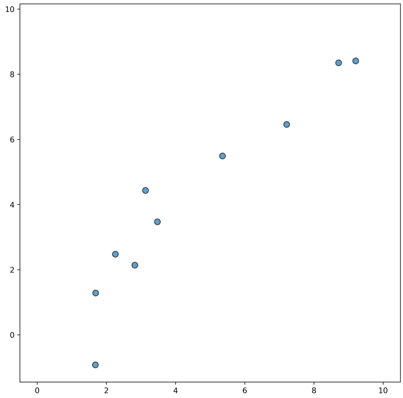
Deep Learning

Lukas Schnelle

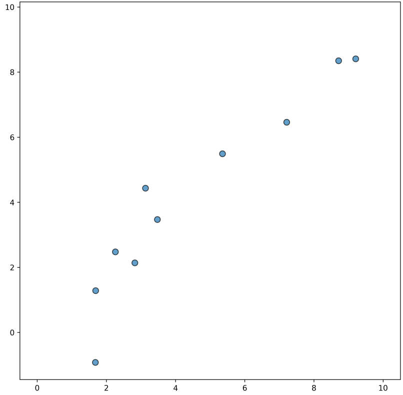
June 2023

An Example



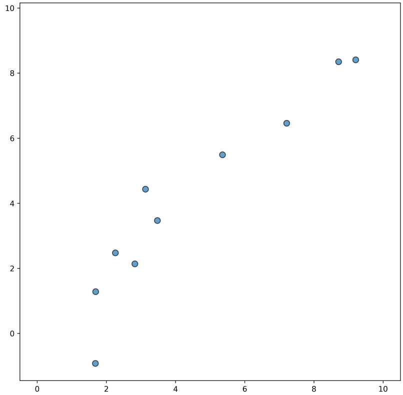
An Example

Let $\forall i \in [n] : a_i \in \mathbb{R}^1, b_i \in \mathbb{R}$ some data.
Goal: Find $x \in \mathbb{R}$ s.th. $ax = -b$



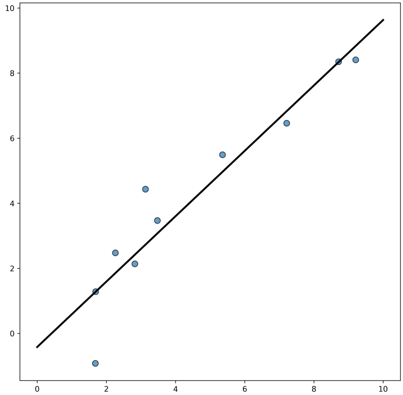
An Example

Let $\forall i \in [n] : a_i \in \mathbb{R}^1, b_i \in \mathbb{R}$ some data.
Goal: Find $x \in \mathbb{R}$ s.th. $min_{x \in \mathbb{R}}(ax - b)$



An Example

Let $\forall i \in [n] : a_i \in \mathbb{R}^1, b_i \in \mathbb{R}$ some data.
Goal: Find $x \in \mathbb{R}$ s.th. $min_{x \in \mathbb{R}}(ax - b)$



Problem Statement

Let $\forall i \in [n] : a_i \in \mathbb{R}^d, b_i \in \mathbb{R}$. Then find:

$$\min_{x \in \mathbb{R}^d} \frac{1}{n} \sum_{i=0}^{n} (a_i^T \cdot x - b_i)^2$$

Problem Statement

Let $\forall i \in [n] : a_i \in \mathbb{R}^d, b_i \in \mathbb{R}$. Then find:

$$\min_{x \in \mathbb{R}^d} \underbrace{\frac{1}{n} \sum_{i=0}^{n} (a_i^T \cdot x - b_i)^2}_{=:g(x)}$$

Problem Statement

Let $\forall i \in [n] : a_i \in \mathbb{R}^d, b_i \in \mathbb{R}$. Then find:

$$\min_{x \in \mathbb{R}^d} \underbrace{\frac{1}{n} \sum_{i=0}^{n} \overbrace{(a_i^T \cdot x - b_i)^2}^{=:f_i(x)}}_{=:g(x)}$$

(Full) Gradient Descent

How to solve such problem?

(Full) Gradient Descent

How to solve such problem?

Idea

Guess x, try to improve: go "down"

(Full) Gradient Descent

How to solve such problem?

Idea

Guess x, try to improve: go "down"
Formally: Guess x^k, set $x^{k+1} = x^k - \nabla g(x^k)$

Called "Gradient Descent" or "Full Gradient"

Properties of Full Gradient

Notations

In the following let x^* the optimal value,

Properties of Full Gradient

Notations

In the following let x^* the optimal value,

$$a := \begin{pmatrix} -a_1^T- \\ \vdots \\ -a_n^T- \end{pmatrix},$$

Properties of Full Gradient

Notations

In the following let x^* the optimal value,

$$a := \begin{pmatrix} -a_1^T- \\ \vdots \\ -a_n^T- \end{pmatrix}, b := \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

Properties of Full Gradient

Notations

In the following let x^* the optimal value,

$$a := \begin{pmatrix} -a_1^T- \\ \vdots \\ -a_n^T- \end{pmatrix}, b := \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

$$g(x) = \sum_{i=1}^{n} \left(a_i^T x - b_i\right)^2 \qquad = \frac{1}{n}\|ax - b\|_2^2$$

Properties of Full Gradient

Notations

In the following let x^* the optimal value,

$$a := \begin{pmatrix} -a_1^T- \\ \vdots \\ -a_n^T- \end{pmatrix}, b := \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

$$g(x) = \sum_{i=1}^{n} \left(a_i^T x - b_i\right)^2 \qquad = \frac{1}{n}\|ax - b\|_2^2$$

$$\nabla g(x) = \frac{1}{n}\sum_{i=1}^{n} a_i \left(a_i^T x - b_i\right) = \frac{1}{n} a^T(ax - b)$$

Properties of Full Gradient

Notations

In the following let x^* the optimal value,

$$a := \begin{pmatrix} -a_1^T- \\ \vdots \\ -a_n^T- \end{pmatrix}, b := \begin{pmatrix} b_1 \\ \vdots \\ b_n \end{pmatrix}$$

$$g(x) = \sum_{i=1}^{n} \left(a_i^T x - b_i\right)^2 \qquad = \frac{1}{n}\|ax - b\|_2^2$$

$$\nabla g(x) = \frac{1}{n}\sum_{i=1}^{n} a_i \left(a_i^T x - b_i\right) = \frac{1}{n} a^T(ax - b)$$

$$\nabla^2 g(x) = \frac{1}{n}\sum_{i=1}^{n} a_i a_i^T \qquad = \frac{1}{n} a^T a$$

L-smoothness

Definition (L-smoothness)

Let $f : \mathbb{R}^d \to \mathbb{R}$ a function, $x_1, x_2 \in \mathbb{R}^d$.Then f is called L-smooth if

L-smoothness

Definition (L-smoothness)

Let $f : \mathbb{R}^d \to \mathbb{R}$ a function, $x_1, x_2 \in \mathbb{R}^d$.Then f is called L-smooth if

$$\exists L : \|\nabla f(x_1) - \nabla f(x_2)\|_2^2 \leq L\|x_1 - x_2\|_2^2$$

L-smoothness

Definition (L-smoothness)

Let $f : \mathbb{R}^d \to \mathbb{R}$ a function, $x_1, x_2 \in \mathbb{R}^d$.Then f is called L-smooth if

$$\exists L : \|\nabla f(x_1) - \nabla f(x_2)\|_2^2 \leq L\|x_1 - x_2\|_2^2$$

Proposition

The function g from our problem is L-smooth.

