BOSTON UNIVERSITY

Lecture II: Accelerated Gradient Descent

ALINA ENE

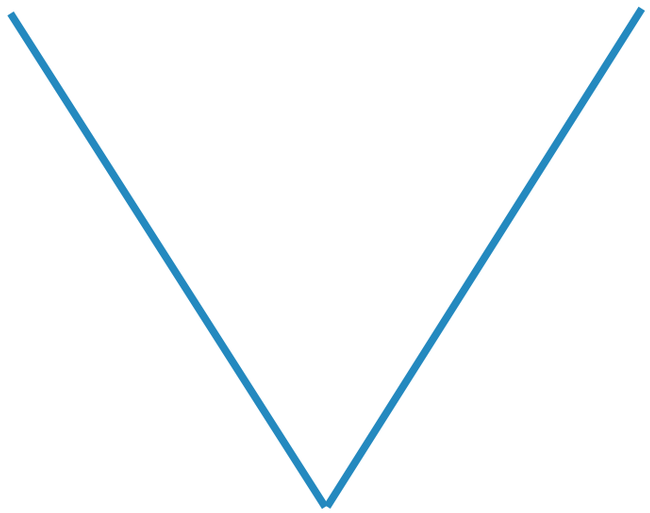
ADFOCS ’21: Convex Optimization and Graph Algorithms

The Unreasonable Effectiveness of Adagrad

It automatically adapts to problem structure

non-smooth

$$\|\nabla f(x)\| \leq G$$

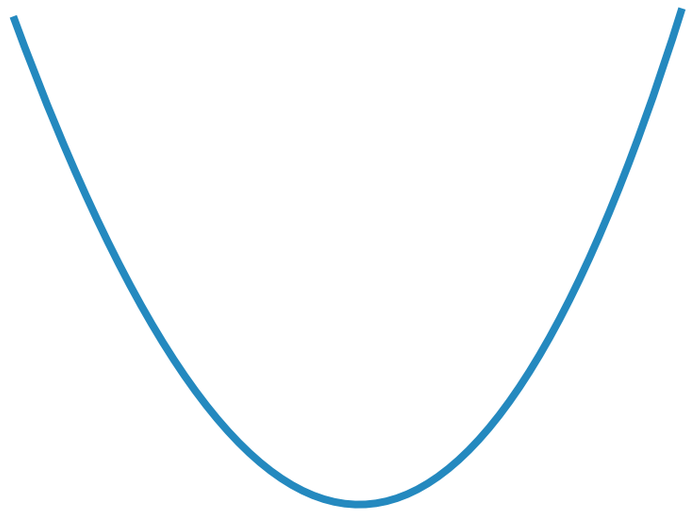


$$T = \frac{G^2 R^2}{\epsilon^2}$$

optimal

smooth

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$



$$T = \frac{\beta R^2}{\epsilon}$$

not optimal

$$R = \max_{t \in [T]} \|x_t - x^*\|$$

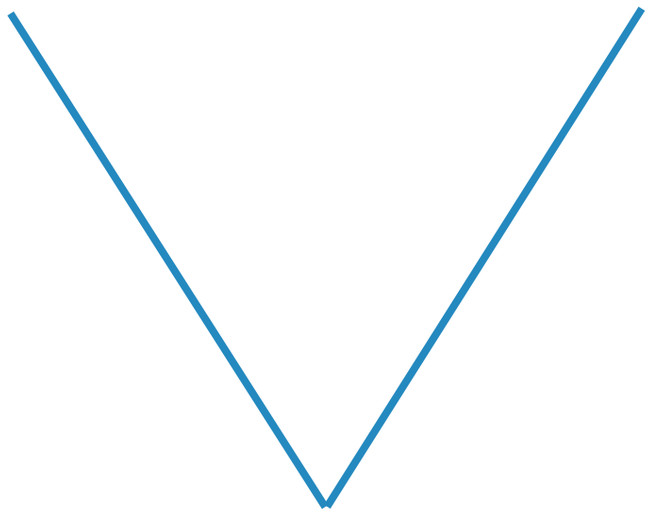
The Unreasonable Effectiveness of Adagrad

