Large Scale & Distributed Optimization

Angelia Nedić

January 20, 2022

ASU
ARIZONA STATE UNIVERSITY

contact: *Angelia.Nedich@asu.edu; angelianedich@gmail.com*

Lecture 4:

Random Algorithms for Problems with many Constraints

Problem

- We consider the problem of the following form:

$$\begin{aligned} &\text{minimize} \quad f(x) \\ &\text{subject to} \quad x \in Y \cap (\cap_{i=1}^m X_i) \end{aligned}$$

  where $f : \mathbb{R}^n \to \mathbb{R}$ is convex, $Y \subseteq \mathbb{R}^n$ is convex and closed set, and each $X_i \subseteq \mathbb{R}^n$ is convex and closed set.

- Throughout we assume that the constraint set $Y \cap (\cap_{i=1}^m X_i)$ is nonempty

- The set Y represents a simple constraint set admitting projections easily (such as non-negative orthant, a ball, or a box)

- *It is assumed that the projection on the entire constraint set $Y \cap (\cap_{i=1}^m X_i)$ is complicated* due to large m (or due to complex structure of some of the sets X_i, such as when X_i is given by a convex inequality)

Sources

- Machine Learning: $f(\cdot)$ is a regularizer, $Y = \mathbb{R}^n$, each X_i is specified from the data (linear classifiers, X_i is given by a linear inequality)

- Robust LP: $f(x) = \langle c, x \rangle$, $Y = \mathbb{R}^n$, constraints: $\langle a_i, x \rangle \leq b_i$, where a_i and b_i take values in some uncertainty sets A_i and B_i.

- Robust Optimization, Robust System Identification and Control:
  - A. Ben-Tal, A. Nemirovski *Lectures on Modern Convex Optimization*, MPS-SIAM Series on Optimization, SIAM, Philadelphia, 2001
  - A. Ben-Tal, L. El Ghaoui, A. Nemirovski, *Robust Optimization*, Princeton University Press 2009
  - A. Ben-Tal, S. Boyd, A. Nemirovski, Extending Scope of Robust Optimization: Comprehensive Robust Counterparts of Uncertain Problems, Math. Prog. 2006

Projection

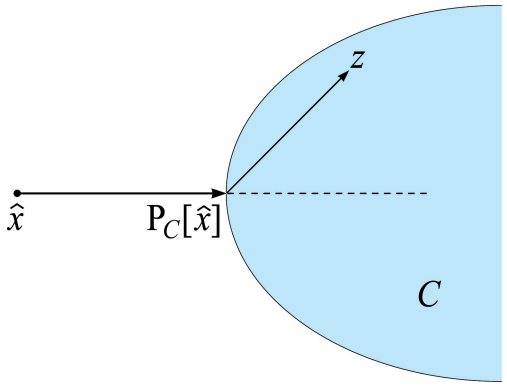
Projection Theorem Let $C \subseteq \mathbb{R}^n$ be a *nonempty closed convex* set and $\widehat{x} \in \mathbb{R}^n$ be arbitrary

(a) There is a unique solution to the following problem

$$\begin{aligned} &\text{minimize} && \|z - \widehat{x}\|^2 \\ &\text{subject to} && z \in C \end{aligned}$$

(b) A vector $z^* \in C$ is the solution if and only if

$$\langle z^* - \widehat{x}, z - z^* \rangle \geq 0 \quad \text{for all } z \in C$$



- The solution is said to be the *projection of* $\widehat{x}$ *on* C, denoted by $\Pi_C[\widehat{x}]$

Projection Properties

Theorem Let $C \subseteq \mathbb{R}^n$ be a *nonempty closed convex set*

(a) The following relation holds

$$\|\Pi_C[x] - z\|^2 + \|x - \Pi_C[x]\|^2 \leq \|x - z\|^2 \qquad \text{for all } x, \in \mathbb{R}^n \text{ and } z \in C.$$

As a consequence, the projection mapping $\Pi_C : \mathbb{R}^n \to C$ is non-expansive, i.e.,

$$\|\Pi_C[x] - \Pi_C[y]\| \leq \|x - y\| \qquad \text{for all } x, y \in \mathbb{R}^n$$

(b) The set distance function $\mathrm{dist}(\cdot, C) : \mathbb{R}^n \to \mathbb{R}$ given by

$$\mathrm{dist}(x, C) = \|\Pi_C[x] - x\| \quad \textbf{is convex on } \mathbb{R}^n$$

(c) The squared distance function

$$\mathrm{dist}^2(x, C) = \|\Pi_C[x] - x\|^2 \qquad \text{for } x \in \mathbb{R}^n$$

is convex and continuously differentiable on $\mathbb{R}^n$

$$\nabla \mathrm{dist}^2(x, C) = 2(x - \Pi_C[x])$$

In fact its gradient mapping is Lipschitz continuous

$$\|\nabla \mathrm{dist}^2(x, C) - \nabla \mathrm{dist}^2(y, C)\| \leq 2\|x - y\| \qquad \text{for } x, y \in \mathbb{R}^n$$

Definition of Sub-gradient and Sub-differential

Def. A vector $\mathbf{s} \in \mathbb{R}^n$ is a **sub-gradient of f at $\widehat{x} \in \mathrm{dom}\, f$** when

$$f(\widehat{x}) + \langle \mathbf{s}, x - \widehat{x} \rangle \leq f(x) \qquad \text{for all } x \in \mathrm{dom}\, f$$