Proof.

Let $x_1, x_2 \in \mathbb{R}^d$,

Proof.

Let $x_1, x_2 \in \mathbb{R}^d$, $B := \max_{i \in [n]} \|a_i\|_2$. Then

$$\|\nabla g(x_1) - \nabla g(x_2)\| =$$

$\square$

Proof.

Let $x_1, x_2 \in \mathbb{R}^d$, $B := \max_{i\in[n]} \|a_i\|_2$. Then

$$\|\nabla g(x_1) - \nabla g(x_2)\| = \|\sum_{i=1}^{n} a_i a_i^T (x_1 - x_2)\|_2$$

$$=$$

$\square$

Proof.

Let $x_1, x_2 \in \mathbb{R}^d$, $B := \max_{i\in[n]} \|a_i\|_2$. Then

$$\begin{aligned}
\|\nabla g(x_1) - \nabla g(x_2)\| &= \|\sum_{i=1}^{n} a_i a_i^T (x_1 - x_2)\|_2 \\
&= \|a^T a(x_1 - x_2)\|_2 \\
&\leq
\end{aligned}$$

□

Proof.

Let $x_1, x_2 \in \mathbb{R}^d$, $B := \max_{i \in [n]} \|a_i\|_2$. Then

$$
\begin{aligned}
\|\nabla g(x_1) - \nabla g(x_2)\| &= \|\sum_{i=1}^{n} a_i a_i^T (x_1 - x_2)\|_2 \\
&= \|a^T a(x_1 - x_2)\|_2 \\
&\leq B^2 n \|(x_1 - x_2)\|_2
\end{aligned}
$$

$\square$

Proof.

Let $x_1, x_2 \in \mathbb{R}^d$, $B := \max_{i\in[n]} \|a_i\|_2$. Then

$$
\begin{aligned}
\|\nabla g(x_1) - \nabla g(x_2)\| &= \|\sum_{i=1}^{n} a_i a_i^T (x_1 - x_2)\|_2 \\
&= \|a^T a(x_1 - x_2)\|_2 \\
&\leq B^2 n \|(x_1 - x_2)\|_2
\end{aligned}
$$

$\implies g$ is L-smooth with $L := B^2 n$ $\square$

Proof.

Let $x_1, x_2 \in \mathbb{R}^d$, $B := \max_{i\in[n]} \|a_i\|_2$. Then

$$\begin{aligned}\|\nabla g(x_1) - \nabla g(x_2)\| &= \|\sum_{i=1}^{n} a_i a_i^T (x_1 - x_2)\|_2 \\ &= \|a^T a(x_1 - x_2)\|_2 \\ &\leq B^2 n\|(x_1 - x_2)\|_2\end{aligned}$$

$\implies$ g is L-smooth with $L := B^2 n$ $\square$

Corollary

For L from before it holds that

$$L = n\left(\max_{i\in[n]} \|a_i\|_2\right)^2$$

Proof.

Let $x_1, x_2 \in \mathbb{R}^d$, $B := \max_{i\in[n]} \|a_i\|_2$. Then

$$\begin{aligned}\|\nabla g(x_1) - \nabla g(x_2)\| &= \|\sum_{i=1}^{n} a_i a_i^T (x_1 - x_2)\|_2 \\ &= \|a^T a(x_1 - x_2)\|_2 \\ &\leq B^2 n\|(x_1 - x_2)\|_2\end{aligned}$$

$\implies$ g is L-smooth with $L := B^2 n$ □

Corollary

For L from before it holds that

$$L = n\left(\max_{i\in[n]} \|a_i\|_2\right)^2$$

Proof.

Let $x_1, x_2 \in \mathbb{R}^d$, $B := \max_{i \in [n]} \|a_i\|_2$. Then

$$\begin{aligned}
\|\nabla g(x_1) - \nabla g(x_2)\| &= \|\sum_{i=1}^{n} a_i a_i^T (x_1 - x_2)\|_2 \\
&= \|a^T a(x_1 - x_2)\|_2 \\
&\leq B^2 n \|(x_1 - x_2)\|_2
\end{aligned}$$

$\implies$ g is L-smooth with $L := B^2 n$ $\square$

Corollary

For L from before it holds that

$$L = n \left(\max_{i \in [n]} \|a_i\|_2 \right)^2 \geq \sigma_{max}(a^T a)$$

Properties

Definition (Convexity)

Let $f : \mathbb{R}^d \to \mathbb{R}$ a function, $x_1, x_2 \in \mathbb{R}^d$.Then

Properties

Definition (Convexity)

Let $f : \mathbb{R}^d \to \mathbb{R}$ a function, $x_1, x_2 \in \mathbb{R}^d$.Then

(i) f is called convex if $\forall 0 \leq t \leq 1$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

Properties

Definition (Convexity)

Let $f : \mathbb{R}^d \to \mathbb{R}$ a function, $x_1, x_2 \in \mathbb{R}^d$. Then

(i) f is called convex if $\forall 0 \leq t \leq 1$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

$$\iff f(x_1) \geq f(x_2) + (\nabla f(x_2))^T (x_1 - x_2)$$

Properties

Definition (Convexity)

Let $f : \mathbb{R}^d \to \mathbb{R}$ a function, $x_1, x_2 \in \mathbb{R}^d$. Then

(i) f is called convex if $\forall 0 \leq t \leq 1$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$

$$\iff f(x_1) \geq f(x_2) + (\nabla f(x_2))^T (x_1 - x_2)$$

(ii) f is called **strongly** convex if

$$\exists \mu > 0 : f(x_1) \geq$$

Properties

Definition (Convexity)

Let $f : \mathbb{R}^d \to \mathbb{R}$ a function, $x_1, x_2 \in \mathbb{R}^d$.Then

(i) f is called convex if $\forall 0 \leq t \leq 1$

$$f(tx_1 + (1-t)x_2) \leq tf(x_1) + (1-t)f(x_2)$$
$$\iff f(x_1) \geq f(x_2) + (\nabla f(x_2))^T (x_1 - x_2)$$

(ii) f is called **strongly** convex if

$$\exists \mu > 0 : f(x_1) \geq f(x_2) + (\nabla f(x_2))^T (x_1 - x_2) + \frac{\mu}{2} \|x_1 - x_2\|_2^2$$

Example: Blackboard

Local to Global

Why these Properties?

Local to Global

Why these Properties?

- Convex $\implies$ solution exists and is unique

Local to Global

Why these Properties?

- Convex $\implies$ solution exists and is unique
- Strongly convex $\implies$ small $g(x')$ is close to $g(x^*)$ implies x' close to x^*

Theorem

The function g from our problem is convex.