AutoML

It automatically adapts to problem structure

non-smooth

$$\|\nabla f(x)\| \leq G$$



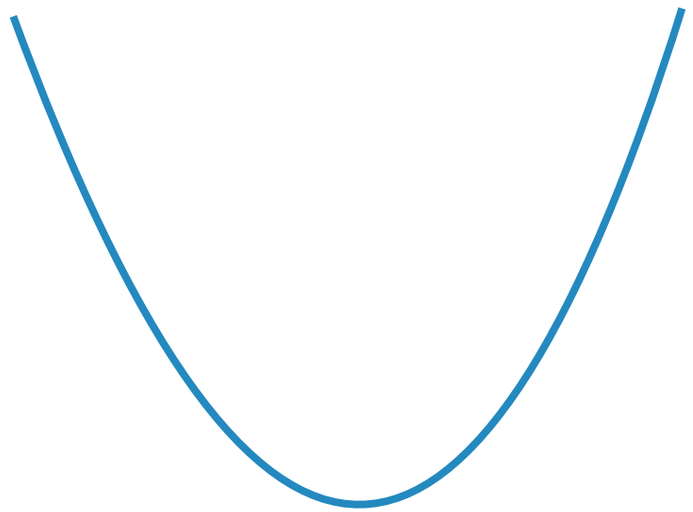
$$R = \max_{t \in [T]} \|x_t - x^*\|$$

$$T = \frac{G^2 R^2}{\epsilon^2}$$

optimal

smooth

$$\|\nabla f(x) - \nabla f(y)\| \leq \beta \|x - y\|$$



$$T = \frac{\beta R^2}{\epsilon}$$

not optimal

$$T = \frac{\beta R^2}{\sqrt{\epsilon}}$$

acceleration

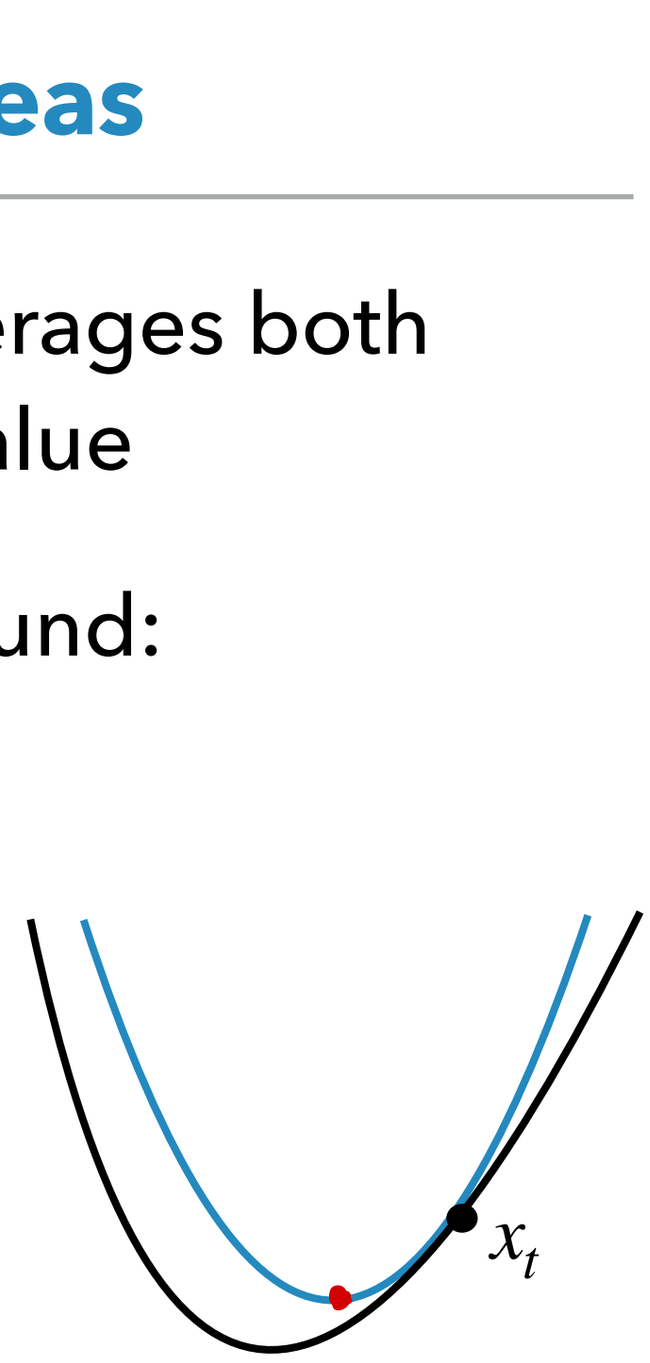
Gradient Descent: Main Ideas

- Gradient descent for smooth functions leverages both upper and lower bounds on the function value

Gradient Descent: Main Ideas

- Gradient descent for smooth functions leverages both upper and lower bounds on the function value
- Smoothness gives us a quadratic upper bound:

$$f(x) \leq \underbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{\beta}{2}\|x - x_t\|^2}_{\text{quadratic upper obound on } f}$$



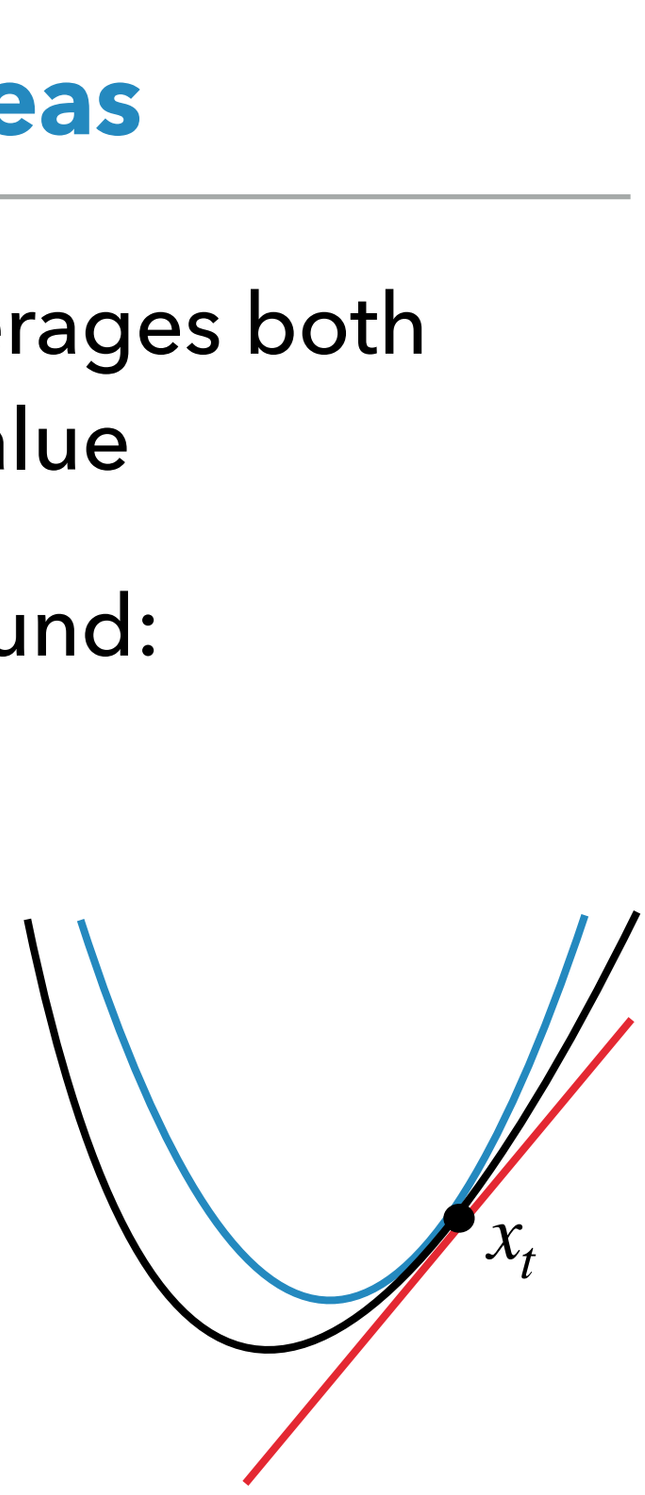
Gradient Descent: Main Ideas

- Gradient descent for smooth functions leverages both upper and lower bounds on the function value
- Smoothness gives us a quadratic upper bound:

$$f(x) \leq \overbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{\beta}{2}\|x - x_t\|^2}^{\text{quadratic upper obound on } f}$$

- Convexity gives us an affine lower bound:

$$f(x) \geq \overbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle}^{\text{affine lower bound on } f}$$



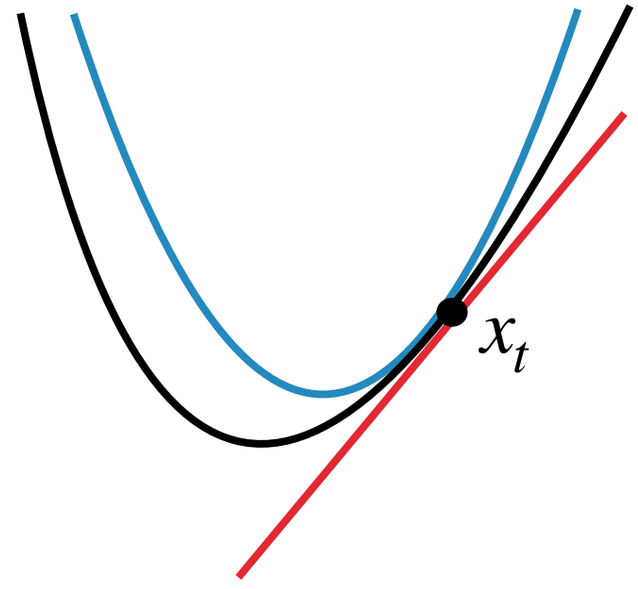
Gradient Descent: Main Ideas

- Gradient descent for smooth functions leverages both upper and lower bounds on the function value
- Smoothness gives us a quadratic upper bound:

$$f(x) \leq \overbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{\beta}{2}\|x - x_t\|^2}^{\text{quadratic upper obound on } f}$$