Def. The **sub-differential of f at $\widehat{x} \in \mathrm{dom}\, f$** is the set of all sub-gradients $\mathbf{s}$ of f at $\widehat{x}$

- The **sub-differential set of f at $\widehat{x}$ is denoted by $\partial f(\widehat{x})$**
- When f is differentiable at $\widehat{x}$, we have $\partial f(\widehat{x}) = \{\nabla f(\widehat{x})\}$
- When f is convex and defined on $\mathbb{R}^n$, the sub-differental set $\partial f(\widehat{x})$ is nonempty, convex, and compact set for every $x \in \mathbb{R}^n$
- Examples
  - $f(x) = \|x\|$,

$$\partial f(x) = \frac{x}{\|x\|}, x \neq 0, \quad \partial f(0) = \{\mathbf{s} \in \mathbb{R}^n \mid \|\mathbf{s}\| \leq 1\}$$

  - $f(x) = \max\{g(x), 0\}$ where $g : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable

$$\partial f(x) = \begin{cases} \nabla g(x) & \text{if } g(x) > 0 \\ \mathrm{conv}\{0, \nabla g(x)\} & \text{if } g(x) = 0 \\ 0 & \text{if } g(x) < 0 \end{cases}$$

where $\mathrm{conv}\{0, \nabla g(x)\} = \{\alpha \nabla g(x) \mid \alpha \in [0, 1]\}$

A Single Convex Inequality

- Let $g : \mathbb{R}^n \to \mathbb{R}$ be convex and continuously differentiable
- Let $X = \{x \in \mathbb{R}^n \mid g(x) \leq 0\}$ and assume $X \neq \emptyset$
- Suppose we just want to find a point x^* that is feasible, i.e., $g(x^*) \leq 0$.
- The projection of x on X would satisfy the following inequality (derived from the KKT conditions associated with the corresponding projection problem $\min_{g(v)\leq 0} \|v - x\|^2$)

$$\Pi_X[x] = x - \frac{\langle \nabla g(\Pi_X[x]), x - \Pi_X[x] \rangle}{\|\nabla g(\Pi_X[x])\|^2} \nabla g(\Pi_X[x])$$

- In general, it is hard to project on such a set X
- A method that finds a feasible point for X is based on **Polyak update rule***

$$v^+ = v - \beta \frac{g(v)}{\|\nabla g(v)\|^2} \nabla g(v) \tag{1}$$

where v is not feasible for X ($g(v) > 0$) and $\beta > 0$

- What motivates this rule?

*B.T. Polyak, Minimization of unsmooth functionals, U.S.S.R. Comput. Math. and Math. Phys. 9: 14?29, 1969

- The intuition comes from the linearization of $g(\cdot)$.
  - At the point v, where $g(v) > 0$, we consider the hyperplane
    $$H_v = \{x \in \mathbb{R}^n \mid \langle \nabla g(v), x - v \rangle = 0\}.$$
    Note that $\nabla g(v) \neq 0$ for the method to work, which is satisfied here since v is not feasible (so it cannot be attaining the minimal value of $g(z)$ over $x \in \mathbb{R}^n$).
  - Suppose now we push the hyperplane until it touches the feasible set X at some point, say z, where we have
    $$g(z) = 0$$
  - The equation for this pushed hyperplane is
    $$H = \{x \in \mathbb{R}^n \mid \langle \nabla g(v), x - z \rangle = 0\}$$
    (since it is just a translation of the hyperplane H_v passing through a point z).
  - The projection of v on the hyperplane H is given by
    $$\Pi_H[v] = v - \frac{\langle \nabla g(v), v - z \rangle}{\|\nabla g(v)\|^2} \nabla g(v).$$
  - We cannot compute this projection since we do not know z.

- The value $\langle \nabla g(v), v - z \rangle$ is lower-bounded by using the convexity of $g(\cdot)$, as follows:

$$\langle \nabla g(v), v - z \rangle \geq g(v) - g(z) = g(v)$$

since $g(x^*) = 0$

- So instead of the true projection $\Pi_H[v]$, we have an "approximation"

$$\Pi_H[v] \approx v - \frac{g(v)}{\|\nabla g(v)\|^2} \nabla g(v),$$

which gives the update rule (1) if we include a stepsize β

- Now we focus on the update rule

$$v^+ = v - \beta \frac{g(v)}{\|\nabla g(v)\|^2} \nabla g(v)$$

- Then for any $z \in X$ $(g(z) \leq 0)$,

$$\begin{aligned}
\|v^+ - z\|^2 &= \|v - z\|^2 - 2\beta \frac{g(v)}{\|\nabla g(v)\|^2} \langle \nabla g(v), v - z \rangle + \beta^2 \frac{g^2(v)}{\|\nabla g(v)\|^2} \\
&\leq \|v - z\|^2 - 2\beta \frac{g(v)}{\|\nabla g(v)\|^2} (g(v) - g(z)) + \beta^2 \frac{g^2(v)}{\|\nabla g(v)\|^2} \\
&\leq \|v - z\|^2 - 2\beta \frac{g^2(v)}{\|\nabla g(v)\|^2} + \beta^2 \frac{g^2(v)}{\|\nabla g(v)\|^2} \\
&\|v - z\|^2 - \beta(2 - \beta) \frac{g^2(v)}{\|\nabla g(v)\|^2}
\end{aligned}$$