Theorem

The function g from our problem is convex.
If $\sigma_{min}(a^T a) > 0$ it is even strongly convex

Proof.

$$\overset{\text{Taylor}}{\Longrightarrow} g(x') = g(x) + (\nabla g(x))^T (x' - x) + \frac{1}{2}(x' - x)^T a^T a(x' - x)$$

Proof.

$$\overset{\text{Taylor}}{\Longrightarrow} g(x') = g(x) + (\nabla g(x))^T (x' - x) + \frac{1}{2}(x' - x)^T a^T a (x' - x)$$

$$\geq g(x) + (\nabla g(x))^T (x' - x) + \frac{1}{2}\sigma_{min}(a^T a)\|x' - x\|_2^2$$

Proof.

$$\overset{\text{Taylor}}{\Longrightarrow} g(x') = g(x) + (\nabla g(x))^T(x'-x) + \frac{1}{2}(x'-x)^T a^T a(x'-x)$$

$$\geq g(x) + (\nabla g(x))^T(x'-x) + \frac{1}{2}\sigma_{min}(a^T a)\|x'-x\|_2^2$$

$$\Longrightarrow g(x) + (\nabla g(x))^T(x'-x) \leq g(x') - \frac{1}{2}\sigma_{min}(a^T a)\|x'-x\|_2^2$$

Proof.

$$\overset{\text{Taylor}}{\Longrightarrow} g(x') = g(x) + (\nabla g(x))^T (x' - x) + \frac{1}{2}(x' - x)^T a^T a (x' - x)$$

$$\geq g(x) + (\nabla g(x))^T (x' - x) + \frac{1}{2}\sigma_{min}(a^T a)\|x' - x\|_2^2$$

$$\Longrightarrow g(x) + (\nabla g(x))^T (x' - x) \leq g(x') - \frac{1}{2}\sigma_{min}(a^T a)\|x' - x\|_2^2$$

$\Longrightarrow$ g is strongly convex, if $\mu := \sigma_{min}(a^T a) > 0$ $\square$

Convergence rate

Theorem

The FG method has for decreasing step size convergence rate of

$$g(x^k) - g(x^*) = \mathcal{O}(1/k)$$

and

Convergence rate

Theorem

The FG method has for decreasing step size convergence rate of

$$g(x^k) - g(x^*) = \mathcal{O}(1/k)$$

and

$$\exists \rho \in (0,1) : g(x^k) - g(x^*) = \mathcal{O}(\rho^k)$$

if g is strongly convex (i.e. $\sigma_{min}(a^T a) > 0$

Proof of convergence of FG

For the following, let η_k the step size of every step, and

$$x_{k+1} := x_k - \eta_k \nabla g(x_k)$$

the update.

Proof of convergence of FG

For the following, let η_k the step size of every step, and

$$x_{k+1} := x_k - \eta_k \nabla g(x_k)$$

the update.
Here: only strongly convex case

$$\|x^{k+1} - x^*\|_2^2 = \langle x^{k+1} - x^*, x^{k+1} - x^* \rangle$$

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \langle x^{k+1} - x^*, x^{k+1} - x^* \rangle \\
=& \langle x^k - \eta_k \nabla g(x^k) - x^*, x^k - \eta_k \nabla g(x^k) - x^* \rangle
\end{aligned}
$$

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \langle x^{k+1} - x^*, x^{k+1} - x^* \rangle \\
=& \langle x^k - \eta_k \nabla g(x^k) - x^*, x^k - \eta_k \nabla g(x^k) - x^* \rangle \\
=& \langle x^k - x^*, x^k - \eta_k \nabla g(x^k) - x^* \rangle \\
& - \langle \eta_k \nabla g(x^k), x^k - \eta_k \nabla g(x^k) - x^* \rangle
\end{aligned}
$$

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \langle x^{k+1} - x^*, x^{k+1} - x^* \rangle \\
=& \langle x^k - \eta_k \nabla g(x^k) - x^*, x^k - \eta_k \nabla g(x^k) - x^* \rangle \\
=& \langle x^k - x^*, x^k - \eta_k \nabla g(x^k) - x^* \rangle \\
& - \langle \eta_k \nabla g(x^k), x^k - \eta_k \nabla g(x^k) - x^* \rangle \\
=& \langle x^k - x^*, x^k - x^* \rangle \\
& - \langle \eta_k \nabla g(x^k), x^k - x^* \rangle \\
& - \langle x^k - x^*, \eta_k \nabla g(x^k) \rangle \\
& + \langle \eta_k \nabla g(x^k), \eta_k \nabla g(x^k) \rangle
\end{aligned}
$$

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 =& \langle x^{k+1} - x^*, x^{k+1} - x^* \rangle \\
=& \langle x^k - \eta_k \nabla g(x^k) - x^*, x^k - \eta_k \nabla g(x^k) - x^* \rangle \\
=& \langle x^k - x^*, x^k - \eta_k \nabla g(x^k) - x^* \rangle \\
& - \langle \eta_k \nabla g(x^k), x^k - \eta_k \nabla g(x^k) - x^* \rangle \\
=& \langle x^k - x^*, x^k - x^* \rangle \\
& - \langle \eta_k \nabla g(x^k), x^k - x^* \rangle \\
& - \langle x^k - x^*, \eta_k \nabla g(x^k) \rangle \\
& + \langle \eta_k \nabla g(x^k), \eta_k \nabla g(x^k) \rangle \\
=& \|x^k - x^*\|_2^2 - 2\eta_k (\nabla g(x^k))^T (x^k - x^*) + \eta_k^2 \|\nabla g(x^k)\|_2^2
\end{aligned}
$$

Let $\delta_k := g(x^*) - g(x^k)$

$\|x^{k+1} - x^*\|_2^2 =$

Let $\delta_k := g(x^*) - g(x^k)$

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - x^*\|_2^2 - 2\eta_k(\nabla g(x^k))^T(x^k - x^*)) + \eta_k^2\|\nabla g(x^k)\|_2^2$$

Let $\delta_k := g(x^*) - g(x^k)$

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - x^*\|_2^2 - 2\eta_k(\nabla g(x^k))^T(x^k - x^*)) + \eta_k^2\|\nabla g(x^k)\|_2^2$$

$$\leq \|x^k - x^*\|_2^2 - 2\eta_k(g(x^*) - g(x^k) - \frac{\mu}{2}\|x^k - x^*\|)$$

$$+ \eta_k^2\|\nabla g(x^k)\|_2^2$$

Let $\delta_k := g(x^*) - g(x^k)$

$$\begin{aligned}
\|x^{k+1} - x^*\|_2^2 = &\|x^k - x^*\|_2^2 - 2\eta_k(\nabla g(x^k))^T(x^k - x^*)) + \eta_k^2\|\nabla g(x^k)\|_2^2 \\
\leq &\|x^k - x^*\|_2^2 - 2\eta_k(g(x^*) - g(x^k) - \frac{\mu}{2}\|x^k - x^*\|) \\
&+ \eta_k^2\|\nabla g(x^k)\|_2^2 \\
= &\|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k)\|_2^2
\end{aligned}$$