- Convexity gives us an affine lower bound:

$$f(x) \geq \overbrace{f(x_t) + \langle \nabla f(x_t), x - x_t \rangle}^{\text{affine lower bound on } f}$$

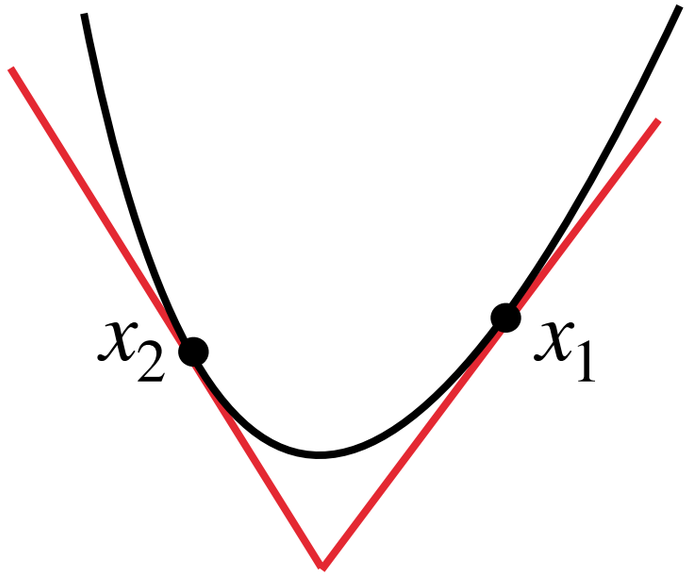


- Today: build better lower bounds, converge faster

Towards Better Lower Bounds

- Suppose we have queried the gradients at $x_1, \ldots, x_t$
- Each gradient $\nabla f(x_i)$ gives us a lower bound on f

$$f(x) \geq f(x_i) + \langle \nabla f(x_i), x - x_i \rangle \quad \forall i = 1, \ldots, t$$

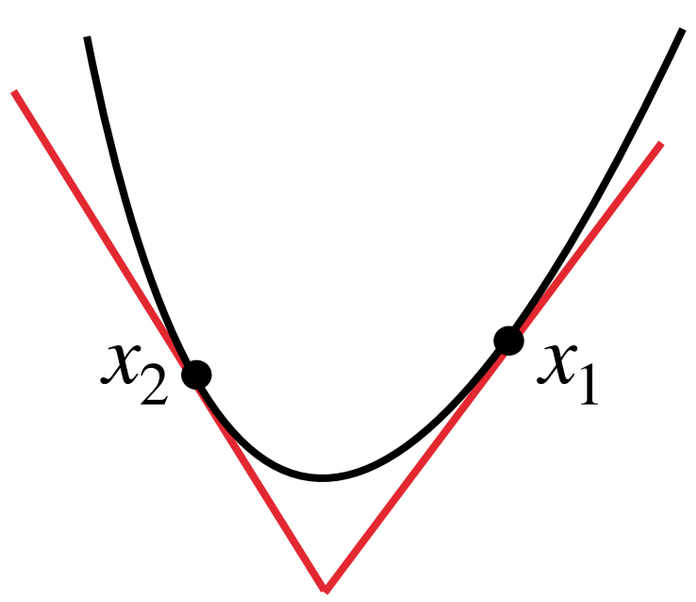


Towards Better Lower Bounds

- Suppose we have queried the gradients at $x_1, \ldots, x_t$
- Each gradient $\nabla f(x_i)$ gives us a lower bound on f

$$f(x) \geq f(x_i) + \langle \nabla f(x_i), x - x_i \rangle \quad \forall i = 1, \ldots, t$$

- Best choice: maximum (downside: complicated)

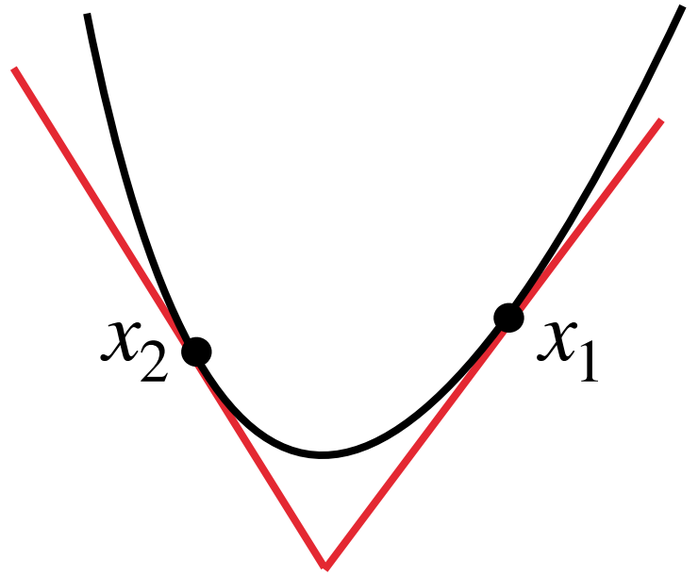


Towards Better Lower Bounds

- Suppose we have queried the gradients at $x_1, \ldots, x_t$
- Each gradient $\nabla f(x_i)$ gives us a lower bound on f

$$f(x) \geq f(x_i) + \langle \nabla f(x_i), x - x_i \rangle \quad \forall i = 1, \ldots, t$$

- Best choice: maximum (downside: complicated)
- Next best choice: a convex combination



Towards Better Lower Bounds

- Suppose we have queried the gradients at $x_1, \ldots, x_t$
- Each gradient $\nabla f(x_i)$ gives us a lower bound on $f(x^*)$

$$f(x) \geq f(x_i) + \langle \nabla f(x_i), x - x_i \rangle \quad \forall i = 1, \ldots, t$$

- Let $a_1, a_2, \ldots, a_t > 0$ be any positive weights and $A_t = \sum_{i=1}^{t} a_i$

$$f(x) \geq \frac{1}{A_t} \sum_{i=1}^{t} \left(a_i f(x_i) + a_i \langle \nabla f(x_i), x - x_i \rangle \right)$$

Towards an Algorithm

- The gradients seen so far give us a combined lower bound:

$$f(x) \geq \frac{1}{A_t} \sum_{i=1}^{t} \left(a_i f(x_i) + a_i \left\langle \nabla f(x_i), x - x_i \right\rangle \right)$$

Towards an Algorithm

- The gradients seen so far give us a combined lower bound:

$$f(x) \geq \frac{1}{A_t} \sum_{i=1}^{t} \left(a_i f(x_i) + a_i \left\langle \nabla f(x_i), x - x_i \right\rangle \right)$$

- Recall the gradient descent approach:

$$\underbrace{x_{t+1}}_{\text{next iterate}} = \arg\min_{x \in K} \left\{ \underbrace{f(x_t) + \left\langle \nabla f(x_t), x - x_t \right\rangle}_{\text{lower bound}} + \frac{1}{2\eta_t} \underbrace{\| x - x_t \|^2}_{\text{proximity term}} \right\}$$

Towards an Algorithm

- The gradients seen so far give us a combined lower bound:

$$f(x) \geq \frac{1}{A_t} \sum_{i=1}^{t} \left(a_i f(x_i) + a_i \left\langle \nabla f(x_i), x - x_i \right\rangle \right)$$

