \in Q} [\hat{f}(x) + \langle A \bar{u}_t, x \rangle - \hat{\phi}(\bar{u}_t)] + \gamma_t D \\ & = -A_t [\hat{\phi}(\bar{u}_t) + \hat{f}_Q^*(-A \bar{u}_t)] + \gamma_t D. \end{aligned}$$

Thus, $\Phi(x_t, \bar{u}_t) \leq \frac{1}{A_t} [\gamma_t D + B_t]$.

Interpretation

Assume $d(x) \leq D$ for all $x \in Q$. Denote $u_k = u(x_k)$. Then

$$
\begin{aligned}
& f(x_k) + \langle \nabla f(x_k), x - x_k \rangle \\
= & \hat{f}(x_k) + \langle A u_k, x_k \rangle - \hat{\phi}(u_k) + \langle \nabla \hat{f}(x_k) + A u_k, x - x_k \rangle \\
\leq & \hat{f}(x) + \langle A u_k, x \rangle - \hat{\phi}(u_k).
\end{aligned}
$$

Denote $\bar{u}_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k u_k \in U$. Then

$$
\begin{aligned}
\ell_t(x) & \leq A_t \hat{f}(x) + A_t \langle A \bar{u}_t, x \rangle - \sum_{k=0}^{t} a_k \hat{\phi}(u_k) \\
& \leq \; A_t [\hat{f}(x) + \langle A_t \bar{u}_t, x \rangle - \hat{\phi}(\bar{u}_t)].
\end{aligned}
$$

Therefore, $\psi_t^* = \min\limits_{x \in Q} \{ \ell_t(x) + \gamma_t d(x) \} \; \leq \; \min\limits_{x \in Q} \ell_t(x) + \gamma_t D$

$$
\begin{aligned}
& \leq \quad A_t \min_{x \in Q} [\hat{f}(x) + \langle A \bar{u}_t, x \rangle - \hat{\phi}(\bar{u}_t)] + \gamma_t D \\
& = \quad -A_t [\hat{\phi}(\bar{u}_t) + \hat{f}_Q^*(-A \bar{u}_t)] + \gamma_t D.
\end{aligned}
$$

Thus, $\Phi(x_t, \bar{u}_t) \leq \frac{1}{A_t} [\gamma_t D + B_t]$. **NB:** No computations of $\hat{f}_Q$!

Dual Lagrangian methods

Dual Lagrangian methods

Problem: $f^* = \min\limits_{x \in Q} \{ f_0(x) : \; f(x) \leq 0_m \},$

Dual Lagrangian methods

Problem: $f^* = \min\limits_{x \in Q} \{ f_0(x) : \; f(x) \leq 0_m \}$, where

- convex set $Q \subset E$ is closed convex set,
- all functional components are closed convex functions.

Dual Lagrangian methods

Problem: $f^* = \min\limits_{x \in Q} \{ f_0(x) : \; f(x) \leq 0_m \}$, where

- convex set $Q \subset E$ is closed convex set,
- all functional components are closed convex functions.

Lagrangian dual problem:

$$\max_{\lambda \geq 0_m} \left\{ \phi(\lambda) \stackrel{\mathrm{def}}{=} \min_{x \in Q} [f_0(x) + \langle \lambda, f(x) \rangle] \right\} \stackrel{\mathrm{def}}{=} f_*.$$

Dual Lagrangian methods

Problem: $f^* = \min\limits_{x \in Q} \{ f_0(x) : \; f(x) \leq 0_m \}$, where

- convex set $Q \subset E$ is closed convex set,
- all functional components are closed convex functions.

Lagrangian dual problem:

$$\max_{\lambda \geq 0_m} \left\{ \phi(\lambda) \stackrel{\mathrm{def}}{=} \min_{x \in Q} [f_0(x) + \langle \lambda, f(x) \rangle] \right\} \stackrel{\mathrm{def}}{=} f_*.$$

Main assumption: Q, f_0 and f are so simple, that $\phi(\lambda)$.

Dual Lagrangian methods

Problem: $f^* = \min\limits_{x \in Q} \{ f_0(x) : \ f(x) \leq 0_m \}$, where

- convex set $Q \subset E$ is closed convex set,
- all functional components are closed convex functions.

Lagrangian dual problem:

$$\max_{\lambda \geq 0_m} \left\{ \phi(\lambda) \stackrel{\mathrm{def}}{=} \min_{x \in Q} [f_0(x) + \langle \lambda, f(x) \rangle] \right\} \stackrel{\mathrm{def}}{=} f_*.$$

Main assumption: Q, f_0 and f are so simple, that $\phi(\lambda)$. Then

$$\nabla \phi(\lambda) = f(x(\lambda)), \quad x(\lambda) \in \mathrm{Arg} \min_{x \in Q} [f_0(x) + \langle \lambda, f(x) \rangle].$$

Dual Lagrangian methods

Problem: $f^* = \min\limits_{x\in Q}\{f_0(x) : \; f(x) \leq 0_m\}$, where

- convex set $Q \subset E$ is closed convex set,
- all functional components are closed convex functions.