Let $\delta_k := g(x^*) - g(x^k)$

$$\begin{aligned}
\|x^{k+1} - x^*\|_2^2 = {} & \|x^k - x^*\|_2^2 - 2\eta_k(\nabla g(x^k))^T(x^k - x^*)) + \eta_k^2\|\nabla g(x^k)\|_2^2 \\
\leq & \|x^k - x^*\|_2^2 - 2\eta_k(g(x^*) - g(x^k) - \frac{\mu}{2}\|x^k - x^*\|) \\
& + \eta_k^2\|\nabla g(x^k)\|_2^2 \\
= & \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k)\|_2^2 \\
= & \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k) - \nabla g(x^*)\|_2^2
\end{aligned}$$

Let $\delta_k := g(x^*) - g(x^k)$

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 = {} & \|x^k - x^*\|_2^2 - 2\eta_k(\nabla g(x^k))^T(x^k - x^*)) + \eta_k^2\|\nabla g(x^k)\|_2^2 \\
\leq & \|x^k - x^*\|_2^2 - 2\eta_k(g(x^*) - g(x^k) - \frac{\mu}{2}\|x^k - x^*\|) \\
& + \eta_k^2\|\nabla g(x^k)\|_2^2 \\
= & \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k)\|_2^2 \\
= & \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k) - \nabla g(x^*)\|_2^2 \\
\leq & \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2 L^2\|x^k - x^*\|_2^2
\end{aligned}
$$

Let $\delta_k := g(x^*) - g(x^k)$

$$\begin{aligned}
\|x^{k+1} - x^*\|_2^2 = \|x^k - x^*\|_2^2 - 2\eta_k(\nabla g(x^k))^T(x^k - x^*)) + \eta_k^2\|\nabla g(x^k)\|_2^2 \\
\leq \|x^k - x^*\|_2^2 - 2\eta_k(g(x^*) - g(x^k) - \frac{\mu}{2}\|x^k - x^*\|) \\
+ \eta_k^2\|\nabla g(x^k)\|_2^2 \\
= \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k)\|_2^2 \\
= \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k) - \nabla g(x^*)\|_2^2 \\
\leq \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2 L^2\|x^k - x^*\|_2^2
\end{aligned}$$

Let $\eta_k \in (0, \min(\frac{\mu}{2L^2}, \frac{2}{\mu}))$

Let $\delta_k := g(x^*) - g(x^k)$

$$
\begin{aligned}
\|x^{k+1} - x^*\|_2^2 = {} & \|x^k - x^*\|_2^2 - 2\eta_k(\nabla g(x^k))^T(x^k - x^*)) + \eta_k^2\|\nabla g(x^k)\|_2^2 \\
\leq & \|x^k - x^*\|_2^2 - 2\eta_k(g(x^*) - g(x^k) - \frac{\mu}{2}\|x^k - x^*\|) \\
& + \eta_k^2\|\nabla g(x^k)\|_2^2 \\
= & \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k)\|_2^2 \\
= & \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k) - \nabla g(x^*)\|_2^2 \\
\leq & \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2 L^2\|x^k - x^*\|_2^2
\end{aligned}
$$

Let $\eta_k \in (0, \min(\frac{\mu}{2L^2}, \frac{2}{\mu}))$

$\implies \exists \rho \in (0,1) : \|x^k - x^*\|_2^2 \leq \rho^k\|x^0 - x^*\|_2^2$

Let $\delta_k := g(x^*) - g(x^k)$

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - x^*\|_2^2 - 2\eta_k(\nabla g(x^k))^T(x^k - x^*)) + \eta_k^2\|\nabla g(x^k)\|_2^2$$

$$\le \|x^k - x^*\|_2^2 - 2\eta_k(g(x^*) - g(x^k) - \frac{\mu}{2}\|x^k - x^*\|)$$

$$+ \eta_k^2\|\nabla g(x^k)\|_2^2$$

$$= \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k)\|_2^2$$

$$= \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k) - \nabla g(x^*)\|_2^2$$

$$\le \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2 L^2\|x^k - x^*\|_2^2$$

Let $\eta_k \in (0, \min(\frac{\mu}{2L^2}, \frac{2}{\mu}))$

$\implies \exists\rho \in (0,1) : \|x^k - x^*\|_2^2 \le \rho^k\|x^0 - x^*\|_2^2$

$\implies g(x^*) - g(x^k) \le \frac{1}{2}\rho^k\|x^0 - x^*\|_2^2(\mu + \eta_k L^2) - \rho^{k+1}\frac{\|x^0 - x^*\|_2^2}{2\eta_k}$

Let $\delta_k := g(x^*) - g(x^k)$

$$\|x^{k+1} - x^*\|_2^2 = \|x^k - x^*\|_2^2 - 2\eta_k(\nabla g(x^k))^T(x^k - x^*)) + \eta_k^2\|\nabla g(x^k)\|_2^2$$

$$\leq \|x^k - x^*\|_2^2 - 2\eta_k(g(x^*) - g(x^k) - \frac{\mu}{2}\|x^k - x^*\|)$$

$$+ \eta_k^2\|\nabla g(x^k)\|_2^2$$

$$= \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k)\|_2^2$$

$$= \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2\|\nabla g(x^k) - \nabla g(x^*)\|_2^2$$

$$\leq \|x^k - x^*\|_2^2(1 - \eta_k\mu) - 2\eta_k\delta_k + \eta_k^2 L^2\|x^k - x^*\|_2^2$$

Let $\eta_k \in (0, \min(\frac{\mu}{2L^2}, \frac{2}{\mu}))$

$\implies \exists\rho \in (0,1) : \|x^k - x^*\|_2^2 \leq \rho^k\|x^0 - x^*\|_2^2$

$\implies g(x^*) - g(x^k) \leq \frac{1}{2}\rho^k\|x^0 - x^*\|_2^2(\mu + \eta_k L^2) - \rho^{k+1}\frac{\|x^0 - x^*\|_2^2}{2\eta_k}$

$\rightarrow$Lyapunov function that controls convergence

Where is the problem?

Idea (FG)

Guess x, try to improve: go "down"
Formally: Guess x^k, set $x^{k+1} = x^k - \eta_k \nabla g(x)$

Where is the problem?

Idea (FG)

Guess x, try to improve: go "down"
Formally: Guess x^k, set $x^{k+1} = x^k - \eta_k \nabla g(x)$

But: $\nabla g(x) \in \mathbb{R}^d$

Where is the problem?

Idea (FG)

Guess x, try to improve: go "down"
Formally: Guess x^k, set $x^{k+1} = x^k - \eta_k \nabla g(x)$

But: $\nabla g(x) \in \mathbb{R}^d$
And: $[\nabla g(x)]_k = \frac{1}{n} \sum_{i=0}^{n} \nabla f_i(x)$

A new challenger

Problem

$[\nabla g(x)]_k = \frac{1}{n} \sum_{i=0}^{n} \nabla f_i(x)$

Calculating gradient of **all** samples

A new challenger

Problem

$[\nabla g(x)]_k = \frac{1}{n}\sum_{i=0}^{n} \nabla f_i(x)$

Calculating gradient of **all** samples

Ansatz

Don't calculate the gradient for every sample. Go into "direction" of **one** sample

A new challenger

Problem

$[\nabla g(x)]_k = \frac{1}{n} \sum_{i=0}^{n} \nabla f_i(x)$
Calculating gradient of **all** samples

Ansatz

Don't calculate the gradient for every sample. Go into "direction" of **one** sample

Formally

$x^{k+1} = x^k - \eta_k \nabla f_i(x)$

A new challenger

Problem

$[\nabla g(x)]_k = \frac{1}{n}\sum_{i=0}^{n} \nabla f_i(x)$
Calculating gradient of **all** samples

Ansatz

Don't calculate the gradient for every sample. Go into "direction" of **one** sample

Formally

$x^{k+1} = x^k - \eta_k \nabla f_i(x)$

But which i?