- Recall the gradient descent approach:

$$\underbrace{x_{t+1}}_{\text{next iterate}} = \arg\min_{x \in K} \left\{ \underbrace{f(x_t) + \left\langle \nabla f(x_t), x - x_t \right\rangle}_{\text{lower bound}} + \frac{1}{2\eta_t} \underbrace{\| x - x_t \|^2}_{\text{proximity term}} \right\}$$

- Following the GD framework:

$$z_t = \arg\min_{x \in K} \left\{ \frac{1}{A_t} \sum_{i=1}^{t} \left(a_i f(x_i) + a_i \left\langle \nabla f(x_i), x - x_i \right\rangle + \frac{1}{2\eta_t} \| x - z_0 \|^2 \right) \right\}$$

$$= \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{1}{2\eta_t} \| x - z_0 \|^2 \right\}$$

Towards an Algorithm

- We have settled on the following main iterates:

$$z_t = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{1}{2\eta_t} \left\| x - z_0 \right\|^2 \right\}$$

Towards an Algorithm

- We have settled on the following main iterates:

$$z_t = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{1}{2\eta_t} \left\| x - z_0 \right\|^2 \right\}$$

- Following GD, we can return the average of the iterates, but we will incorporate the weights $\{a_t\}$:

$$\bar{z}_t = \frac{\sum_{i=1}^{t} a_i z_i}{A_t}$$

return $\bar{z}_T$

Towards an Algorithm

- We have settled on the following main iterates:

$$z_t = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{1}{2\eta_t} \left\| x - z_0 \right\|^2 \right\}$$

- Following GD, we can return the average of the iterates, but we will incorporate the weights $\{a_t\}$:

$$z_t = \frac{\sum_{i=1}^{t} a_i z_i}{A_t}$$

- For $\{x_t\}$, we use a similar approach, but with a twist:

$$x_t = \frac{\sum_{i=1}^{t-1} a_i z_i + a_t z_{t-1}}{A_t}$$

Towards an Algorithm

- We have settled on the following main iterates:

$$z_t = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{1}{2\eta_t} \left\| x - z_0 \right\|^2 \right\}$$

- Following GD, we can return the average of the iterates, but we will incorporate the weights $\{a_t\}$:

$$z_t = \frac{\sum_{i=1}^{t} a_i z_i}{A_t}$$

- For $\{x_t\}$, we use a similar approach, but with a twist:

$$x_t = \frac{\sum_{i=1}^{t-1} a_i z_i + a_t z_{t-1}}{A_t}$$

- The choice of weights $\{a_t\}$ will follow organically from the analysis

Accelerated Methods

AGD+ algorithm [Gasnikov, Nesterov 2016; Cohen et al. 2018]

Choose $z_0 \in K$, weights $a_t \geq 0, A_t = \sum_{i=1}^{t} a_i$

For $t = 1, \ldots, T$:

$$x_t = \frac{\sum_{i=1}^{t-1} a_i z_i + a_t z_{t-1}}{A_t}$$

$$z_t = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{\beta}{2} \| x - z_0 \|^2 \right\}$$

Return $\bar{z}_T := \sum_{t=1}^{T} \frac{a_t}{A_T} z_t$

Accelerated Methods

AGD+ algorithm [Gasnikov, Nesterov 2016; Cohen et al. 2018]

Choose $z_0 \in K$, weights $a_t = \Theta(t)$, $A_t = \sum_{i=1}^{t} a_i = \Theta(t^2)$

For $t = 1, \ldots, T$:

$$x_t = \frac{\sum_{i=1}^{t-1} a_i z_i + a_t z_{t-1}}{A_t}$$

$$z_t = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{\beta}{2} \| x - z_0 \|^2 \right\}$$

Return $\bar{z}_T := \sum_{t=1}^{T} \frac{a_t}{A_T} z_t$

$$T = O\left(\sqrt{\frac{1}{\epsilon}}\right) \text{ optimal}$$

AGD+ Analysis

Our goal is to upper bound $f(\overline{z}_t) - f(x^*)$

AGD+ Analysis

Our goal is to upper bound $f(\overline{z}_t) - f(x^*)$

Our combined lower bound gives us:

$$A_t f(x^*) \geq \sum_{i=1}^{t} a_i f(x_i) + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle$$

AGD+ Analysis

Our goal is to upper bound $f(\bar{z}_t) - f(x^*)$

Our combined lower bound gives us:

$$A_t f(x^*) \geq \sum_{i=1}^{t} a_i f(x_i) + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle$$

A key step is to connect the lower bound to the main update:

$$z_t = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{\beta}{2} \| x - z_0 \|^2 \right\}$$

AGD+ Analysis

Our goal is to upper bound $f(\bar{z}_t) - f(x^*)$

Our combined lower bound gives us:

$$A_t f(x^*) \geq \sum_{i=1}^{t} a_i f(x_i) + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle$$

A key step is to connect the lower bound to the main update:

$$z_t = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{\beta}{2} \| x - z_0 \|^2 \right\}$$

$$A_t f(x^*) \geq \sum_{i=1}^{t} a_i f(x_i) + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle$$

AGD+ Analysis

Our goal is to upper bound $f(\bar{z}_t) - f(x^*)$

Our combined lower bound gives us:

$$A_t f(x^*) \geq \sum_{i=1}^{t} a_i f(x_i) + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle$$

A key step is to connect the lower bound to the main update:

$$z_t = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{\beta}{2} \left\| x - z_0 \right\|^2 \right\}$$

$$\begin{aligned} A_t f(x^*) &\geq \sum_{i=1}^{t} a_i f(x_i) + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle \\ &= \sum_{i=1}^{t} a_i f(x_i) - \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle + \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 \end{aligned}$$

AGD+ Analysis

Our goal is to upper bound $f(\bar{z}_t) - f(x^*)$

Our combined lower bound gives us:

$$A_t f(x^*) \geq \sum_{i=1}^{t} a_i f(x_i) + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle$$

A key step is to connect the lower bound to the main update:

$$z_t = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{\beta}{2} \left\| x - z_0 \right\|^2 \right\}$$

$$\begin{aligned}
A_t f(x^*) &\geq \sum_{i=1}^{t} a_i f(x_i) + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle \\
&= \sum_{i=1}^{t} a_i f(x_i) - \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle + \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 \\
&\geq \sum_{i=1}^{t} a_i f(x_i) - \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 + \min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x - x_i \right\rangle + \frac{\beta}{2} \left\| x - z_0 \right\|^2 \right\}
\end{aligned}$$

AGD+ Analysis

Our goal is to upper bound $f(\bar{z}_t) - f(x^*)$

Our combined lower bound gives us:

$$A_t f(x^*) \geq \sum_{i=1}^{t} a_i f(x_i) + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle$$

A key step is to connect the lower bound to the main update:

$$z_t = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{\beta}{2} \left\| x - z_0 \right\|^2 \right\}$$

$$\begin{aligned}
A_t f(x^*) &\geq \sum_{i=1}^{t} a_i f(x_i) + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle \\
&= \sum_{i=1}^{t} a_i f(x_i) - \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x^* - x_i \right\rangle + \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 \\
&\geq \sum_{i=1}^{t} a_i f(x_i) - \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 + \min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x - x_i \right\rangle + \frac{\beta}{2} \left\| x - z_0 \right\|^2 \right\} \\
&= \sum_{i=1}^{t} a_i f(x_i) - \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), z_t - x_i \right\rangle + \frac{\beta}{2} \left\| z_t - z_0 \right\|^2
\end{aligned}$$