Lagrangian dual problem:

$$\max_{\lambda \geq 0_m} \left\{ \phi(\lambda) \stackrel{\mathrm{def}}{=} \min_{x\in Q}[f_0(x) + \langle \lambda, f(x)\rangle] \right\} \stackrel{\mathrm{def}}{=} f_*.$$

Main assumption: Q, f_0 and f are so simple, that $\phi(\lambda)$. Then

$$\nabla\phi(\lambda) = f(x(\lambda)), \quad x(\lambda) \in \mathrm{Arg}\min_{x\in Q}[f_0(x) + \langle \lambda, f(x)\rangle].$$

Dual problem: $\max\limits_{\lambda\geq 0_m} \phi(\lambda).$

Dual Lagrangian methods

Problem: $f^* = \min\limits_{x \in Q} \{ f_0(x) : \; f(x) \leq 0_m \}$, where

- convex set $Q \subset E$ is closed convex set,
- all functional components are closed convex functions.

Lagrangian dual problem:

$$\max_{\lambda \geq 0_m} \left\{ \phi(\lambda) \stackrel{\mathrm{def}}{=} \min_{x \in Q} [f_0(x) + \langle \lambda, f(x) \rangle] \right\} \stackrel{\mathrm{def}}{=} f_* .$$

Main assumption: Q, f_0 and f are so simple, that $\phi(\lambda)$. Then

$$\nabla \phi(\lambda) = f(x(\lambda)), \quad x(\lambda) \in \operatorname{Arg} \min_{x \in Q} [f_0(x) + \langle \lambda, f(x) \rangle].$$

Dual problem: $\max\limits_{\lambda \geq 0_m} \phi(\lambda)$.

Prox-function: $d(\lambda) = \frac{1}{2} \| \lambda \|_2^2$, $\lambda_0 \; = \; 0_m$.

Analysis

Analysis

Note that

Analysis

Note that

$$\phi(\lambda_t) + \langle \nabla \phi(\lambda_t), \lambda - \lambda_t \rangle$$

Analysis

Note that

$$\begin{aligned}&\phi(\lambda_t) + \langle \nabla \phi(\lambda_t), \lambda - \lambda_t \rangle \\ = &f_0(x(\lambda_t)) + \langle \lambda_t, f(x(\lambda_t)) \rangle + \langle f(x(\lambda_t)), \lambda - \lambda_t \rangle\end{aligned}$$

Analysis

Note that

$$\begin{aligned} &\phi(\lambda_t) + \langle \nabla \phi(\lambda_t), \lambda - \lambda_t \rangle \\ &= f_0(x(\lambda_t)) + \langle \lambda_t, f(x(\lambda_t)) \rangle + \langle f(x(\lambda_t)), \lambda - \lambda_t \rangle \\ &= f_0(x(\lambda_t)) + \langle f(x(\lambda_t)), \lambda \rangle. \end{aligned}$$

Analysis

Note that

$$\begin{aligned}
&\phi(\lambda_t) + \langle \nabla\phi(\lambda_t), \lambda - \lambda_t\rangle \\
&= f_0(x(\lambda_t)) + \langle \lambda_t, f(x(\lambda_t))\rangle + \langle f(x(\lambda_t)), \lambda - \lambda_t\rangle \\
&= f_0(x(\lambda_t)) + \langle f(x(\lambda_t)), \lambda\rangle.
\end{aligned}$$

Denote $x_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k x(\lambda_k))$.

Analysis

Note that

$$\begin{aligned} &\phi(\lambda_t) + \langle \nabla \phi(\lambda_t), \lambda - \lambda_t \rangle \\ &= f_0(x(\lambda_t)) + \langle \lambda_t, f(x(\lambda_t)) \rangle + \langle f(x(\lambda_t)), \lambda - \lambda_t \rangle \\ &= f_0(x(\lambda_t)) + \langle f(x(\lambda_t)), \lambda \rangle. \end{aligned}$$

Denote $x_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k x(\lambda_k))$. Then

$$f_0(x_t) + \frac{A_t}{2\gamma_t} \| \left(f(x_t) \right)_+ \|_2^2 - \phi(\lambda_t) \le \frac{1}{A_t} B_t,$$

where $(a)_+ = \max\{a, 0\}$.

Analysis

Note that

$$\begin{aligned}
&\phi(\lambda_t) + \langle \nabla\phi(\lambda_t), \lambda - \lambda_t\rangle \\
&= f_0(x(\lambda_t)) + \langle \lambda_t, f(x(\lambda_t))\rangle + \langle f(x(\lambda_t)), \lambda - \lambda_t\rangle \\
&= f_0(x(\lambda_t)) + \langle f(x(\lambda_t)), \lambda\rangle.
\end{aligned}$$

Denote $x_t = \frac{1}{A_t}\sum\limits_{k=0}^{t} a_k x(\lambda_k))$. Then

$$f_0(x_t) + \frac{A_t}{2\gamma_t}\|\,(f(x_t))_+\,\|_2^2 - \phi(\lambda_t) \le \frac{1}{A_t}B_t,$$

where $(a)_+ = \max\{a, 0\}$.

Boundedness of subgradients:

Analysis

Note that

$$\begin{aligned} &\phi(\lambda_t) + \langle \nabla \phi(\lambda_t), \lambda - \lambda_t \rangle \\ = &f_0(x(\lambda_t)) + \langle \lambda_t, f(x(\lambda_t)) \rangle + \langle f(x(\lambda_t)), \lambda - \lambda_t \rangle \\ = &f_0(x(\lambda_t)) + \langle f(x(\lambda_t)), \lambda \rangle. \end{aligned}$$

Denote $x_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k x(\lambda_k))$. Then

$$f_0(x_t) + \frac{A_t}{2\gamma_t} \| \left(f(x_t) \right)_+ \|_2^2 - \phi(\lambda_t) \le \frac{1}{A_t} B_t,$$

where $(a)_+ = \max\{a, 0\}$.