- Implication: whenever v is not feasible, the new point v^+ is closer to the set X than the point v, if we choose $\beta \in (0, 2)$

- Re-iterating, we can obtain a sequence of infeasible points converging to a point in X

- If at any time it happens that we hit a feasible v^+, the update process stops

- To capture this possibility, write the update rule in a form

$$v^+ = v - \beta \frac{g^+(v)}{\|d\|^2} d$$

where $g^+(v) = \max\{g(v), 0\}$ and

- $d = \nabla g(v)$ if $g(v) > 0$
- d is any nonzero vector if $g(v) \leq 0$. Note that d does not affect the update in this case, since $g^+(v) = 0$ and we get and $v^+ = v$

Set given by Multiple Convex Inequalities

Parallel Batch Processing

- Let $X = \cap_{i=1}^m \{x \in \mathbb{R}^n \mid g_i(x) \leq 0\}$ and assume $X \neq \emptyset$, where each $g_i : \mathbb{R}^n \to \mathbb{R}$ is convex and continuously differentiable

- Assume that the following "error-bound" condition is satisfied: for some $h > 0$,

$$\frac{1}{m}\sum_{i=1}^m g_i^+(x) \geq h \cdot \mathrm{dist}(x, X) \qquad \text{for all } x \in \mathbb{R}^n$$

  NOTE: the condition is satisfied for example if all $g_i(\cdot)$'s are affine function (linear inequalities). The result was established by Hoffman 1952, and known as Hoffman bound

- Let $\{i_1, \ldots, i_N\} \subset [m]$ be a set of indices that will be used in an update step

$$v_{k+1} = v_k - \frac{\beta}{N}\sum_{j=1}^N \frac{g_{i_j}^+(v_k)}{\|\nabla g_{i_j}(v_k)\|^2}\nabla g_{i_j}(v_k)$$

- Similar to the preceding, we can see that for any $z \in X$,

$$\|v_{k+1} - z\|^2 \leq \|v_k - z\|^2 - \frac{\beta(2-\beta)}{N} \sum_{j=1}^{N} \frac{(g_{i_j}^+(v_k))^2}{\|\nabla g_{i_j}(v_k)\|^2}$$

  Hence, $\{v_k\}$ is bounded (deterministically if v_0 is deterministic).

- It follows that $\{\|\nabla g_i(v_k)\|\}$ is bounded for any $i \in [m]$, by some $M_g > 0$.

- Thus, using this bound and taking the infimum over $z \in X$,

$$\mathrm{dist}^2(v_{k+1}, X) \leq \mathrm{dist}^2(v_k, X) - \frac{\beta(2-\beta)}{N} \sum_{j=1}^{N} \frac{(g_{i_j}^+(v_k))^2}{M_g^2}$$

- Suppose each i_j is selected randomly with a uniform distribution on $[m]$

$$\mathbb{E}\left[g_{i_j}^+(x)\right] = \frac{1}{m} \sum_{i=1}^{m} g_i^+(x)$$

- We have

$$\mathbb{E}\left[\mathrm{dist}^2(v_{k+1},X)\mid v_k\right] \le \mathrm{dist}^2(v_k,X) - \frac{\beta(2-\beta)}{mM_g^2}\sum_{i=1}^m (g_i^+(v_k))^2$$
$$\le \mathrm{dist}^2(v_k,X) - \frac{h^2\beta(2-\beta)}{M_g^2}\mathrm{dist}^2(v_k,X)$$

  where the last inequality is obtained using the convexity of $s\mapsto s^2$ and the Hoffman bound to get

$$\frac{1}{m}\sum_{i=1}^m (g_i^+(v_k))^2 \ge \left(\frac{1}{m}\sum_{i=1}^m g_i^+(v_k)\right)^2 \ge h^2\mathrm{dist}(v_k,X)^2$$

- Hence

$$\mathbb{E}\left[\mathrm{dist}^2(v_{k+1},X)\right] \le \left(1-\frac{h^2\beta(2-\beta)}{M_g^2}\right)\mathbb{E}\left[\mathrm{dist}^2(v_k,X)\right]$$

- Expected distance shrinks geometrically in k.

- More careful analysis can capture the effect of the batch size b (the relation among the average and variance instead of the "blue" inequality), resulting in $2-\beta L_N$, with $L_N \le 1$, instead of $2-\beta$ in the last inequality.

Sequential Batch Processing:

- At time k, we have v_k

$$z_1 = v_k \qquad z_{j+1} = z_j - \beta \frac{g_{i_j}^+(z_j)}{\|\nabla g_{i_j}(z_j)\|^2} \nabla g_{i_j}(z_j),\ j = 1, \ldots, N, \qquad v_{k+1} = z_{N+1}$$

- Similar to the preceding we can see that $\{v_k\}$ and the intermittent iterates are bounded and we have for each $j = 1, \ldots, N$

$$\mathbb{E}\left[\mathrm{dist}^2(z_{j+1}, X)\right] \leq \left(1 - \frac{h^2\beta(2-\beta)}{M_g^2}\right) \mathbb{E}\left[\mathrm{dist}^2(z_j, X)\right]$$

implying

$$\mathbb{E}\left[\mathrm{dist}^2(v_{k+1}, X)\right] \leq \left(1 - \frac{h^2\beta(2-\beta)}{M_g^2}\right)^N \mathbb{E}\left[\mathrm{dist}^2(v_k, X)\right]$$

- Compared to the result for parallel batch processing as in the paper[†]

$$\mathbb{E}\left[\mathrm{dist}^2(v_{k+1}, X)\right] \leq \left(1 - \frac{h^2\beta(2-\beta L_N)}{M_g^2}\right) \mathbb{E}\left[\mathrm{dist}^2(v_k, X)\right]$$

with $L_N \leq 1$

- It appears that sequential is better; also easier for deciding on the size N of the batch,
- When $g_i(\cdot)$'s are linear, the Hoffman constant satisfies $h \geq 1$.