AGD+ Analysis

We have shown:

$$f(x^*) \geq \underbrace{\frac{1}{A_t}\left(\sum_{i=1}^{t} a_i f(x_i) - \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), z_t - x_i \right\rangle + \frac{\beta}{2} \left\| z_t - z_0 \right\|^2\right)}_{:=L_t}$$

Thus $f(\bar{z}_t) - f(x^*) \leq f(\bar{z}_t) - L_t$

$A_t L_t - A_{t-1} L_{t-1}$

AGD+ Analysis

We have shown:

$$f(x^*) \geq \underbrace{\frac{1}{A_t}\left(\sum_{i=1}^{t} a_i f(x_i) - \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 + \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), z_t - x_i \right\rangle + \frac{\beta}{2} \left\| z_t - z_0 \right\|^2 \right)}_{:=L_t}$$

Thus $f(\bar{z}_t) - f(x^*) \leq f(\bar{z}_t) - L_t$

Next, we analyze how the lower bounds are evolving:

$$\begin{aligned} A_{t-1}L_{t-1} - A_t L_t = & - a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle \\ & + \sum_{i=1}^{t-1} a_i \left\langle \nabla f(x_i), z_{t-1} - z_t \right\rangle \\ & + \frac{\beta}{2} \left\| z_{t-1} - z_0 \right\|^2 - \frac{\beta}{2} \left\| z_t - z_0 \right\|^2 \end{aligned}$$

AGD+ Analysis

We have

$$A_{t-1}L_{t-1} - A_tL_t = -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle$$
$$+ \sum_{i=1}^{t-1} a_i \left\langle \nabla f(x_i), z_{t-1} - z_t \right\rangle$$
$$+ \frac{\beta}{2} \| z_{t-1} - z_0 \|^2 - \frac{\beta}{2} \| z_t - z_0 \|^2$$

AGD+ Analysis

We have

$$A_{t-1}L_{t-1} - A_tL_t = -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle$$

$$+ \sum_{i=1}^{t-1} a_i \left\langle \nabla f(x_i), z_{t-1} - z_t \right\rangle$$

$$+ \frac{\beta}{2} \| z_{t-1} - z_0 \|^2 - \frac{\beta}{2} \| z_t - z_0 \|^2$$

AGD+ Analysis

We have

$$A_{t-1}L_{t-1} - A_tL_t = -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle$$
$$+ \sum_{i=1}^{t-1} a_i \left\langle \nabla f(x_i), z_{t-1} - z_t \right\rangle$$
$$+ \frac{\beta}{2} \| z_{t-1} - z_0 \|^2 - \frac{\beta}{2} \| z_t - z_0 \|^2$$

We can bound the above term using the optimality condition:

$$z_{t-1} = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t-1} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{\beta}{2} \| x - z_0 \|^2 \right\}$$

AGD+ Analysis

We have

$$A_{t-1}L_{t-1} - A_tL_t = -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle$$
$$+ \sum_{i=1}^{t-1} a_i \left\langle \nabla f(x_i), z_{t-1} - z_t \right\rangle$$
$$+ \frac{\beta}{2} \| z_{t-1} - z_0 \|^2 - \frac{\beta}{2} \| z_t - z_0 \|^2$$

We can bound the above term using the optimality condition:

$$z_{t-1} = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t-1} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{\beta}{2} \| x - z_0 \|^2 \right\}$$

$$\left\langle \sum_{i=1}^{t-1} a_i \nabla f(x_i) + \beta \left(z_{t-1} - z_0 \right), z_{t-1} - z_t \right\rangle \leq 0$$

AGD+ Analysis

The optimality condition for z_{t-1} gives us:

$$\left\langle \sum_{i=1}^{t-1} a_i \nabla f(x_i) + \beta \left(z_{t-1} - z_0 \right), z_{t-1} - z_t \right\rangle \leq 0$$

Rearranging and using the identity $ab = \frac{1}{2}(a+b)^2 - \frac{1}{2}a^2 - \frac{1}{2}b^2$:

$$\sum_{i=1}^{t-1} a_i \left\langle \nabla f(x_i), z_{t-1} - z_t \right\rangle \leq \beta \left\langle z_0 - z_{t-1}, z_{t-1} - z_t \right\rangle$$

$$= \frac{\beta}{2} \left(\| z_0 - z_t \|^2 - \| z_0 - z_{t-1} \|^2 - \| z_{t-1} - z_t \|^2 \right)$$

$ab = \frac{1}{2}(a+b)^2 - \frac{1}{2}a^2 - \frac{1}{2}b^2$

AGD+ Analysis

We have

$$A_{t-1}L_{t-1} - A_tL_t = -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle$$
$$+ \sum_{i=1}^{t-1} a_i \left\langle \nabla f(x_i), z_{t-1} - z_t \right\rangle$$
$$+ \frac{\beta}{2} \left\| z_{t-1} - z_0 \right\|^2 - \frac{\beta}{2} \left\| z_t - z_0 \right\|^2$$

The optimality condition for z_{t-1} gives us:

$$\sum_{i=1}^{t-1} a_i \left\langle \nabla f(x_i), z_{t-1} - z_t \right\rangle \leq \frac{\beta}{2} \left(\left\| z_0 - z_t \right\|^2 - \left\| z_0 - z_{t-1} \right\|^2 - \left\| z_{t-1} - z_t \right\|^2 \right)$$

We combine the two and obtain

$$A_{t-1}L_{t-1} - A_tL_t \leq -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \left\| z_{t-1} - z_t \right\|^2$$

AGD+ Analysis

We have shown:

$$A_{t-1}L_{t-1} - A_tL_t \leq -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \left\| z_{t-1} - z_t \right\|^2$$

AGD+ Analysis

We have shown:

$$A_{t-1}L_{t-1} - A_tL_t \leq -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \left\| z_{t-1} - z_t \right\|^2$$

Thus

$$\begin{aligned}
&A_t \left(f(\overline{z}_t) - L_t\right) - A_{t-1} \left(f(\overline{z}_{t-1}) - L_{t-1}\right) \\
&= A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) + A_{t-1}L_{t-1} - A_tL_t \\
&\leq A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \left\| z_{t-1} - z_t \right\|^2
\end{aligned}$$

AGD+ Analysis

We have shown:

$$A_{t-1}L_{t-1} - A_tL_t \leq -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \| z_{t-1} - z_t \|^2$$

Thus

$$\begin{aligned}
&A_t \left(f(\bar{z}_t) - L_t\right) - A_{t-1} \left(f(\bar{z}_{t-1}) - L_{t-1}\right) \\
&= A_t f(\bar{z}_t) - A_{t-1} f(\bar{z}_{t-1}) + A_{t-1}L_{t-1} - A_tL_t \\
&\leq \boxed{A_t f(\bar{z}_t) - A_{t-1} f(\bar{z}_{t-1}) - a_t f(x_t)} - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \| z_{t-1} - z_t \|^2
\end{aligned}$$

$$A_t = \sum_{i=1}^{t} a_i \qquad a_t = A_t - A_{t-1}$$

AGD+ Analysis

We have shown:

$$A_{t-1}L_{t-1} - A_tL_t \leq -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \left\| z_{t-1} - z_t \right\|^2$$