Boundedness of subgradients: $\|f(x)\|_2 \le L$, $x \in Q$.

Analysis

Note that

$$\begin{aligned} &\phi(\lambda_t) + \langle \nabla\phi(\lambda_t), \lambda - \lambda_t \rangle \\ &= f_0(x(\lambda_t)) + \langle \lambda_t, f(x(\lambda_t)) \rangle + \langle f(x(\lambda_t)), \lambda - \lambda_t \rangle \\ &= f_0(x(\lambda_t)) + \langle f(x(\lambda_t)), \lambda \rangle. \end{aligned}$$

Denote $x_t = \frac{1}{A_t} \sum\limits_{k=0}^{t} a_k x(\lambda_k))$. Then

$$f_0(x_t) + \frac{A_t}{2\gamma_t} \| \left(f(x_t) \right)_+ \|_2^2 - \phi(\lambda_t) \le \frac{1}{A_t} B_t,$$

where $(a)_+ = \max\{a, 0\}$.

Boundedness of subgradients: $\|f(x)\|_2 \le L$, $x \in Q$.

NB: This is difficult to get by primal methods.

Privacy-respecting taxation

Privacy-respecting taxation

Goal: bound pollution produced by industry.

Privacy-respecting taxation

Goal: bound pollution produced by industry.

Producer i**:** chooses production volume $u_i \in \mathcal{U}_i \subset \mathbb{R}^{m_i}_+$, $i = 1 \dots n$.

Privacy-respecting taxation

Goal: bound pollution produced by industry.

Producer i**:** chooses production volume $u_i \in \mathcal{U}_i \subset \mathbb{R}^{m_i}_+$, $i = 1 \dots n$.
Utility function: concave function $\phi_i(u_i)$, $u_i \in \mathcal{U}_i$.

Privacy-respecting taxation

Goal: bound pollution produced by industry.

Producer i**:** chooses production volume $u_i \in \mathcal{U}_i \subset \mathbb{R}^{m_i}_+$, $i = 1 \dots n$.
Utility function: concave function $\phi_i(u_i)$, $u_i \in \mathcal{U}_i$.
NB: It is usually *unknown* to the coordination center.

Privacy-respecting taxation

Goal: bound pollution produced by industry.

Producer i**:** chooses production volume $u_i \in \mathcal{U}_i \subset \mathbb{R}^{m_i}_+$, $i = 1 \dots n$.
Utility function: concave function $\phi_i(u_i)$, $u_i \in \mathcal{U}_i$.
NB: It is usually *unknown* to the coordination center.

Social goal:

$$\max_{\{u_i\}_{i=1}^n} \left\{ \sum_{i=1}^n \phi_i(u_i) : \sum_{i=1}^n P_i u_i \leq b, \; u_i \in \mathcal{U}_i, \; i = 1, \dots, n \right\}, \text{ where}$$

- $b \in \mathbb{R}^m_+$ is the upper limits on pollution,
- P_i transforms the production u_i into the generated pollution.

Privacy-respecting taxation

Goal: bound pollution produced by industry.

Producer i**:** chooses production volume $u_i \in \mathcal{U}_i \subset \mathbb{R}^{m_i}_+$, $i = 1 \dots n$.
Utility function: concave function $\phi_i(u_i)$, $u_i \in \mathcal{U}_i$.
NB: It is usually *unknown* to the coordination center.

Social goal:

$$\max_{\{u_i\}_{i=1}^n} \left\{ \sum_{i=1}^n \phi_i(u_i) : \ \sum_{i=1}^n P_i u_i \leq b, \ u_i \in \mathcal{U}_i, \ i = 1, \dots, n \right\}, \text{ where}$$

- $b \in \mathbb{R}^m_+$ is the upper limits on pollution,
- P_i transforms the production u_i into the generated pollution.

Coordination tool: taxes $p \in \mathbb{R}^m_+$.

Privacy-respecting taxation

Goal: bound pollution produced by industry.

Producer i**:** chooses production volume $u_i \in \mathcal{U}_i \subset \mathbb{R}^{m_i}_+$, $i = 1 \dots n$.
Utility function: concave function $\phi_i(u_i)$, $u_i \in \mathcal{U}_i$.
NB: It is usually *unknown* to the coordination center.

Social goal:

$$\max_{\{u_i\}_{i=1}^n} \left\{ \sum_{i=1}^n \phi_i(u_i) : \ \sum_{i=1}^n P_i u_i \le b, \ u_i \in \mathcal{U}_i, \ i = 1, \dots, n \right\}, \text{ where}$$

- $b \in \mathbb{R}^m_+$ is the upper limits on pollution,
- P_i transforms the production u_i into the generated pollution.

Coordination tool: taxes $p \in \mathbb{R}^m_+$.

Reaction of producers: $f_i(p) = \max\limits_{u_i} [\phi_i(u_i) - \langle p, P_i u_i \rangle : \ u_i \in \mathcal{U}_i]$.

Privacy-respecting taxation

Goal: bound pollution produced by industry.

Producer i**:** chooses production volume $u_i \in \mathcal{U}_i \subset \mathbb{R}^{m_i}_+$, $i = 1 \dots n$.
Utility function: concave function $\phi_i(u_i)$, $u_i \in \mathcal{U}_i$.
NB: It is usually ***unknown*** to the coordination center.