Which i

1. Idea

Just count

Which i

1. Idea

Just count $\rightarrow$ Samples could be ordered

Which i

1. Idea

Just count $\rightarrow$ Samples could be ordered

2. Idea

Choose random one

Which i

1. Idea

Just count $\rightarrow$ Samples could be ordered

2. Idea

Choose random one
Called "Stochastic gradient"

Convergence

Theorem

For k iterations, decreasing step size, g convex:

$$\mathbb{E}[g(x^k)] - g(x^*) = \mathcal{O}(1/\sqrt{k})$$

and

Convergence

Theorem

For k iterations, decreasing step size, g convex:

$$\mathbb{E}[g(x^k)] - g(x^*) = \mathcal{O}(1/\sqrt{k})$$

and

$$\mathbb{E}[g(x^k)] - g(x^*) = \mathcal{O}(1/k)$$

if g strongly-convex

Convergence

Theorem

For k iterations, decreasing step size, g convex:

$$\mathbb{E}[g(x^k)] - g(x^*) = \mathcal{O}(1/\sqrt{k})$$

and

$$\mathbb{E}[g(x^k)] - g(x^*) = \mathcal{O}(1/k)$$

if g strongly-convex

Can be shown similar as for FG.

Comparison to FG

	convex	strongly convex
FG		
SG		

Comparison to FG

	convex	strongly convex
FG	$\mathcal{O}(1/k)$	$\mathcal{O}(\rho^k)$
SG		

Comparison to FG

	convex	strongly convex
FG	$\mathcal{O}(1/k)$	$\mathcal{O}(\rho^k)$
SG	$\mathcal{O}(1/\sqrt{k})$	$\mathcal{O}(1/k)$

For SG the convergence is in $\mathbb{E}$

What about the other data?

SG Method does only consider **one** sample.

Question: how to consider more without depending on the sample size?

What about the other data?

SG Method does only consider **one** sample.

Question: how to consider more without depending on the sample size?

Idea

Take an average

What about the other data?

SG Method does only consider **one** sample.
Question: how to consider more without depending on the sample size?

Idea

Take an average
Formally:

$$x^{k+1} := x^k - \frac{\alpha_k}{n} \sum_{i=0}^{n} y_i^k$$

with

What about the other data?

SG Method does only consider **one** sample.
Question: how to consider more without depending on the sample size?

Idea

Take an average
Formally:

$$x^{k+1} := x^k - \frac{\alpha_k}{n} \sum_{i=0}^{n} y_i^k$$

with

$$y_i^k := \begin{cases} \nabla f_i(x^k) & i = i_k \\ y_i^{k-1} & else \end{cases}$$

where i_k is the random variable from the SG method.

Convergence

Theorem 1

Let L the Lipschitz constant of g, step size $\alpha_k := \frac{1}{16L}$. The SAG method has convergence rate of:

Convergence

Theorem 1

Let L the Lipschitz constant of g, step size $\alpha_k := \frac{1}{16L}$. The SAG method has convergence rate of:

$$\mathbb{E}[g(x^k)] - g(x^*) \leq \frac{32n}{k} C_0$$

with

Convergence

Theorem 1

Let L the Lipschitz constant of g, step size $\alpha_k := \frac{1}{16L}$. The SAG method has convergence rate of:

$$\mathbb{E}[g(x^k)] - g(x^*) \leq \frac{32n}{k} C_0$$

with

$$C_0 = \frac{3}{2}\left(g(x^0) - g(x^*)\right) + \frac{4L}{n}\|x^0 - x^*\|^2$$

for $y_i^0 = f_i'(x^0) - g'(x^0)$ and

Convergence

Theorem 1

Let L the Lipschitz constant of g, step size $\alpha_k := \frac{1}{16L}$. The SAG method has convergence rate of:

$$\mathbb{E}[g(x^k)] - g(x^*) \leq \frac{32n}{k} C_0$$

with

$$C_0 = \frac{3}{2}\left(g(x^0) - g(x^*)\right) + \frac{4L}{n}\|x^0 - x^*\|^2$$

for $y_i^0 = f_i'(x^0) - g'(x^0)$ and

$$\mathbb{E}[g(x^k)] - g(x^*) \leq \left(1 - min\left\{\frac{\mu}{16L}, \frac{1}{8n}\right\}\right)^k C_0$$

if g strongly convex w.r.t. μ.

Sketch of a proof

Goal: Want to find Lyapunov function again, s.th. it controls the convergence.

Sketch of a proof

Goal: Want to find Lyapunov function again, s.th. it controls the convergence.

$$\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$$

s.th. $\mathcal{L}(x) \geq g(x^*) - g(x^k)$.

Sketch of a proof

Goal: Want to find Lyapunov function again, s.th. it controls the convergence.

$$\mathcal{L} : \mathbb{R}^d \to \mathbb{R}$$

s.th. $\mathcal{L}(x) \geq g(x^*) - g(x^k)$. For this, let

$$y^k := \begin{pmatrix} y_1^k \\ \vdots \\ y_n^k \end{pmatrix}, \theta^k := \begin{pmatrix} y_1^k \\ \vdots \\ y_n^k \\ x^k \end{pmatrix}, \theta^* := \begin{pmatrix} f_1'(x^*) \\ \vdots \\ f_n'(x^*) \\ x^* \end{pmatrix}$$

and

$$e := \begin{pmatrix} I \\ \vdots \\ I \end{pmatrix}, f'(x) := \begin{pmatrix} f_1'(x) \\ \vdots \\ f_n'(x) \end{pmatrix}$$

Here the Lyapunov function has the following form:

$$\mathcal{L}(\theta^k) = 2hg(x^k + de^T y^k) - 2hg(x^*) + (\theta^k - \theta^*)^T \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} (\theta^k - \theta^*)$$

Here the Lyapunov function has the following form:

$$\mathcal{L}(\theta^k) = 2hg(x^k + de^T y^k) - 2hg(x^*) + (\theta^k - \theta^*)^T \begin{pmatrix} A & B \\ B^T & C \end{pmatrix} (\theta^k - \theta^*)$$

In the paper: solved by identifying coefficients, finding a valid ones with CAS.