[†]AN , I. Necoara Random minibatch projection algorithms for convex problems with functional constraints 2019

Problem of Interest[‡]

- Consider canonical convex minimization problem

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X = \bigcap_{i=1}^{m} X_i \end{array}$$

- Convex $f : \mathbb{R}^n \to \mathbb{R}$, closed convex X_i for all i, and $X \neq \emptyset$.
- Interested in the case when the projection $\Pi_X[x]$ of x on X is not readily available:
  - $\Pi_X[x]$ cannot be given in a closed form
  - The sets $X_i, i = 1, \ldots, m$ may not be known a priori, but are revealed over time
- Problems of this nature arise in
  - Dynamic rate control in communication networks
  - Robust control (feasibility problem)
  - Online learning

[‡]A. N. Random algorithms for convex minimization problems, Math. Program., 2011

Issues

$$\begin{aligned} \text{minimize} \quad & f(x) \\ \text{subject to} \quad & x \in X = \bigcap_{i=1}^{m} X_i \end{aligned}$$

- We focus on **first order methods**
- When the sets X_i are not a priori available: **there is no algorithm**
- When the sets X_i are available, but $\Pi_X[x]$ is "unavailable" (expensive or not possible)
  - **Standard gradient projection algorithm cannot be used**

$$x_k = \Pi_X[x_{k-1} - \alpha_k \nabla f(x_{k-1})]$$

  - Possible alternative (Han and Lou 1988, Han 1989)
    - "Approximate" the projection on X by a sequence of alternate projections

$$\Pi_X[x] \approx \Pi_{X_m}[\cdots \Pi_{X_1}[x] \cdots]$$

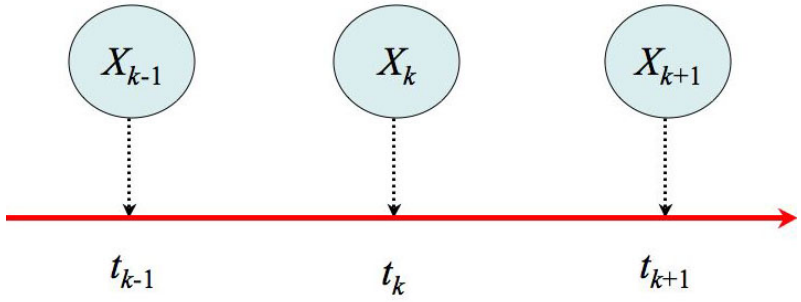
  - Large number of constraints can make this projection "approximation" **practically infeasible**

Proposed Approach: Random Projection Algorithm

- We propose a simple gradient projection algorithm using random projections

$$x_k = \Pi_{\omega_k}[x_{k-1} - \alpha_k \nabla f(x_{k-1})]$$

  - Π_{ω_k} is the projection on the set X_{ω_k}, which is a random realization of the constraint set at time k

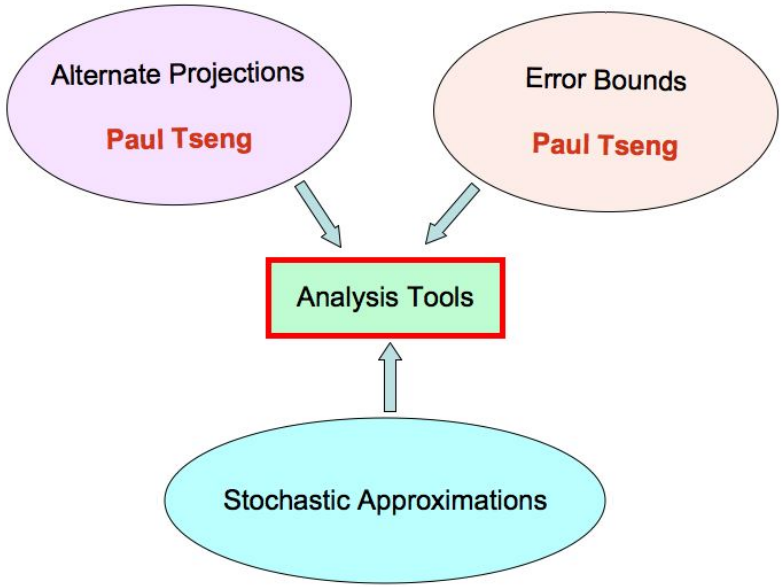


  - $\alpha_k > 0$ is a stepsize
  - The initial point $x_0 \in \mathbb{R}^n$ can be selected at random

- When the collection $\{X_i, i = 1, \ldots, m\}$ is known, we can use uniform distribution over the index set.