Social goal:

$$\max_{\{u_i\}_{i=1}^n} \left\{ \sum_{i=1}^n \phi_i(u_i) : \ \sum_{i=1}^n P_i u_i \le b, \ u_i \in \mathcal{U}_i, \ i = 1, \dots, n \right\}, \text{ where}$$

- $b \in \mathbb{R}^m_+$ is the upper limits on pollution,
- P_i transforms the production u_i into the generated pollution.

Coordination tool: taxes $p \in \mathbb{R}^m_+$.

Reaction of producers: $f_i(p) = \max\limits_{u_i} [\phi_i(u_i) - \langle p, P_i u_i \rangle : \ u_i \in \mathcal{U}_i]$.

Denote by $u_i(p)$ its optimal solution. Then $-P_i u_i(p) \in \partial f_i(p)$.

Coordination problem

Coordination problem

We assume that coordinator is solving the problem

$$\min_{p \geq 0} \left\{ f(p) \stackrel{\mathrm{def}}{=} \langle b, p \rangle + \sum_{i=1}^{n} f_i(p) \right\}.$$

Coordination problem

We assume that coordinator is solving the problem

$$\min_{p \geq 0} \left\{ f(p) \stackrel{\mathrm{def}}{=} \langle b, p \rangle + \sum_{i=1}^{n} f_i(p) \right\}.$$

(It is dual to the socially optimal distribution.)

Coordination problem

We assume that coordinator is solving the problem

$$\min_{p \geq 0} \left\{ f(p) \stackrel{\mathrm{def}}{=} \langle b, p \rangle + \sum_{i=1}^{n} f_i(p) \right\}.$$

(It is dual to the socially optimal distribution.)

Gradient: $\nabla f(p) = b - v(p)$, $v(p) \stackrel{\mathrm{def}}{=} \sum_{i=1}^{n} P_i u_i(p)$.

Coordination problem

We assume that coordinator is solving the problem

$$\min_{p \geq 0} \left\{ f(p) \stackrel{\mathrm{def}}{=} \langle b, p \rangle + \sum_{i=1}^{n} f_i(p) \right\}.$$

(It is dual to the socially optimal distribution.)

Gradient: $\nabla f(p) = b - v(p)$, $v(p) \stackrel{\mathrm{def}}{=} \sum_{i=1}^{n} P_i u_i(p)$.

Interpretation: $-\nabla f(p)$ is the *excessive pollution*.

Coordination problem

We assume that coordinator is solving the problem

$$\min_{p \geq 0} \left\{ f(p) \stackrel{\text{def}}{=} \langle b, p \rangle + \sum_{i=1}^{n} f_i(p) \right\}.$$

(It is dual to the socially optimal distribution.)

Gradient: $\nabla f(p) = b - v(p)$, $v(p) \stackrel{\text{def}}{=} \sum_{i=1}^{n} P_i u_i(p)$.

Interpretation: $-\nabla f(p)$ is the *excessive pollution*.

Optimality condition: $\langle \nabla f(p_*), p - p_* \rangle \geq 0$, $p \in \mathbb{R}^m_+$.

Coordination problem

We assume that coordinator is solving the problem

$$\min_{p \geq 0} \left\{ f(p) \stackrel{\mathrm{def}}{=} \langle b, p \rangle + \sum_{i=1}^{n} f_i(p) \right\}.$$

(It is dual to the socially optimal distribution.)

Gradient: $\nabla f(p) = b - v(p)$, $v(p) \stackrel{\mathrm{def}}{=} \sum\limits_{i=1}^{n} P_i u_i(p)$.

Interpretation: $-\nabla f(p)$ is the *excessive pollution*.

Optimality condition: $\langle \nabla f(p_*), p - p_* \rangle \geq 0$, $p \in \mathbb{R}^m_+$.

- Positive optimal tax $\Rightarrow$ no excessive pollution.

Coordination problem

We assume that coordinator is solving the problem

$$\min_{p \geq 0} \left\{ f(p) \stackrel{\mathrm{def}}{=} \langle b, p \rangle + \sum_{i=1}^{n} f_i(p) \right\}.$$

(It is dual to the socially optimal distribution.)

Gradient: $\nabla f(p) = b - v(p)$, $v(p) \stackrel{\mathrm{def}}{=} \sum\limits_{i=1}^{n} P_i u_i(p)$.

Interpretation: $-\nabla f(p)$ is the *excessive pollution*.

Optimality condition: $\langle \nabla f(p_*), p - p_* \rangle \geq 0$, $p \in \mathbb{R}^m_+$.

- Positive optimal tax $\Rightarrow$ no excessive pollution.
- Zero tax $\Rightarrow$ excessive pollution is non-positive.

Coordination problem

We assume that coordinator is solving the problem

$$\min_{p \geq 0} \left\{ f(p) \stackrel{\mathrm{def}}{=} \langle b, p \rangle + \sum_{i=1}^{n} f_i(p) \right\}.$$

(It is dual to the socially optimal distribution.)

Gradient: $\nabla f(p) = b - v(p)$, $v(p) \stackrel{\mathrm{def}}{=} \sum\limits_{i=1}^{n} P_i u_i(p)$.

Interpretation: $-\nabla f(p)$ is the *excessive pollution*.

Optimality condition: $\langle \nabla f(p_*), p - p_* \rangle \geq 0$, $p \in \mathbb{R}^m_+$.