Term	Scalar s	Matrix M	Source	Group
$sg(x^*)$	$-2h$		$\mathbb{E}\big(\mathcal{L}(\theta^k)\|\mathcal{F}_{k-1}\big)$	0
$sg(x^*)$	$2h(1-\delta)$		$L(\theta^{k-1})$	0
$s(y^{k-1}-f'(x^*))^\top M(y^{k-1}-f'(x^*))$	$-(1-\delta)$	$a_1ee^\top + a_2I$	$L(\theta^{k-1})$	3 and 4
$s(y^{k-1}-f'(x^*))^\top M(x^{k-1}-x^*)$	$-2(1-\delta)$	be	$L(\theta^{k-1})$	5
$s(x^{k-1}-x^*)\top M(x^{k-1}-x^*)$	$-(1-\delta)$	cI	$L(\theta^{k-1})$	9
$sg(x^{k-1})$	$-2h(1-\delta)$		Inequality (12)	0
$sg'(x^{k-1})^\top My^{k-1}$	$-2h(1-\delta)d$	$e^\top$	Inequality (12)	7
$s(y^{k-1})^\top My^{k-1}$	$-h(1-\delta)\mu d^2$	$ee^\top$	Inequality (12)	4
$sg(x^{k-1})$	$2h$		Inequality (13)	0
$sg'(x^{k-1})^\top My^{k-1}$	$2h(d-\frac{\alpha}{n})$	$(1-\frac{1}{n})e^\top$	Inequality (13)	7
$sg'(x^{k-1})^\top Mf'(x^{k-1})$	$2h(d-\frac{\alpha}{n})$	$\frac{1}{n}e^\top$	Inequality (13)	2
$s(y^{k-1}-f'(x^*))^\top M(y^{k-1}-f'(x^*))$	$Lh(d-\frac{\alpha}{n})^2$	$\left(1-\frac{2}{n}\right)ee^\top + \frac{1}{n}I$	Inequality (13)	3 and 4
$s(f'(x^{k-1})-f'(x^*))^\top M(f'(x^{k-1})-f'(x^*))$	$\frac{Lh}{n}(d-\frac{\alpha}{n})^2$	I	Inequality (13)	8
$s(y^{k-1}-f'(x^*))^\top M(f'(x^{k-1})-f'(x^*))$	$\frac{2Lh}{n}(d-\frac{\alpha}{n})^2$	$ee^\top - I$	Inequality (13)	6 and 7
$s(y^{k-1}-f'(x^*))^\top M(y^{k-1}-f'(x^*))$	1	$\left(1-\frac{2}{n}\right)S + \frac{1}{n}\mathrm{Diag}(\mathrm{diag}(S))$	Lemma 1	3 and 4
$s(y^{k-1}-f'(x^*))^\top M(x^{k-1}-x^*)$	$2\left(1-\frac{1}{n}\right)$	$(b-\frac{\alpha}{n}c)e$	Lemma 1	5
$s(x^{k-1}-x^*)\top M(x^{k-1}-x^*)$	1	cI	Lemma 1	9
$s(f'(x^{k-1})-f'(x^*))^\top M(f'(x^{k-1})-f'(x^*))$	$\frac{1}{n}$	$\mathrm{Diag}(\mathrm{diag}(S))$	Lemma 1	8
$s(y^{k-1}-f'(x^*))^\top M(f'(x^{k-1})-f'(x^*))$	$\frac{2}{n}$	$S-\mathrm{Diag}(\mathrm{diag}(S))$	Lemma 1	6 and 7
$s(f'(x^{k-1})-f'(x^*))^\top M(x^{k-1}-x^*)$	$\frac{2}{n}$	$(b-\frac{\alpha}{n}c)e$	Lemma 1	1

Table 3: Expressions in upper bound on $\mathbb{E}\big(\mathcal{L}(\theta^k)|\mathcal{F}_{k-1}\big) - (1-\delta)\mathcal{L}(\theta^{k-1})$.

with

$$
\begin{aligned}
B_0 &= 2\delta h \\
B_1 &= 2(b - \frac{\alpha}{n}c) \\
B_2 &= 2(\frac{\alpha}{n} - d)h \\
B_3 &= -\Big[\Big(1 - \frac{2}{n}\Big)a_2 + \frac{1}{n}[a_1 + a_2 - 2\frac{\alpha}{n}b + \frac{\alpha^2}{n^2}c] - (1-\delta)a_2 + Lh\frac{1}{n}(d - \frac{\alpha}{n})^2\Big] \\
B_4 &= B_3 \\
&- n\Big[\Big(1 - \frac{2}{n}\Big)(a_1 - 2\frac{\alpha}{n}b + \frac{\alpha^2}{n^2}c) - (1-\delta)a_1 + L(1 - \frac{2}{n})h(d - \frac{\alpha}{n})^2 - (1-\delta)\mu h d^2\Big] \\
B_5 &= 2\Big[(\delta - \frac{1}{n})b - \frac{\alpha}{n}(1 - \frac{1}{n})c\Big] \\
B_6 &= -\frac{2}{n}\Big(hL(d - \frac{\alpha}{n})^2 + a_1 - \frac{2\alpha}{n}b + \frac{\alpha^2}{n^2}c\Big) \\
B_7 &= \Big(\Big[2\Big(hL(d - \frac{\alpha}{n})^2 + a_1 - \frac{2\alpha}{n}b + \frac{\alpha^2}{n^2}c\Big) + 2\Big(h(d - \frac{\alpha}{n})(1 - \frac{1}{n}) - h(1-\delta)d\Big)\Big]\Big) \\
B_8 &= \Big[\frac{1}{n}(a_1 + a_2 - 2\frac{\alpha}{n}b + \frac{\alpha^2}{n^2}c) + \frac{L}{n}h(d - \frac{\alpha}{n})^2\Big] \\
B_9 &= c\delta.
\end{aligned}
$$

$$
\begin{aligned}
a_1 &= \frac{1}{32nL}(1 - \frac{1}{2n}) \\
a_2 &= \frac{1}{16nL}(1 - \frac{1}{2n}) \\
b &= -\frac{1}{4n}(1 - \frac{1}{n}) \\
c &= \frac{4L}{n} \\
h &= \frac{1}{2} - \frac{1}{n} \\
d &= \frac{\alpha}{n} \\
\alpha &= \frac{1}{16L} \\
\delta &= \min\left(\frac{1}{8n}, \frac{\mu}{16L}\right) \\
\gamma &= 1 \\
C_3 &= \frac{1}{32n}.
\end{aligned}
$$

A lemma

Lemma (2)

Let $I \in \mathbb{R}^{p \times p}$ *identity matrix,* $e = \begin{pmatrix} I \\ \vdots \\ I \end{pmatrix} \in \mathbb{R}^{np \times p}$, $\alpha, \beta \in \mathbb{R} \setminus \{0\}$

Then

A lemma

Lemma (2)

Let $I \in \mathbb{R}^{p \times p}$ *identity matrix,* $e = \begin{pmatrix} I \\ \vdots \\ I \end{pmatrix} \in \mathbb{R}^{np \times p}$, $\alpha, \beta \in \mathbb{R} \setminus \{0\}$

Then

$$\left(\alpha\left(I - \frac{1}{n}ee^T\right) + \beta\left(\frac{1}{n}ee^T\right)\right)^{-1} = \frac{1}{\alpha}\left(I - \frac{1}{n}ee^T\right) + \frac{1}{\beta}\left(\frac{1}{n}ee^T\right)$$

A lemma

Lemma (2)

Let $I \in \mathbb{R}^{p \times p}$ *identity matrix,* $e = \begin{pmatrix} I \\ \vdots \\ I \end{pmatrix} \in \mathbb{R}^{np \times p}$, $\alpha, \beta \in \mathbb{R} \setminus \{0\}$

Then

$$\left(\alpha\left(I - \frac{1}{n}ee^T\right) + \beta\left(\frac{1}{n}ee^T\right)\right)^{-1} = \frac{1}{\alpha}\left(I - \frac{1}{n}ee^T\right) + \frac{1}{\beta}\left(\frac{1}{n}ee^T\right)$$

Proof.