Thus

$$\begin{aligned} & A_t \left(f(\overline{z}_t) - L_t \right) - A_{t-1} \left(f(\overline{z}_{t-1}) - L_{t-1} \right) \\ & = A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) + A_{t-1}L_{t-1} - A_tL_t \\ & \leq \boxed{A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - a_t f(x_t)} - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \left\| z_{t-1} - z_t \right\|^2 \end{aligned}$$

We now leverage smoothness and convexity:

$$\begin{aligned} & A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - a_t f(x_t) \\ & = A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - \left(A_t - A_{t-1} \right) f(x_t) \end{aligned}$$

AGD+ Analysis

We have shown:

$$A_{t-1}L_{t-1} - A_tL_t \leq -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \left\| z_{t-1} - z_t \right\|^2$$

Thus

$$\begin{aligned}
&A_t \left(f(\overline{z}_t) - L_t\right) - A_{t-1} \left(f(\overline{z}_{t-1}) - L_{t-1}\right) \\
&= A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) + A_{t-1}L_{t-1} - A_tL_t \\
&\leq \boxed{A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - a_t f(x_t)} - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \left\| z_{t-1} - z_t \right\|^2
\end{aligned}$$

We now leverage smoothness and convexity:

$$\begin{aligned}
&A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - a_t f(x_t) \\
&= A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - \left(A_t - A_{t-1}\right) f(x_t) \\
&= A_t \left(f(\overline{z}_t) - f(x_t)\right) - A_{t-1} \left(f(\overline{z}_{t-1}) - f(x_t)\right)
\end{aligned}$$

AGD+ Analysis

We have shown:

$$A_{t-1}L_{t-1} - A_tL_t \leq -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \left\| z_{t-1} - z_t \right\|^2$$

Thus

$$\begin{aligned}
&A_t \left(f(\overline{z}_t) - L_t\right) - A_{t-1} \left(f(\overline{z}_{t-1}) - L_{t-1}\right) \\
&= A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) + A_{t-1}L_{t-1} - A_tL_t \\
&\leq \boxed{A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - a_t f(x_t)} - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \left\| z_{t-1} - z_t \right\|^2
\end{aligned}$$

We now leverage smoothness and convexity:

$$\begin{aligned}
&A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - a_t f(x_t) \\
&= A_t f(\overline{z}_t) - A_{t-1} f(\overline{z}_{t-1}) - \left(A_t - A_{t-1}\right) f(x_t) \\
&= \underbrace{A_t \left(f(\overline{z}_t) - f(x_t)\right)}_{\text{smoothness}} - \underbrace{A_{t-1} \left(f(\overline{z}_{t-1}) - f(x_t)\right)}_{\text{convexity}}
\end{aligned}$$

AGD+ Analysis

We have shown:

$$A_{t-1}L_{t-1} - A_tL_t \leq -a_t f(x_t) - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \left\| z_{t-1} - z_t \right\|^2$$

Thus

$$\begin{aligned}
&A_t \left(f(\bar{z}_t) - L_t\right) - A_{t-1} \left(f(\bar{z}_{t-1}) - L_{t-1}\right) \\
&= A_t f(\bar{z}_t) - A_{t-1} f(\bar{z}_{t-1}) + A_{t-1}L_{t-1} - A_tL_t \\
&\leq \boxed{A_t f(\bar{z}_t) - A_{t-1} f(\bar{z}_{t-1}) - a_t f(x_t)} - a_t \left\langle \nabla f(x_t), z_t - x_t \right\rangle - \frac{\beta}{2} \left\| z_{t-1} - z_t \right\|^2
\end{aligned}$$

We now leverage smoothness and convexity:

$$\begin{aligned}
&A_t f(\bar{z}_t) - A_{t-1} f(\bar{z}_{t-1}) - a_t f(x_t) \\
&= A_t f(\bar{z}_t) - A_{t-1} f(\bar{z}_{t-1}) - \left(A_t - A_{t-1}\right) f(x_t) \\
&= A_t \underbrace{\left(f(\bar{z}_t) - f(x_t)\right)}_{\text{smoothness}} - A_{t-1} \underbrace{\left(f(\bar{z}_{t-1}) - f(x_t)\right)}_{\text{convexity}} \\
&\leq A_t \left(\left\langle \nabla f(x_t), \bar{z}_t - x_t \right\rangle + \frac{\beta}{2} \left\| \bar{z}_t - x_t \right\|^2 \right) - A_{t-1} \left\langle \nabla f(x_t), \bar{z}_{t-1} - x_t \right\rangle
\end{aligned}$$

AGD+ Analysis

Thus we have:

$$A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right)$$

$$\leq\left\langle\nabla f(x_t),A_t\left(\bar{z}_t-x_t\right)-A_{t-1}\left(\bar{z}_{t-1}-x_t\right)-a_t\left(z_t-x_t\right)\right\rangle+\frac{\beta}{2}A_t\parallel\bar{z}_t-x_t\parallel^2-\frac{\beta}{2}\parallel z_{t-1}-z_t\parallel^2$$

AGD+ Analysis

Thus we have:

$$A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right)$$

$$\leq\left\langle \nabla f(x_t), \underbrace{A_t\left(\bar{z}_t-x_t\right)-A_{t-1}\left(\bar{z}_{t-1}-x_t\right)-a_t\left(z_t-x_t\right)}_{=0}\right\rangle+\frac{\beta}{2}A_t\parallel\bar{z}_t-x_t\parallel^2-\frac{\beta}{2}\parallel z_{t-1}-z_t\parallel^2$$

AGD+ Analysis

Thus we have:

$$
\begin{aligned}
& A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right) \\
& \leq \Big\langle \nabla f(x_t), \underbrace{A_t\left(\bar{z}_t-x_t\right)-A_{t-1}\left(\bar{z}_{t-1}-x_t\right)-a_t\left(z_t-x_t\right)}_{=0} \Big\rangle + \frac{\beta}{2}A_t \parallel \bar{z}_t-x_t \parallel^2 - \frac{\beta}{2} \parallel z_{t-1}-z_t \parallel^2 \\
& = \frac{\beta}{2}A_t \parallel \bar{z}_t-x_t \parallel^2 - \frac{\beta}{2} \parallel z_{t-1}-z_t \parallel^2
\end{aligned}
$$

AGD+ Analysis

Thus we have:

$$A_t \left(f(\overline{z}_t) - L_t\right) - A_{t-1} \left(f(\overline{z}_{t-1}) - L_{t-1}\right) \leq \frac{\beta}{2} A_t \parallel \overline{z}_t - x_t \parallel^2 - \frac{\beta}{2} \parallel z_{t-1} - z_t \parallel^2$$

AGD+ Analysis

Thus we have:

$$A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right)\leq \underbrace{\frac{\beta}{2}A_t\parallel \bar{z}_t-x_t\parallel^2-\frac{\beta}{2}\parallel z_{t-1}-z_t\parallel^2}_{\text{want it to be small}}$$