Contribution

- Almost sure convergence and convergence rate of the algorithm:
  - Case I: a closed form for the projection on each component X_i is available
  - Case II: each set X_i is given by an algebraic relation
  - In each of the cases, we study the method for *differentiable* and *nondifferentiable* f
  - We analyze all the cases for *diminishing and constant stepsize*
    - Diminishing stepsize: to establish almost sure convergence
    - Constant stepsize: to determine the convergence rate



Related Literature

- *Random Projection Algorithms* for **convex feasibility problems**

$$\text{determine} \quad \bar{x} \in X = \bigcap_{i=1}^{m} X_i$$

  - Amemiya and Ando 1965, Polyak 2001

- *Alternate (or Cyclic) Projection Algorithms*

  - **Projection problem**

    Von Neumann 1950, Aronszajn 1950, Halperin 1962, Han 1988

  - **Feasibility problem**

    Gubin, Polyak and Raik 1967, Deutch 1983, Bauschke and Borwein 1996, Combettes 1997, **Tseng** 1990 (see Bauschke 2001).

    These algorithms solve a feasibility problem by **cyclically projecting** on the sets $X_1, \ldots, X_m$

$$x_k = \Pi_{X_m}[\cdots[\Pi_{X_1}[x_{k-1}]]\cdots]$$

- *Error Bounds* initially studied by Hoffman 1952, and more recently by Hu and Wang 1989, Luo and **Tseng** 1992, 1993, Burke and Ferris 1993, Mangasarian 1996, Burke and **Tseng** 1996, Lewis and Pang 1998, **Tseng** 1999, Bauschke, Borwein, **Tseng** 2000, Bertsekas 1999, Luo 2000 (see also Facchinei and Pang 2003), **Tseng 2009**.

- Error Bounds are related to the question of **determining whether**

$$\text{for a given set} \quad S = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0, i = 1, \ldots, m\}$$

there exists a scalar $\gamma > 0$ such that

$$dist(x, S) \leq \gamma \max_{i \in [m]} g_i^+(x) \qquad \text{for all } x \in \mathbb{R}^n$$

Such a set S is often referred to as **"regular"**.

- Error-bounds are related to "set regularity" and "metric regularity" literature

- The random projection method could be viewed as a counterpart of the random *Incremental Method*.
- **Random Incremental Method solves the following problem**

$$\text{minimize } f(x) = \sum_{i=1}^{m} f_i(x) \quad \text{over } x \in X$$

$$x_k = \Pi_X[x_{k-1} - \alpha_k \nabla f_{\omega_k}(x_{k-1})]$$

where f_{ω_k} is a randomly selected f_i from $f_1, \ldots, f_m$

- Incremental methods have been studied starting with Kibardin 1980 and later on by Luo and **Tseng** 1994, Bertsekas 1997, **Tseng** 1998, Solodov 1998, Bertsekas and Tsitsiklis 2000, Nedić and Bertsekas 2001, Kiwiel 2003, Sundhar et. all 2009, **Tseng** 2009.

Assumptions

Assumption 1 **Convexity** *The sets $X_i \subseteq \mathbb{R}^n$ are closed and convex. The function f is defined and convex over $\mathbb{R}^n$.*

Assumption 2 **IID** *$\{\omega_k\}$ is iid and independent of the initial random point x_0.*

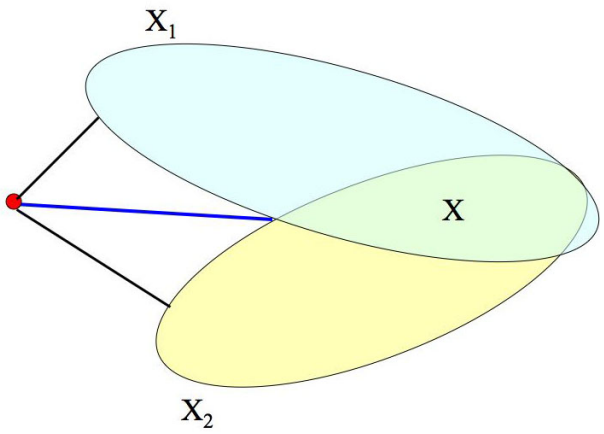
Assumption 3 **Set Regularity** *There exists a constant $c > 0$ such that*

$$\mathrm{dist}^2(x, X) \leq c\, \mathsf{E}\big[\mathrm{dist}^2(x, X_\omega)\big] \qquad \textit{for all } x \in \mathbb{R}^n.$$

The condition is satisfied when the sets are "regular"

- There exists a constant $\gamma > 0$ such that

$$\mathrm{dist}(x, X) \leq \gamma \max_{1 \leq i \leq m} \mathrm{dist}(x, X_i) \qquad \text{for all } x \in \mathbb{R}^n.$$



Random Projection Method for Nondifferentiable Objective

When f is not differentiable, the method is using a subgradient instead of a gradient:

$$x_k = \Pi_{\omega_k}[x_{k-1} - \alpha_k s_f(x_{k-1})] \qquad \text{for all } k \geq 1.$$

The **basic subgradient property**: for any $x \in \mathbb{R}^n$,

$$f(x) + \langle s_f(x), y - x \rangle \leq f(y) \qquad \text{for all } y \in \mathbb{R}^n$$

We assume that **the subgradients of f are uniformly bounded**: there is a scalar C_f,

$$\|s_f(x)\| \leq C_f \qquad \text{for all } s_f(x) \in \partial f(x) \text{ and } x \in \mathbb{R}^n.$$