First, notice that $e^T e = nI$.

A lemma

Lemma (2)

Let $I \in \mathbb{R}^{p\times p}$ *identity matrix,* $e = \begin{pmatrix} I \\ \vdots \\ I \end{pmatrix} \in \mathbb{R}^{np\times p}$, $\alpha, \beta \in \mathbb{R} \setminus \{0\}$

Then

$$\left(\alpha\left(I - \frac{1}{n}ee^T\right) + \beta\left(\frac{1}{n}ee^T\right)\right)^{-1} = \frac{1}{\alpha}\left(I - \frac{1}{n}ee^T\right) + \frac{1}{\beta}\left(\frac{1}{n}ee^T\right)$$

Proof.

First, notice that $e^Te = nI$. Therefore we get
$\left(\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right) =$

A lemma

Lemma (2)

Let $I \in \mathbb{R}^{p \times p}$ *identity matrix,* $e = \begin{pmatrix} I \\ \vdots \\ I \end{pmatrix} \in \mathbb{R}^{np \times p}$, $\alpha, \beta \in \mathbb{R} \setminus \{0\}$

Then

$$\left(\alpha \left(I - \frac{1}{n} ee^T\right) + \beta \left(\frac{1}{n} ee^T\right)\right)^{-1} = \frac{1}{\alpha}\left(I - \frac{1}{n} ee^T\right) + \frac{1}{\beta}\left(\frac{1}{n} ee^T\right)$$

Proof.

First, notice that $e^T e = nI$. Therefore we get
$\left(\frac{1}{n} ee^T\right)\left(\frac{1}{n} ee^T\right) = \frac{1}{n^2} e \underbrace{e^T e}_{=nI} e^T = \frac{1}{n} ee^T$ □

A Lemma

$$\left(\alpha\left(I-\frac{1}{n}ee^T\right)+\beta\left(\frac{1}{n}ee^T\right)\right)\cdot\left(\frac{1}{\alpha}\left(I-\frac{1}{n}ee^T\right)+\frac{1}{\beta}\left(\frac{1}{n}ee^T\right)\right)$$

A Lemma

$$
\left(\alpha\left(I-\frac{1}{n}ee^T\right)+\beta\left(\frac{1}{n}ee^T\right)\right)\cdot\left(\frac{1}{\alpha}\left(I-\frac{1}{n}ee^T\right)+\frac{1}{\beta}\left(\frac{1}{n}ee^T\right)\right)
$$

$$
=\left(I-\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+
$$

A Lemma

$$\left(\alpha\left(I-\frac{1}{n}ee^T\right)+\beta\left(\frac{1}{n}ee^T\right)\right)\cdot\left(\frac{1}{\alpha}\left(I-\frac{1}{n}ee^T\right)+\frac{1}{\beta}\left(\frac{1}{n}ee^T\right)\right)$$

$$=\left(I-\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+\frac{\alpha}{\beta}\left(I-\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right)$$

$$+$$

A Lemma

$$\left(\alpha\left(I-\frac{1}{n}ee^T\right)+\beta\left(\frac{1}{n}ee^T\right)\right)\cdot\left(\frac{1}{\alpha}\left(I-\frac{1}{n}ee^T\right)+\frac{1}{\beta}\left(\frac{1}{n}ee^T\right)\right)$$

$$=\left(I-\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+\frac{\alpha}{\beta}\left(I-\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right)$$

$$+\frac{\beta}{\alpha}\left(\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+$$

A Lemma

$$\left(\alpha\left(I-\frac{1}{n}ee^T\right)+\beta\left(\frac{1}{n}ee^T\right)\right)\cdot\left(\frac{1}{\alpha}\left(I-\frac{1}{n}ee^T\right)+\frac{1}{\beta}\left(\frac{1}{n}ee^T\right)\right)$$

$$=\left(I-\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+\frac{\alpha}{\beta}\left(I-\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right)$$

$$+\frac{\beta}{\alpha}\left(\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+\left(\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right)$$

$$=$$

A Lemma

$$\left(\alpha\left(I-\frac{1}{n}ee^T\right)+\beta\left(\frac{1}{n}ee^T\right)\right)\cdot\left(\frac{1}{\alpha}\left(I-\frac{1}{n}ee^T\right)+\frac{1}{\beta}\left(\frac{1}{n}ee^T\right)\right)$$

$$=\left(I-\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+\frac{\alpha}{\beta}\left(I-\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right)$$

$$+\frac{\beta}{\alpha}\left(\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+\left(\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right)$$

$$=\left(I-\frac{2}{n}ee^T+\frac{1}{n}ee^T\right)+$$

A Lemma

$$
\begin{aligned}
&\left(\alpha\left(I-\frac{1}{n}ee^T\right)+\beta\left(\frac{1}{n}ee^T\right)\right)\cdot\left(\frac{1}{\alpha}\left(I-\frac{1}{n}ee^T\right)+\frac{1}{\beta}\left(\frac{1}{n}ee^T\right)\right)\\
&=\left(I-\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+\frac{\alpha}{\beta}\left(I-\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right)\\
&\quad+\frac{\beta}{\alpha}\left(\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+\left(\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right)\\
&=\left(I-\frac{2}{n}ee^T+\frac{1}{n}ee^T\right)+\frac{\alpha}{\beta}\left(\frac{1}{n}ee^T-\frac{1}{n}ee^T\right)\\
&\quad+
\end{aligned}
$$

A Lemma

$$\left(\alpha\left(I-\frac{1}{n}ee^T\right)+\beta\left(\frac{1}{n}ee^T\right)\right)\cdot\left(\frac{1}{\alpha}\left(I-\frac{1}{n}ee^T\right)+\frac{1}{\beta}\left(\frac{1}{n}ee^T\right)\right)$$

$$=\left(I-\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+\frac{\alpha}{\beta}\left(I-\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right)$$

$$+\frac{\beta}{\alpha}\left(\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+\left(\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right)$$

$$=\left(I-\frac{2}{n}ee^T+\frac{1}{n}ee^T\right)+\frac{\alpha}{\beta}\left(\frac{1}{n}ee^T-\frac{1}{n}ee^T\right)$$

$$+\frac{\beta}{\alpha}\left(\frac{1}{n}ee^T-\frac{1}{n}ee^T\right)+$$

A Lemma

$$
\begin{aligned}
&\left(\alpha\left(I-\frac{1}{n}ee^T\right)+\beta\left(\frac{1}{n}ee^T\right)\right)\cdot\left(\frac{1}{\alpha}\left(I-\frac{1}{n}ee^T\right)+\frac{1}{\beta}\left(\frac{1}{n}ee^T\right)\right)\\
=&\left(I-\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+\frac{\alpha}{\beta}\left(I-\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right)\\
&+\frac{\beta}{\alpha}\left(\frac{1}{n}ee^T\right)\left(I-\frac{1}{n}ee^T\right)+\left(\frac{1}{n}ee^T\right)\left(\frac{1}{n}ee^T\right)\\
=&\left(I-\frac{2}{n}ee^T+\frac{1}{n}ee^T\right)+\frac{\alpha}{\beta}\left(\frac{1}{n}ee^T-\frac{1}{n}ee^T\right)\\
&+\frac{\beta}{\alpha}\left(\frac{1}{n}ee^T-\frac{1}{n}ee^T\right)+\frac{1}{n}ee^T\\
=&I \quad \square
\end{aligned}
$$