- Positive optimal tax $\Rightarrow$ no excessive pollution.
- Zero tax $\Rightarrow$ excessive pollution is non-positive.

Main difficulty: utility functions of producers are hidden.

Coordination problem

We assume that coordinator is solving the problem

$$\min_{p \geq 0} \left\{ f(p) \stackrel{\mathrm{def}}{=} \langle b, p \rangle + \sum_{i=1}^{n} f_i(p) \right\}.$$

(It is dual to the socially optimal distribution.)

Gradient: $\nabla f(p) = b - v(p)$, $v(p) \stackrel{\mathrm{def}}{=} \sum\limits_{i=1}^{n} P_i u_i(p)$.

Interpretation: $-\nabla f(p)$ is the *excessive pollution*.

Optimality condition: $\langle \nabla f(p_*), p - p_* \rangle \geq 0$, $p \in \mathbb{R}^m_+$.

- Positive optimal tax $\Rightarrow$ no excessive pollution.
- Zero tax $\Rightarrow$ excessive pollution is non-positive.

Main difficulty: utility functions of producers are hidden.
We can observe only the *aggregated pollution* $v(p)$.

Double Simple Averaging for Taxation

Double Simple Averaging for Taxation

Prox-function: $d(p) = \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\varkappa_j}(p^{(j)})^2$, where $\varkappa_j > 0$ are scaling coefficients.

Double Simple Averaging for Taxation

Prox-function: $d(p) = \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\varkappa_j}(p^{(j)})^2$, where $\varkappa_j > 0$ are scaling coefficients. (Hence, $p[0] = p_0 = 0$.)

Double Simple Averaging for Taxation

Prox-function: $d(p) = \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\varkappa_j}(p^{(j)})^2$, where $\varkappa_j > 0$ are scaling coefficients. (Hence, $p[0] = p_0 = 0$.)

Denote $s[t] = \frac{1}{t+1}\sum\limits_{k=0}^{t} \nabla f(p[k]) \; = \; b - \frac{1}{t+1}\sum\limits_{k=0}^{t} v(p[k])$,

Double Simple Averaging for Taxation

Prox-function: $d(p) = \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\varkappa_j}(p^{(j)})^2$, where $\varkappa_j > 0$ are scaling coefficients. (Hence, $p[0] = p_0 = 0$.)

Denote $s[t] = \frac{1}{t+1}\sum\limits_{k=0}^{t} \nabla f(p[k]) \; = \; b - \frac{1}{t+1}\sum\limits_{k=0}^{t} v(p[k])$, and

$S[t] = -(t+1)s[t] = \sum\limits_{k=0}^{t}(v(p[k]) - b)$, the aggregated pollution.

Double Simple Averaging for Taxation

Prox-function: $d(p) = \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\varkappa_j}(p^{(j)})^2$, where $\varkappa_j > 0$ are scaling coefficients. (Hence, $p[0] = p_0 = 0$.)

Denote $s[t] = \frac{1}{t+1}\sum\limits_{k=0}^{t} \nabla f(p[k]) \; = \; b - \frac{1}{t+1}\sum\limits_{k=0}^{t} v(p[k])$, and

$S[t] = -(t+1)s[t] = \sum\limits_{k=0}^{t} (v(p[k]) - b)$, the aggregated pollution.

1. Measure the total pollution volume $v(p[k])$, and update aggregate excessive pollution $S[t] = S[t-1] + v(p[k]) - b$.

Double Simple Averaging for Taxation

Prox-function: $d(p) = \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\varkappa_j}(p^{(j)})^2$, where $\varkappa_j > 0$ are scaling coefficients. (Hence, $p[0] = p_0 = 0$.)

Denote $s[t] = \frac{1}{t+1}\sum\limits_{k=0}^{t} \nabla f(p[k]) \; = \; b - \frac{1}{t+1}\sum\limits_{k=0}^{t} v(p[k])$, and

$S[t] = -(t+1)s[t] = \sum\limits_{k=0}^{t} (v(p[k]) - b)$, the aggregated pollution.

1. Measure the total pollution volume $v(p[k])$, and update aggregate excessive pollution $S[t] = S[t-1] + v(p[k]) - b$.
2. Compute the tax predictions $p_+^{(j)}[t] = \frac{\varkappa_j}{\gamma_t}\left(S^{(j)}[t]\right)_+$, $j = 1, \dots, m$.

Double Simple Averaging for Taxation

Prox-function: $d(p) = \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\varkappa_j}(p^{(j)})^2$, where $\varkappa_j > 0$ are scaling coefficients. (Hence, $p[0] = p_0 = 0$.)

Denote $s[t] = \frac{1}{t+1}\sum\limits_{k=0}^{t} \nabla f(p[k]) \;=\; b - \frac{1}{t+1}\sum\limits_{k=0}^{t} v(p[k])$, and

$S[t] = -(t+1)s[t] = \sum\limits_{k=0}^{t} (v(p[k]) - b)$, the aggregated pollution.

1. Measure the total pollution volume $v(p[k])$, and update aggregate excessive pollution $S[t] = S[t-1] + v(p[k]) - b$.
2. Compute the tax predictions $p_+^{(j)}[t] = \frac{\varkappa_j}{\gamma_t}\left(S^{(j)}[t]\right)_+$, $j = 1, \dots, m$.
3. Define new vector of taxes $p[t+1] = \frac{t+1}{t+2} p[t] + \frac{1}{t+2} p_+[t]$.