The subgradient boundedness essentially plays the role of the Lipschitz gradient continuity condition for differentiable $f(\cdot)$

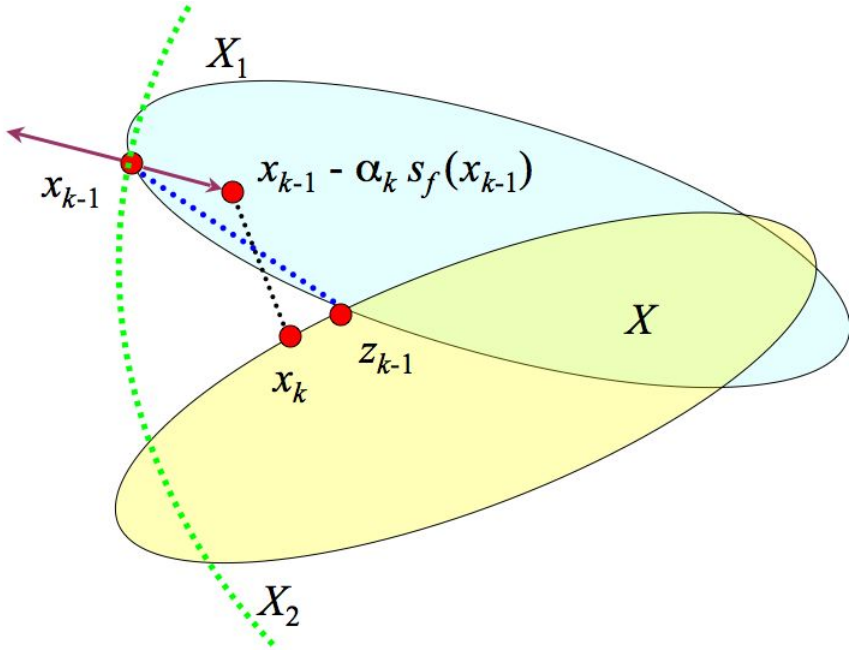
Basic Relation

Let $\mathscr{F}_k$ be the sigma-field generated by x_0 and the realizations of ω_ℓ for $\ell = 1, \ldots, k$.

Lemma: Convexity, IID and Set Regularity assumptions hold.
The subgradients of f are uniformly bounded by C_f, and $\alpha_k > 0$ is arbitary.
Then, for any $\bar{x} \in X$ and $k \geq 1$,

$$\mathsf{E}\left[\|x_k - \bar{x}\|^2 \mid \mathscr{F}_{k-1}\right] \leq \|x_{k-1} - \bar{x}\|^2 - 2\alpha_k \left(f(z_{k-1}) - f(\bar{x})\right) - \frac{1}{2c}\|x_{k-1} - z_{k-1}\|^2 + 4\alpha_k^2(1+c)C_f^2,$$

- $z_{k-1} = \Pi_X[x_{k-1}]$ and c is "regularity constant"



Diminishing Stepsize: Almost Sure Convergence

Proposition 1

Let Convexity, IID and Set Regularity assumptions hold.

Let the subgradients of f be uniformly bounded by C_f.

Let the stepsize be such that $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$.

Assume that problem $\min_{x \in X} f(x)$ has a nonempty optimal set X^*.

Then, the iterates $\{x_k\}$ generated by the method converge almost surely to some random point in the optimal set X^*.

Proof Outline:

Super-martingale convergence result: (Robbins and Siegmund 1971)

Let $\{v_k\}$, $\{u_k\}$, and $\{b_k\}$ be sequences of nonnegative random variables such that $\sum_{k=0}^{\infty} b_k < +\infty$ $a.s.$ and

$$\mathsf{E}[v_{k+1} \mid \mathcal{F}_k] \leq v_k - u_k + b_k \quad \text{for all } k \geq 0 \ \ a.s.,$$

where $\mathcal{F}_k$ denotes the collection $v_0, \ldots, v_k$, $u_0, \ldots, u_k$, $b_0, \ldots, b_k$.

Then, we have $\lim_{k\to\infty} v_k = v$ for a random variable $v \geq 0$ $a.s.$, and $\sum_{k=0}^{\infty} u_k < \infty$ $a.s.$

From our lemma, we have

$$\begin{aligned}\mathsf{E}\left[\|x_k - \bar{x}\|^2 \mid \mathscr{F}_{k-1}\right] \ \leq \ & \|x_{k-1} - \bar{x}\|^2 - 2\alpha_k \left(f(z_{k-1}) - f(\bar{x})\right) \\ & -\frac{1}{2c}\|x_{k-1} - z_{k-1}\|^2 + 4\alpha_k^2(1+c)C_f^2,\end{aligned}$$

Hence it follows:

$$\{\|x_k - \bar{x}\|^2\} \quad \text{is convergent for every } \bar{x} \in X$$

$$\sum_{k=1}^{\infty} \left(2\alpha_k \left(f(z_{k-1}) - f(\bar{x})\right) + \frac{1}{2c}\|x_{k-1} - z_{k-1}\|^2\right) < \infty$$

Rate Analysis

Proposition 2 Let Convexity, IID and Set Regularity assumptions hold.

Let the subgradients of f be uniformly bounded by C_f.

Assume that problem $\min_{x\in X} f(x)$ has a nonempty optimal set X^*.

Let $\{x_k\}$ be the iterate sequence generated by method, and $z_{k-1} = \Pi_X[x_{k-1}]$ for all k.