AGD+ Analysis

Thus we have:

$$A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right)\leq \underbrace{\frac{\beta}{2}A_t\parallel \bar{z}_t-x_t\parallel^2-\frac{\beta}{2}\parallel z_{t-1}-z_t\parallel^2}_{\text{want it to be small}}$$

We can choose x_t to make the upper bound small

AGD+ Analysis

Thus we have:

$$A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right)\leq \underbrace{\frac{\beta}{2}A_t\parallel \bar{z}_t-x_t\parallel^2-\frac{\beta}{2}\parallel z_{t-1}-z_t\parallel^2}_{\text{want it to be small}}$$

We can choose x_t to make the upper bound small

We need to make the two distances become related:

$$\bar{z}_t-x_t=\lambda_t\left(z_t-z_{t-1}\right)$$

AGD+ Analysis

Thus we have:

$$A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right)\leq \underbrace{\frac{\beta}{2}A_t\left\|\bar{z}_t-x_t\right\|^2-\frac{\beta}{2}\left\|z_{t-1}-z_t\right\|^2}_{\text{want it to be small}}$$

We can choose x_t to make the upper bound small

We need to make the two distances become related:

$$\bar{z}_t-x_t=\lambda_t\left(z_t-z_{t-1}\right)$$

$$\Rightarrow x_t=\bar{z}_t+\lambda_t\left(z_{t-1}-z_t\right)$$

$$=\frac{A_{t-1}\bar{z}_{t-1}+A_t\lambda_t z_{t-1}+\left(a_t-A_t\lambda_t\right)z_t}{A_t}$$

$$\bar{z}_t=\frac{A_{t-1}\,\bar{z}_{t-1}+a_t z_t}{A_t}$$

AGD+ Analysis

Thus we have:

$$A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right)\leq \underbrace{\frac{\beta}{2}A_t\parallel \bar{z}_t-x_t\parallel^2-\frac{\beta}{2}\parallel z_{t-1}-z_t\parallel^2}_{\text{want it to be small}}$$

We can choose x_t to make the upper bound small

We need to make the two distances become related:

$$\bar{z}_t-x_t=\lambda_t\left(z_t-z_{t-1}\right)$$

$$\Rightarrow x_t=\bar{z}_t+\lambda_t\left(z_{t-1}-z_t\right)$$

$$=\frac{A_{t-1}\bar{z}_{t-1}+A_t\lambda_t z_{t-1}+\left(a_t-A_t\lambda_t\right)z_t}{A_t}$$

We need to compute x_t without access to z_t:

$$a_t-A_t\lambda_t=0\Rightarrow\lambda_t=\frac{a_t}{A_t}$$

AGD+ Analysis

Thus we have:

$$A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right)\leq \underbrace{\frac{\beta}{2}A_t\parallel \bar{z}_t-x_t\parallel^2-\frac{\beta}{2}\parallel z_{t-1}-z_t\parallel^2}_{\text{want it to be small}}$$

We can choose x_t to make the upper bound small:

$$x_t=\frac{A_{t-1}\bar{z}_{t-1}+a_t z_{t-1}}{A_t}=\frac{\sum_{i=1}^{t-1}a_i z_i+a_t z_{t-1}}{A_t}$$

AGD+ Analysis

Thus we have:

$$A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right)\le \underbrace{\frac{\beta}{2}A_t\parallel \bar{z}_t-x_t\parallel^2-\frac{\beta}{2}\parallel z_{t-1}-z_t\parallel^2}_{\text{want it to be small}}$$

We can choose x_t to make the upper bound small:

$$x_t=\frac{A_{t-1}\bar{z}_{t-1}+a_t z_{t-1}}{A_t}=\frac{\sum_{i=1}^{t-1}a_i z_i+a_t z_{t-1}}{A_t}$$

Thus we obtain

$$A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right)\le \frac{\beta}{2}\underbrace{\left(\frac{a_t^2}{A_t}-1\right)}_{\le 0\ ?}\parallel z_t-z_{t-1}\parallel^2$$

Want: $a_t^2\le A_t=a_1+\dots+a_t$

Recall: $\sum_{i=1}^{t} i=i^2 \rightarrow$ set $a_t=ct \Rightarrow c^2t^2\le c\frac{t(t+1)}{2} \rightarrow c=\frac{1}{2}$

AGD+ Analysis

Thus we have:

$$A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right)\leq \underbrace{\frac{\beta}{2}A_t\parallel \bar{z}_t-x_t\parallel^2-\frac{\beta}{2}\parallel z_{t-1}-z_t\parallel^2}_{\text{want it to be small}}$$

We can choose x_t to make the upper bound small:

$$x_t=\frac{A_{t-1}\bar{z}_{t-1}+a_t z_{t-1}}{A_t}=\frac{\sum_{i=1}^{t-1}a_i z_i+a_t z_{t-1}}{A_t}$$

Thus we obtain

$$A_t\left(f(\bar{z}_t)-L_t\right)-A_{t-1}\left(f(\bar{z}_{t-1})-L_{t-1}\right)\leq\frac{\beta}{2}\left(\frac{a_t^2}{A_t}-1\right)\parallel z_t-z_{t-1}\parallel^2$$

We now choose a_t to make the coefficient ≤ 0:

$$a_t=\frac{1}{2}t \qquad A_t=\sum_{i=1}^{t}a_i=\frac{1}{4}t(t+1)$$

AGD+ Analysis

We have:

$$A_t \left(f(\overline{z}_t) - L_t\right) - A_{t-1} \left(f(\overline{z}_{t-1}) - L_{t-1}\right) \leq 0$$

Summing up, we obtain

$$A_T \left(f(\overline{z}_T) - L_T\right) \leq A_1 \left(f(z_1) - L_1\right)$$

AGD+ Analysis

We have:

$$A_t \left(f(\overline{z}_t) - L_t\right) - A_{t-1} \left(f(\overline{z}_{t-1}) - L_{t-1}\right) \leq 0$$

Summing up, we obtain

$$A_T \left(f(\overline{z}_T) - L_T\right) \leq A_1 \left(f(z_1) - L_1\right)$$

$$a_1 = A_1 = \frac{1}{2}$$

AGD+ Analysis

We have:

$$A_t \left(f(\bar{z}_t) - L_t \right) - A_{t-1} \left(f(\bar{z}_{t-1}) - L_{t-1} \right) \leq 0$$

Summing up, we obtain

$$A_T \left(f(\bar{z}_T) - L_T \right) \leq A_1 \left(f(z_1) - L_1 \right)$$

$$a_1 = A_1 = \frac{1}{2}$$

$$L_1 = \frac{1}{A_1} \left(a_1 f(x_1) - \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 + a_1 \left\langle \nabla f(x_1), z_1 - x_1 \right\rangle + \frac{\beta}{2} \left\| z_1 - z_0 \right\|^2 \right)$$

AGD+ Analysis

We have:

$$A_t \left(f(\bar{z}_t) - L_t\right) - A_{t-1} \left(f(\bar{z}_{t-1}) - L_{t-1}\right) \leq 0$$