Double Simple Averaging for Taxation

Prox-function: $d(p) = \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\varkappa_j}(p^{(j)})^2$, where $\varkappa_j > 0$ are scaling coefficients. (Hence, $p[0] = p_0 = 0$.)

Denote $s[t] = \frac{1}{t+1}\sum\limits_{k=0}^{t} \nabla f(p[k]) \;=\; b - \frac{1}{t+1}\sum\limits_{k=0}^{t} v(p[k])$, and

$S[t] = -(t+1)s[t] = \sum\limits_{k=0}^{t}(v(p[k]) - b)$, the aggregated pollution.

1. Measure the total pollution volume $v(p[k])$, and update aggregate excessive pollution $S[t] = S[t-1] + v(p[k]) - b$.
2. Compute the tax predictions $p_+^{(j)}[t] = \frac{\varkappa_j}{\gamma_t}\left(S^{(j)}[t]\right)_+$, $j = 1, \dots, m$.
3. Define new vector of taxes $p[t+1] = \frac{t+1}{t+2}p[t] + \frac{1}{t+2}p_+[t]$.

Theorem. Let $\gamma_t = O(\sqrt{t})$.

Double Simple Averaging for Taxation

Prox-function: $d(p) = \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\varkappa_j}(p^{(j)})^2$, where $\varkappa_j > 0$ are scaling coefficients. (Hence, $p[0] = p_0 = 0$.)

Denote $s[t] = \frac{1}{t+1}\sum\limits_{k=0}^{t} \nabla f(p[k]) \;=\; b - \frac{1}{t+1}\sum\limits_{k=0}^{t} v(p[k])$, and

$S[t] = -(t+1)s[t] = \sum\limits_{k=0}^{t}(v(p[k]) - b)$, the aggregated pollution.

1. Measure the total pollution volume $v(p[k])$, and update aggregate excessive pollution $S[t] = S[t-1] + v(p[k]) - b$.
2. Compute the tax predictions $p_+^{(j)}[t] = \frac{\varkappa_j}{\gamma_t}\left(S^{(j)}[t]\right)_+$, $j = 1, \dots, m$.
3. Define new vector of taxes $p[t+1] = \frac{t+1}{t+2}p[t] + \frac{1}{t+2}p_+[t]$.

Theorem. Let $\gamma_t = O(\sqrt{t})$. Then:

- Taxes $p[t]$ converge to the optimal solution of dual problem.

Double Simple Averaging for Taxation

Prox-function: $d(p) = \frac{1}{2}\sum_{j=1}^{m} \frac{1}{\varkappa_j}(p^{(j)})^2$, where $\varkappa_j > 0$ are scaling coefficients. (Hence, $p[0] = p_0 = 0$.)

Denote $s[t] = \frac{1}{t+1}\sum\limits_{k=0}^{t} \nabla f(p[k]) \; = \; b - \frac{1}{t+1}\sum\limits_{k=0}^{t} v(p[k])$, and

$S[t] = -(t+1)s[t] = \sum\limits_{k=0}^{t}(v(p[k]) - b)$, the aggregated pollution.

1. Measure the total pollution volume $v(p[k])$, and update aggregate excessive pollution $S[t] = S[t-1] + v(p[k]) - b$.
2. Compute the tax predictions $p_+^{(j)}[t] = \frac{\varkappa_j}{\gamma_t}\left(S^{(j)}[t]\right)_+$, $j = 1, \dots, m$.
3. Define new vector of taxes $p[t+1] = \frac{t+1}{t+2}p[t] + \frac{1}{t+2}p_+[t]$.

Theorem. Let $\gamma_t = O(\sqrt{t})$. Then:

- Taxes $p[t]$ converge to the optimal solution of dual problem.
- Historical averages of personal production $\frac{1}{t+1}\sum_{k=0}^{t} u_i(p[k])$, converge to the socially optimal solution.

Can we do this by another scheme?

Can we do this by another scheme?

1. **Best value.**

Can we do this by another scheme?

1. **Best value.** $p_*[t] = \arg \min\limits_{0 \leq k \leq t} f(p[k])$.

Can we do this by another scheme?

1. Best value. $p_*[t] = \arg\min_{0 \le k \le t} f(p[k])$. **No!**

Can we do this by another scheme?

1. **Best value.** $p_*[t] = \arg \min\limits_{0 \le k \le t} f(p[k])$. **No!**

   - We cannot compute the objective function.

Can we do this by another scheme?

1. **Best value.** $p_*[t] = \arg\min_{0 \le k \le t} f(p[k])$. **No!**

- We cannot compute the objective function.
- We need to stop the process and use a tax from the past.

Can we do this by another scheme?

1. **Best value.** $p_*[t] = \arg \min\limits_{0 \le k \le t} f(p[k])$. **No!**
   - We cannot compute the objective function.
   - We need to stop the process and use a tax from the past.

2. **Average value.**

Can we do this by another scheme?

1. **Best value.** $p_*[t] = \arg\min_{0 \le k \le t} f(p[k])$. **No!**
   - We cannot compute the objective function.
   - We need to stop the process and use a tax from the past.

2. **Average value.** $p_*[t] = \frac{1}{t+1} \sum_{k=0}^{t} p[k]$.

Can we do this by another scheme?