Comparison to FG

	convex	strongly convex
FG	$\mathcal{O}(1/k)$	$\mathcal{O}(\rho^k)$
SG	$\mathcal{O}(1/\sqrt{k})$	$\mathcal{O}(1/k)$
SAG		

Comparison to FG

	convex	strongly convex
FG	$\mathcal{O}(1/k)$	$\mathcal{O}(\rho^k)$
SG	$\mathcal{O}(1/\sqrt{k})$	$\mathcal{O}(1/k)$
SAG	$\mathcal{O}(1/k)$	$\mathcal{O}(\rho^k)$

For SG and SAG the convergence is in $\mathbb{E}$

Basic Algorithm

The basic Algorithm just updates every time with

$$x^{k+1} = x^k - \frac{\alpha_k}{n} \sum_{i=1}^{n} y_i^k$$

Basic Algorithm

The basic Algorithm just updates every time with

$$x^{k+1} = x^k - \frac{\alpha_k}{n} \sum_{i=1}^{n} y_i^k$$

Potential improvements are:

- Just in time parameter updates

Basic Algorithm

The basic Algorithm just updates every time with

$$x^{k+1} = x^k - \frac{\alpha_k}{n} \sum_{i=1}^{n} y_i^k$$

Potential improvements are:

- Just in time parameter updates
- Adding weigths to already seen samples

Basic Algorithm

The basic Algorithm just updates every time with

$$x^{k+1} = x^k - \frac{\alpha_k}{n} \sum_{i=1}^{n} y_i^k$$

Potential improvements are:

- Just in time parameter updates
- Adding weigths to already seen samples
- Warm starting

Basic Algorithm

The basic Algorithm just updates every time with

$$x^{k+1} = x^k - \frac{\alpha_k}{n} \sum_{i=1}^{n} y_i^k$$

Potential improvements are:

- Just in time parameter updates
- Adding weigths to already seen samples
- Warm starting
- Line-searching L

Basic Algorithm

The basic Algorithm just updates every time with

$$x^{k+1} = x^k - \frac{\alpha_k}{n} \sum_{i=1}^{n} y_i^k$$

Potential improvements are:

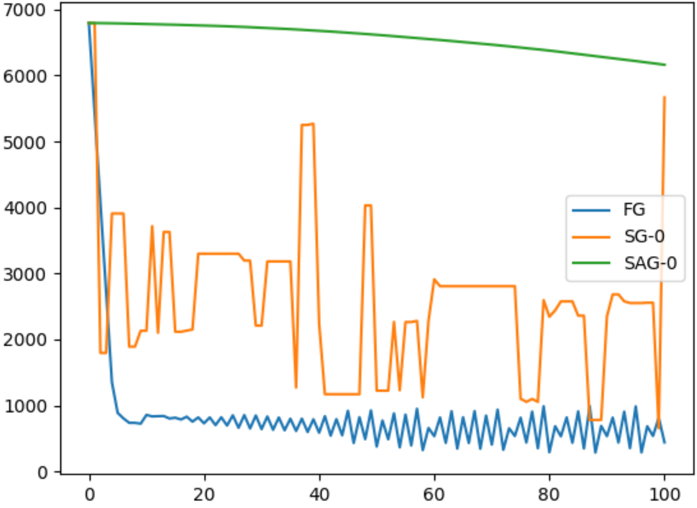
- Just in time parameter updates
- Adding weigths to already seen samples
- Warm starting
- Line-searching L
- Mini-batches

Experimental results

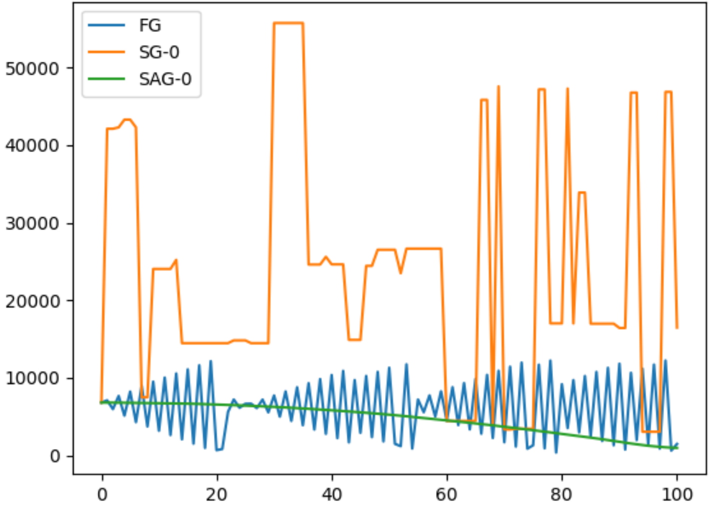
Dataset

KDD Cup 2004
50.000 samples, 78 features/dimensions

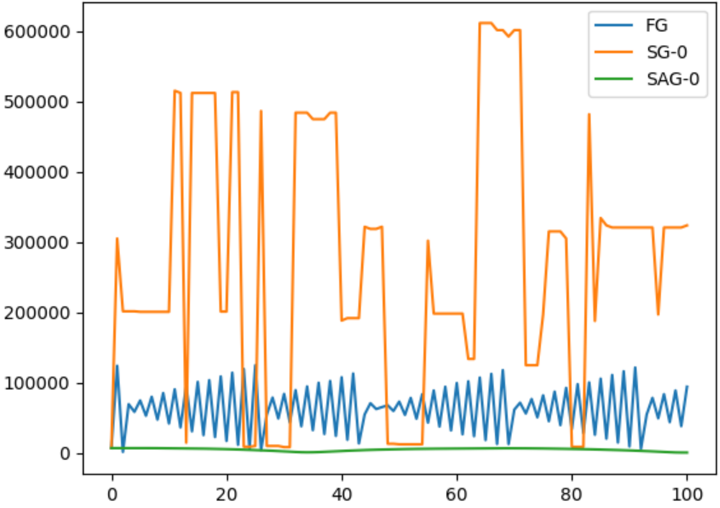
Logistic regression, with same learning rate for all algorithms



With $\eta_k = 1/10.000$



With $\eta_k = 1/1.000$



With $\eta_k = 1/100$

Big thanks to Adrian Gallus and Jonas Nießen for helping with the simulations.

Thank you for your attention

Are there questions?