Define the **weighted averages**

$$\widehat{x}_t = \frac{1}{S_t}\sum_{k=1}^{t} \alpha_k x_{k-1}, \qquad \widehat{z}_t = \frac{1}{S_t}\sum_{k=1}^{t} \alpha_k z_{k-1} \qquad \text{for } t \geq 1$$

$$\text{with} \qquad S_t = \sum_{k=1}^{t} \alpha_k \qquad \text{for } t \geq 1$$

Then:

- If the **stepsize is constant**, i.e., $\alpha_k = \bar{\alpha}$ for all k, then we have the following error bound **per iteration** $t \geq 1$,

$$\mathsf{E}\big[\mathrm{dist}^2(\widehat{x}_t, X)\big] \leq \mathsf{E}\big[\|\widehat{x}_t - \widehat{z}_t\|^2\big] \leq \frac{2c}{t}\mathsf{E}\big[\mathrm{dist}^2(x_0, X^*)\big] + 8c(1+c)C_f^2\bar{\alpha}^2,$$

$$|\mathsf{E}[f(\widehat{x}_t)] - f^*| \leq \frac{1}{2t\bar{\alpha}}\mathsf{E}\big[\mathrm{dist}^2(x_0, X^*)\big] + 2(1+c)C_f^2\bar{\alpha} + C_f\sqrt{\mathsf{E}[\|\widehat{x}_t - \widehat{z}_t\|^2]}$$

with $f^* = \min_{x\in X} f(x)$

$$\mathsf{E}\big[\mathrm{dist}^2(\widehat{x}_t, X)\big] \le \mathsf{E}\big[\|\widehat{x}_t - \widehat{z}_t\|^2\big] \le \frac{2c}{t}\,\mathsf{E}\big[\mathrm{dist}^2(x_0, X^*)\big] + 8c(1+c)C_f^2\bar{\alpha}^2,$$

$$|\mathsf{E}[f(\widehat{x}_t)] - f^*| \le \frac{1}{2t\bar{\alpha}}\,\mathsf{E}\big[\mathrm{dist}^2(x_0, X^*)\big] + 2(1+c)C_f^2\bar{\alpha} + C_f\sqrt{\mathsf{E}[\|\widehat{x}_t - \widehat{z}_t\|^2]}$$

Optimal stepsize for a fixed number of iterations t: (Polyak 2001)

$$\bar{\alpha}_t = \frac{D}{\sqrt{2tB}} \quad \text{for} \quad B = 2C_f^2\left(1 + c + \sqrt{2c(1+c)}\right), \quad D \ge \mathsf{E}\big[\mathrm{dist}^2(x_0, X)\big]$$

$$|\mathsf{E}[f(\widehat{x}_t)] - f^*| \le \frac{\sqrt{2B}}{\sqrt{t}}\,D$$

- If the stepsize satisfies $\lim_{k\to\infty}\alpha_k = \widehat{\alpha} \ge 0$ and $\sum_{k=1}^{\infty}\alpha_k = \infty$, then we have the following **asymptotic error bounds**:

$$\limsup_{t\to\infty}\ \mathsf{E}\big[\|\widehat{x}_t - \widehat{z}_t\|^2\big] \le 8c(1+c)C_f^2\bar{\alpha}\widehat{\alpha},$$

$$\limsup_{t\to\infty}\ |\mathsf{E}[f(\widehat{x}_t)] - f^*| \le 2C_f^2\sqrt{2c(1+c)\bar{\alpha}\widehat{\alpha}} + 2(1+c)C_f^2\widehat{\alpha}.$$

Here $\bar{\alpha} = \max_k \alpha_k$

Constraint Set: Convex Inequalities

$$\begin{array}{ll} \text{minimize} & f(x) \\ \text{subject to} & x \in X, \quad X = \cap_{i=1}^m X_i \\ & X_i = \{x \in \mathbb{R}^n \mid g_i(x) \leq 0\} \qquad \text{for every } i \end{array}$$

where each $g_i : \mathbb{R}^n \to \mathbb{R}$ is a convex function.

The random projection cannot be used

$$x_k = \Pi_{\omega_k}[x_{k-1} - \alpha_k \, s_f(x_{k-1})]$$

- Approximate the projection Π_{ω_k} on X_{ω_k} by a feasibility step striving to reach the set X_i (given by g_i)
- Results in the random algorithm of the following the form:

$$v_k = x_{k-1} - \alpha_k \, s_f(x_{k-1})$$

$$x_k = v_k - \beta \, \frac{g^+_{\omega_k}(v_k)}{\|d_k\|^2} \, d_k \qquad \text{for all } k \geq 1$$

$g^+(x) = \max\{0, g(x)\}$ and $d_k \in \partial g^+_{\omega_k}(v_k)$

$\alpha_k > 0$ is a deterministic stepsize,

and β is a deterministic parameter with $0 < \beta < 2$.

Assumptions

Assumption 4 *The function f and all g_i are defined and convex over $\mathbb{R}^n$. The subgradients $s_{g_i}(x)$ are uniformly bounded, i.e., there is a scalar C_g such that*

$$\|s_{g_i}(x)\| \leq C_g \qquad \text{for all } x \text{ and all } i.$$

In addition, the functions $g_i^+(x)$ satisfy the following condition.