1. **Best value.** $p_*[t] = \arg\min\limits_{0 \le k \le t} f(p[k])$. **No!**
   - We cannot compute the objective function.
   - We need to stop the process and use a tax from the past.

2. **Average value.** $p_*[t] = \frac{1}{t+1} \sum\limits_{k=0}^{t} p[k]$. **No!**

Can we do this by another scheme?

1. **Best value.** $p_*[t] = \arg\min_{0 \le k \le t} f(p[k])$. **No!**

   - We cannot compute the objective function.
   - We need to stop the process and use a tax from the past.

2. **Average value.** $p_*[t] = \frac{1}{t+1} \sum_{k=0}^{t} p[k]$. **No!**

   - We know the good taxes

Can we do this by another scheme?

1. **Best value.** $p_*[t] = \arg\min_{0 \le k \le t} f(p[k])$. **No!**
   - We cannot compute the objective function.
   - We need to stop the process and use a tax from the past.

2. **Average value.** $p_*[t] = \frac{1}{t+1} \sum_{k=0}^{t} p[k]$. **No!**
   - We know the good taxes, but we never use them in the real life.

Numerical experiments: Test function

Numerical experiments: Test function

Problem: $f(x) = \max \left\{ |x^{(1)}|, \max\limits_{2 \leq i \leq n} |x^{(i)} - 2x^{(i-1)}| \right\}.$

Numerical experiments: Test function

Problem: $$f(x) = \max\left\{ |x^{(1)}|, \max_{2 \le i \le n} |x^{(i)} - 2x^{(i-1)}| \right\}.$$

It is a homogeneous convex function of degree one.

Numerical experiments: Test function

Problem: $$f(x) = \max \left\{ |x^{(1)}|, \max_{2 \le i \le n} |x^{(i)} - 2x^{(i-1)}| \right\}.$$

It is a homogeneous convex function of degree one.

Thus, $f_* = \min\limits_{x \in \mathbb{R}^n} f(x) = 0$ and $x_* = 0_n$.

Numerical experiments: Test function

Problem: $$f(x) = \max\left\{ |x^{(1)}|, \max_{2 \le i \le n} |x^{(i)} - 2x^{(i-1)}| \right\}.$$

It is a homogeneous convex function of degree one.

Thus, $f_* = \min\limits_{x \in \mathbb{R}^n} f(x) = 0$ and $x_* = 0_n$.

Condition number.

Numerical experiments: Test function

Problem: $$f(x) = \max\left\{ |x^{(1)}|, \max_{2 \le i \le n} |x^{(i)} - 2x^{(i-1)}| \right\}.$$

It is a homogeneous convex function of degree one.

Thus, $f_* = \min\limits_{x \in \mathbb{R}^n} f(x) = 0$ and $x_* = 0_n$.

Condition number. Consider $\bar{x} \in \mathbb{R}^n$:
$$\bar{x}^{(1)} = 1, \quad \bar{x}^{(i+1)} = 2\bar{x}^{(i)} + 1, \quad i = 1, \dots, n-1.$$

Numerical experiments: Test function

Problem: $$f(x) = \max\left\{|x^{(1)}|, \max_{2\leq i\leq n} |x^{(i)} - 2x^{(i-1)}|\right\}.$$

It is a homogeneous convex function of degree one.

Thus, $f_* = \min\limits_{x\in\mathbb{R}^n} f(x) = 0$ and $x_* = 0_n$.

Condition number. Consider $\bar{x} \in \mathbb{R}^n$:

$$\bar{x}^{(1)} = 1, \quad \bar{x}^{(i+1)} = 2\bar{x}^{(i)} + 1, \quad i = 1,\ldots,n-1.$$

Then $\bar{x}^{(i)} = 2^{i+1} - 1$, $i = 1,\ldots,n$.

Numerical experiments: Test function

Problem: $f(x) = \max\left\{ |x^{(1)}|, \max\limits_{2 \le i \le n} |x^{(i)} - 2x^{(i-1)}| \right\}.$

It is a homogeneous convex function of degree one.

Thus, $f_* = \min\limits_{x \in \mathbb{R}^n} f(x) = 0$ and $x_* = 0_n$.

Condition number. Consider $\bar{x} \in \mathbb{R}^n$:

$$\bar{x}^{(1)} = 1, \quad \bar{x}^{(i+1)} = 2\bar{x}^{(i)} + 1, \quad i = 1, \dots, n-1.$$

Then $\bar{x}^{(i)} = 2^{i+1} - 1$, $i = 1, \dots, n$. Therefore $f(\bar{x}) = f(1_n) = 1$.

Numerical experiments: Test function

Problem: $f(x) = \max \left\{ |x^{(1)}|, \max\limits_{2 \le i \le n} |x^{(i)} - 2x^{(i-1)}| \right\}.$

It is a homogeneous convex function of degree one.

Thus, $f_* = \min\limits_{x \in \mathbb{R}^n} f(x) = 0$ and $x_* = 0_n$.

Condition number. Consider $\bar{x} \in \mathbb{R}^n$:

$$\bar{x}^{(1)} = 1, \quad \bar{x}^{(i+1)} = 2\bar{x}^{(i)} + 1, \quad i = 1, \dots, n-1.$$

Then $\bar{x}^{(i)} = 2^{i+1} - 1$, $i = 1, \dots, n$. Therefore $f(\bar{x}) = f(1_n) = 1$.

Thus, $\kappa_\infty(f) \ge 2^{n+1} - 1$.