Summing up, we obtain

$$A_T \left(f(\bar{z}_T) - L_T\right) \leq A_1 \left(f(z_1) - L_1\right)$$

$$a_1 = A_1 = \frac{1}{2}$$

$$L_1 = \frac{1}{A_1}\left(a_1 f(x_1) - \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 + a_1 \left\langle \nabla f(x_1), z_1 - x_1 \right\rangle + \frac{\beta}{2} \left\| z_1 - z_0 \right\|^2\right)$$

$$= f(x_1) - \beta \left\| x^* - z_0 \right\|^2 + \left\langle \nabla f(x_1), z_1 - x_1 \right\rangle + \beta \left\| z_1 - z_0 \right\|^2$$

$$x_1 = \frac{a_1 z_0}{A_1} = z_0$$

AGD+ Analysis

We have:

$$A_t \left(f(\bar{z}_t) - L_t\right) - A_{t-1} \left(f(\bar{z}_{t-1}) - L_{t-1}\right) \leq 0$$

Summing up, we obtain

$$A_T \left(f(\bar{z}_T) - L_T\right) \leq A_1 \left(f(z_1) - L_1\right)$$

$$a_1 = A_1 = \frac{1}{2}$$

$$L_1 = \frac{1}{A_1}\left(a_1 f(x_1) - \frac{\beta}{2} \left\| x^* - z_0 \right\|^2 + a_1 \left\langle \nabla f(x_1), z_1 - x_1 \right\rangle + \frac{\beta}{2} \left\| z_1 - z_0 \right\|^2\right)$$

$$= f(x_1) - \beta \left\| x^* - z_0 \right\|^2 + \left\langle \nabla f(x_1), z_1 - x_1 \right\rangle + \beta \left\| z_1 - \underbrace{z_0}_{=x_1} \right\|^2$$

AGD+ Analysis

We have:

$$A_t \left(f(\overline{z}_t) - L_t\right) - A_{t-1} \left(f(\overline{z}_{t-1}) - L_{t-1}\right) \leq 0$$

Summing up, we obtain

$$A_T \left(f(\overline{z}_T) - L_T\right) \leq A_1 \left(f(z_1) - L_1\right)$$

$$a_1 = A_1 = \frac{1}{2}$$

$$L_1 = \frac{1}{A_1}\left(a_1 f(x_1) - \frac{\beta}{2} \parallel x^* - z_0 \parallel^2 + a_1 \left\langle \nabla f(x_1), z_1 - x_1 \right\rangle + \frac{\beta}{2} \parallel z_1 - z_0 \parallel^2\right)$$

$$= f(x_1) - \beta \parallel x^* - z_0 \parallel^2 + \left\langle \nabla f(x_1), z_1 - x_1 \right\rangle + \beta \parallel z_1 - \underbrace{z_0}_{=x_1} \parallel^2$$

$$= f(x_1) - \beta \parallel x^* - z_0 \parallel^2 + \left\langle \nabla f(x_1), z_1 - x_1 \right\rangle + \beta \parallel z_1 - x_1 \parallel^2$$

AGD+ Analysis

We have:

$$A_t \left(f(\bar{z}_t) - L_t \right) - A_{t-1} \left(f(\bar{z}_{t-1}) - L_{t-1} \right) \leq 0$$

Summing up, we obtain

$$A_T \left(f(\bar{z}_T) - L_T \right) \leq A_1 \left(f(z_1) - L_1 \right)$$

$$= \frac{1}{2} \left(\beta \parallel x^* - z_0 \parallel^2 + f(z_1) - f(x_1) - \left\langle \nabla f(x_1), z_1 - x_1 \right\rangle - \beta \parallel z_1 - x_1 \parallel^2 \right)$$

AGD+ Analysis

We have:

$$A_t \left(f(\bar{z}_t) - L_t\right) - A_{t-1} \left(f(\bar{z}_{t-1}) - L_{t-1}\right) \leq 0$$

Summing up, we obtain

$$A_T \left(f(\bar{z}_T) - L_T\right) \leq A_1 \left(f(z_1) - L_1\right)$$

$$= \frac{1}{2}\left(\beta \parallel x^* - z_0 \parallel^2 + \underbrace{f(z_1) - f(x_1) - \left\langle \nabla f(x_1), z_1 - x_1 \right\rangle}_{\leq \frac{\beta}{2} \parallel z_1 - x_1 \parallel^2 \text{ by smoothness}} - \beta \parallel z_1 - x_1 \parallel^2 \right)$$

AGD+ Analysis

We have:

$$A_t\left(f(\bar{z}_t) - L_t\right) - A_{t-1}\left(f(\bar{z}_{t-1}) - L_{t-1}\right) \leq 0$$

Summing up, we obtain

$$A_T\left(f(\bar{z}_T) - L_T\right) \leq A_1\left(f(z_1) - L_1\right)$$

$$= \frac{1}{2}\left(\beta \parallel x^* - z_0 \parallel^2 + \underbrace{f(z_1) - f(x_1) - \left\langle \nabla f(x_1), z_1 - x_1 \right\rangle}_{\leq \frac{\beta}{2} \parallel z_1 - x_1 \parallel^2 \text{ by smoothness}} - \beta \parallel z_1 - x_1 \parallel^2\right)$$

$$\leq \frac{\beta \parallel x^* - z_0 \parallel^2}{2}$$

AGD+ Analysis

We have:

$$A_t \left(f(\bar{z}_t) - L_t\right) - A_{t-1} \left(f(\bar{z}_{t-1}) - L_{t-1}\right) \leq 0$$

Summing up, we obtain

$$A_T \left(f(\bar{z}_T) - L_T\right) \leq A_1 \left(f(z_1) - L_1\right)$$

$$= \frac{1}{2}\left(\beta \parallel x^* - z_0 \parallel^2 + \underbrace{f(z_1) - f(x_1) - \left\langle \nabla f(x_1), z_1 - x_1 \right\rangle}_{\leq \frac{\beta}{2} \parallel z_1 - x_1 \parallel^2 \text{ by smoothness}} - \beta \parallel z_1 - x_1 \parallel^2 \right)$$

$$\leq \frac{\beta \parallel x^* - z_0 \parallel^2}{2}$$

Thus we have our final convergence guarantee:

$$f(\bar{z}_T) - f(x^*) \leq f(\bar{z}_T) - L_T \leq \Theta\left(\frac{\beta \parallel x^* - z_0 \parallel^2}{T^2} \right) \qquad A_T = \Theta(T^2)$$

Adaptive AGD+

Set the step size based on the iterate movement $\|z_t - z_{t-1}\|^2$

AdaAGD+ algorithm

$$x_t = \frac{\sum_{i=1}^{t-1} a_i z_i + a_t z_{t-1}}{A_t}$$

$$z_t = \arg\min_{x \in K} \left\{ \sum_{i=1}^{t} a_i \left\langle \nabla f(x_i), x \right\rangle + \frac{1}{2\eta_t} \left\| x - z_0 \right\|^2 \right\}$$

$$\frac{1}{\eta_t^2} = \frac{1}{\eta_{t-1}^2} \left(1 + \frac{\left\| z_{t-1} - z_{t-2} \right\|^2}{R^2} \right) \qquad \forall t \geq 2$$