Assumption 5 *There exists a constant $c > 0$ such that*

$$\text{dist}^2(x, X) \leq c\, \mathsf{E}\Big[\big(g_\omega^+(x)\big)^2\Big] \qquad \text{for all } x.$$

This relation is implied by a condition known as *linear metric regularity* (see for example Facchinei and Pang 2003) i.e., when

$$\text{dist}(x, X) \leq \gamma \max_i g_i^+(x) \qquad \text{for all } x \in \mathbb{R}^n.$$

Basic Relation

Lemma 1 Let the IID, and modified Convexity and Set Regularity assumptions hold. Let the subgradient norms of f and all g_i be uniformly bounded by C_f and C_g, resp. Then, for the iterates of the "modified" subgradient method we have for all $\bar{x} \in X$ and $k \geq 1$,

$$\mathsf{E}\left[\|x_k - \bar{x}\|^2 \mid \mathscr{F}_{k-1}\right] \leq \|x_{k-1} - \bar{x}\|^2 - 2\alpha_k \left(f(z_{k-1}) - f(\bar{x})\right)$$

$$-\frac{1}{2}\,\frac{\beta(2-\beta)}{cC_g^2}\,\|x_{k-1} - z_{k-1}\|^2 + D_\beta\,\alpha_k^2,$$

where $z_{k-1} = \Pi_X[x_{k-1}]$ and $D_\beta = 4\beta(2-\beta)C_f^2 + 4\left(1 + \frac{cC_g^2}{\beta(2-\beta)}\right)C_f^2$.

Almost Sure Convergence

Proposition 3 Let the IID, and modified Convexity and Set Regularity assumptions hold. Let the subgradient norms of f be uniformly bounded by C_f.
Let the stepsize be such that $\sum_{k=1}^{\infty} \alpha_k = \infty$ and $\sum_{k=1}^{\infty} \alpha_k^2 < \infty$.
Assume that problem $\min_{x \in X} f(x)$ has a nonempty optimal set X^*. Then, the iterates $\{x_k\}$ generated by the modified method converge almost surely to some random point in the optimal set X^*.

Error Estimate

Proposition 4 Let IID and modified Convexity and Set Regularity assumptions hold.
Let the subgradients of f be uniformly bounded by C_f.
Assume that problem $\min_{x\in X} f(x)$ has a nonempty optimal set X^*.
Let $\{x_k\}$ be the sequence generated by the modified method, and $z_{k-1} = \Pi_X[x_{k-1}]$.
Define the **weighted averages**

$$\widehat{x}_t = \frac{1}{S_t}\sum_{k=1}^{t} \alpha_k x_{k-1}, \qquad \widehat{z}_t = \frac{1}{S_t}\sum_{k=1}^{t} \alpha_k z_{k-1} \qquad \text{for } t \geq 1$$

$$\text{with} \qquad S_t = \sum_{k=1}^{t} \alpha_k \qquad \text{for } t \geq 1$$

- If the stepsize is constant ($\alpha_k = \bar{\alpha}$), the following error bound holds **per iterate** $t \geq 1$,

$$\mathsf{E}\big[\mathrm{dist}^2(\widehat{x}_y, X)\big] \leq \mathsf{E}\big[\|\widehat{x}_t - \widehat{z}_t\|^2\big] \leq \frac{2cC_g^2}{\beta(2-\beta)t}\,\mathsf{E}\big[\mathrm{dist}^2(x_0, X^*)\big] + \frac{2D_\beta\, cC_g^2}{\beta(2-\beta)}\,\bar{\alpha}^2$$

$$|\mathsf{E}[f(\widehat{x}_t)] - f^*| \leq \frac{1}{2t\bar{\alpha}}\,\mathsf{E}\big[\mathrm{dist}^2(x_0, X^*)\big] + \frac{D_\beta}{2}\,\bar{\alpha} + C_f\sqrt{\mathsf{E}[\|\widehat{x}_t - \widehat{z}_t\|^2]}$$

where $f^* = \min_{x\in X} f(x)$ and $D_\beta = 4\beta(2-\beta)C_f^2 + 4\left(1 + \frac{cC_g^2}{\beta(2-\beta)}\right)C_f^2$.

- If the stepsize satisfies $\lim_{k\to\infty}\alpha_k = \widehat{\alpha} \geq 0$ and $\sum_{k=1}^{\infty}\alpha_k = \infty$, then we have the following **asymptotic error bounds**:

$$\limsup_{t\to\infty} \mathsf{E}\left[\|\widehat{x}_t - \widehat{z}_t\|^2\right] \leq \frac{2D_\beta c C_g^2}{\beta(2-\beta)}\,\bar{\alpha}\widehat{\alpha},$$

$$\limsup_{t\to\infty}\ |\mathsf{E}[f(\widehat{x}_t)] - f^*| \leq C_f C_g \sqrt{\frac{2cD_\beta}{\beta(2-\beta)}\,\bar{\alpha}\widehat{\alpha}} + \frac{D_\beta}{2}\widehat{\alpha}.$$

Here $\bar{\alpha} = \max_k \alpha_k$

In summary: convergence rate for a diminishing stepsize is $O(1/\sqrt{t})$ in both the case when we can project on each X_i and when we cannot project on the sets X_i (i.e., when we use Polyak update rule to reduce the infeasibility)

The rate is not for the last iterate x_t but for the weighted-averages of the iterates (weighted with the stepsize values).