Numerical experiments: Test function

Problem: $$f(x) = \max\left\{ |x^{(1)}|, \max_{2\le i\le n} |x^{(i)} - 2x^{(i-1)}| \right\}.$$

It is a homogeneous convex function of degree one.

Thus, $f_* = \min\limits_{x\in\mathbb{R}^n} f(x) = 0$ and $x_* = 0_n$.

Condition number. Consider $\bar{x} \in \mathbb{R}^n$:

$$\bar{x}^{(1)} = 1, \quad \bar{x}^{(i+1)} = 2\bar{x}^{(i)} + 1, \quad i = 1, \dots, n-1.$$

Then $\bar{x}^{(i)} = 2^{i+1} - 1$, $i = 1, \dots, n$. Therefore $f(\bar{x}) = f(1_n) = 1$.

Thus, $\kappa_\infty(f) \ge 2^{n+1} - 1$.

Let us choose $x_0 = 1_n$. Then $R \stackrel{\text{def}}{=} \|x_0 - x_*\|_2 = \sqrt{n}$ and

$$\|\nabla f(x)\|_* \le L \stackrel{\text{def}}{=} \sqrt{5}, \; x \in \mathbb{R}^n.$$

Numerical experiments: Test function

Problem: $$f(x) = \max\left\{ |x^{(1)}|, \max_{2 \le i \le n} |x^{(i)} - 2x^{(i-1)}| \right\}.$$

It is a homogeneous convex function of degree one.

Thus, $f_* = \min\limits_{x \in \mathbb{R}^n} f(x) = 0$ and $x_* = 0_n$.

Condition number. Consider $\bar{x} \in \mathbb{R}^n$:

$$\bar{x}^{(1)} = 1, \quad \bar{x}^{(i+1)} = 2\bar{x}^{(i)} + 1, \quad i = 1, \dots, n-1.$$

Then $\bar{x}^{(i)} = 2^{i+1} - 1$, $i = 1, \dots, n$. Therefore $f(\bar{x}) = f(1_n) = 1$.

Thus, $\kappa_\infty(f) \ge 2^{n+1} - 1$.

Let us choose $x_0 = 1_n$. Then $R \stackrel{\text{def}}{=} \|x_0 - x_*\|_2 = \sqrt{n}$ and

$$\|\nabla f(x)\|_* \le L \stackrel{\text{def}}{=} \sqrt{5}, \; x \in \mathbb{R}^n.$$

We assume R and L be known for the methods.

Numerical experiments: Results for $\epsilon = 2^{-6} = 0.0156$

Dim.	PGM	SDA	SA_2	$SA_2(\%)$	L^2R^2/ϵ^2
10	51 204	9 254	586	0.29	204 800
20	102 405	65 536	1 587	0.39	409 600
40	204 805	131 072	4 094	0.50	819 200
80	409 616	262 144	6 655	0.41	1 638 400
160	819 209	524 288	16 484	0.50	3 276 800
320	1 638 409	1 048 576	35 184	0.54	6 553 600
640	3 276 807	2 097 152	73 390	0.56	13 107 200
1 280	6 553 612	4 194 304	143 475	0.55	26 214 400
2 560	13 107 205	8 388 608	309 681	0.59	52 428 800
5 120	26 214 405	16 777 216	579 893	0.55	104 857 600
10 240	52 428 810	33 554 432	1 181 849	0.56	209 715 200

Numerical experiments: Results for $\epsilon = 2^{-6} = 0.0156$

Dim.	PGM	SDA	SA_2	SA_2(%)	L^2R^2/ϵ^2
10	51 204	9 254	586	0.29	204 800
20	102 405	65 536	1 587	0.39	409 600
40	204 805	131 072	4 094	0.50	819 200
80	409 616	262 144	6 655	0.41	1 638 400
160	819 209	524 288	16 484	0.50	3 276 800
320	1 638 409	1 048 576	35 184	0.54	6 553 600
640	3 276 807	2 097 152	73 390	0.56	13 107 200
1 280	6 553 612	4 194 304	143 475	0.55	26 214 400
2 560	13 107 205	8 388 608	309 681	0.59	52 428 800
5 120	26 214 405	16 777 216	579 893	0.55	104 857 600
10 240	52 428 810	33 554 432	1 181 849	0.56	209 715 200

NB: SA_2 is a clear winner.

Conclusion

Conclusion

- We presented the first converging SGM.

Conclusion

- We presented the first converging SGM.
- It demonstrates a high practical efficiency.

Conclusion

- We presented the first converging SGM.
- It demonstrates a high practical efficiency.
- It can be applied for real-life real-time adjustment.

Conclusion

- We presented the first converging SGM.
- It demonstrates a high practical efficiency.
- It can be applied for real-life real-time adjustment.
- Some questions are not clear:

Conclusion

- We presented the first converging SGM.
- It demonstrates a high practical efficiency.
- It can be applied for real-life real-time adjustment.
- Some questions are not clear:
  - Stochastic version.

Conclusion

- We presented the first converging SGM.
- It demonstrates a high practical efficiency.
- It can be applied for real-life real-time adjustment.
- Some questions are not clear:
  - Stochastic version.
  - Online optimization, etc.

Conclusion

- We presented the first converging SGM.
- It demonstrates a high practical efficiency.
- It can be applied for real-life real-time adjustment.
- Some questions are not clear:
  - Stochastic version.
  - Online optimization, etc.

THANK YOU FOR YOUR ATTENTION!