Lemma: Any algorithm satisfying assumption 2

$$x_{i} \in \mathrm{Lin}\{\nabla f_k(x_0), \ldots, \nabla f_K(x_{i-1})\}$$

has $\mathrm{Lin}\{\nabla f_k(x_0), \ldots, \nabla f_K(x_i)\} = \mathbb{R}^{i+1} \times \{0\}^{d-i-1}$.

for $i \leq K$.

Proof. Trivially for $i = 0$. $\nabla f_k(x_0) = A_K x_0 - e_1$
$= -e_1$.

Inductively, for $i > 0$, $\nabla f_k(x_i) = A_K \cdot x_{i-1} - e_1$

$$\in A_K \cdot \mathrm{Lin}\{\nabla f(x_0), \ldots, \nabla f(x_{i-1})\}$$

$$= A_K \cdot \mathbb{R}^i \times \{0\}^{d-i}$$

(since A_K is tridiagonal) $= \mathbb{R}^{i+1} \times \{0\}^{d-i-1}$. $\square$

Check f_K is convex: for $s \in \mathbb{R}^d$

$$s^T \nabla^2 f_k(x) s = \frac{L}{4} s^T A_K s = \frac{L}{4}\left[(s^{(1)})^2 + \sum_{i=1}^{K-1} (s^{(i+1)} - s^{(i)})^2 + (s^{(K)})^2\right]$$

$$\geq 0. \checkmark$$

Check f_K is L-smooth: for $s \in \mathbb{R}^d$

$$s^T \nabla^2 f_k(x) s = \cdots = \frac{L}{4}\left[(s^{(1)})^2 + \sum (s^{(i+1)} - s^{(i)})^2 + (s^{(K)})^2\right]$$

$(a-b)^2 \leq 2a^2 + 2b^2$

$$\leq \frac{L}{4}\left[(s^{(1)})^2 + \sum \left[2(s^{(i)})^2 + 2(s^{(i+1)})^2\right] + (s^{(K)})^2\right]$$

$$\leq \frac{L}{4} \sum 4 (s^{(i)})^2 = L \|s\|_2^2. \checkmark$$

How poorly are solutions with only the first i coordinates nonzero?

only in first k variables

Claim: $\nabla f_k(x) = 0 = \frac{L}{4}(A_k x - e_1)$ has ~~unique~~ sol.

$$\bar{x}_k^{(i)} = \begin{cases} 1 - \frac{i}{k+1} & \text{for } i = 1, \dots, k \\ 0 & \text{otherwise.} \end{cases}$$

$\Rightarrow$ Optimal obj of $f_k(\cdot)$ comes from plugging in

$$\begin{aligned} f_k^* &:= f_k(\bar{x}_k) \\ &= \frac{L}{4}\left(\frac{1}{2}\bar{x}_k^T A_k \bar{x}_k - e_1^T \bar{x}_k\right) \\ &= \frac{L}{4}\left(\frac{1}{2}\left(\left(1 - \frac{1}{k+1}\right)^2 + (k-1)\left(\frac{1}{k+1}\right)^2 + \left(1 - \frac{k}{k+1}\right)^2\right) - \left(1 - \frac{1}{k+1}\right)\right) \\ &= \frac{L}{8}\left(-1 + \frac{1}{k+1}\right) \end{aligned}$$

Likewise, $$\|x_0 - \bar{x}_k\|_2^2 = \sum_{i=1}^{k}\left(1 - \frac{i}{k+1}\right)^2 \le \frac{1}{3}(k-1)$$

(see Nesterov's book for gory details)

Proof of Complexity Lower Bound. Consider fixed $L>0, k>0$.

Then consider $f(x) = f_{2k+1}(x)$, $d = 2k+1$.

$$f(x_k) = f_{2k+1}(x_k) = f_k(x_k) \geq f_k^*$$

$$\Rightarrow \quad \frac{f(x_k) - f(x^*)}{\|x_0 - x^*\|_2^2} \geq \frac{f_k^* - f_{2k+1}^*}{\|x_0 - \bar{x}_{2k+1}\|_2^2}$$

$$\geq \frac{\frac{L}{8}\left(\cancel{-1} + \frac{1}{k+1} \cancel{+1} - \frac{1}{2k+2}\right)}{\frac{1}{3}(2k_{+2})}$$

$$\geq \frac{3L}{32(k+1)^2} \quad \checkmark$$

The second part of our theorem holds as

$$\underset{\substack{\| \\ \bar{x}_{2k+1}}}{\|x_k - x^*\|_2^2} \geq \sum_{i=k+1}^{2k+1} \left(\bar{x}_{2k+1}^{(i)}\right)^2 = \sum_{i=k+1}^{2k+1} \left(1 - \frac{i}{2k+2}\right)^2$$

(if $\frac{3}{2}k$ is an integer)
$$\geq \sum_{i=k+1}^{\frac{3}{2}k+1} \left(1 - \frac{i}{2k+2}\right)^2$$

$$\geq \left(\frac{1}{2}k\right)\left(1 - \frac{3/2\,k+1}{2k+2}\right)^2$$

$$\geq \left(\frac{1}{2}k\right)\left(1/4\right)^2$$

$$\geq \frac{1}{32}k$$

$$\left(\text{since } \|x_0 - x_{2k+1}\|^2 \geq \frac{1}{3}(2k+1)\right) \geq \frac{3}{64}\|x_0 - x^*\|^2. \qquad \square$$

Recap of First-Order Smooth Optimization Results

For any function f with L-Lipschitz gradient, we found the following guarantees:

Method	Generic Rate	Speed-up from $\mu > 0$.
Gradient Descent (for nonconvex f)	$\frac{1}{T}\sum_{k=0}^{T-1} \lVert\nabla f(x_k)\rVert_2^2 \leq O\left(\frac{1}{T}\right)$	Linear Rate Under Strict Positive Definite $\nabla^2 f$
Gradient Descent (for convex f)	$f(x_T) - f(x^*) \leq O\left(\frac{1}{T}\right)$	Linear Rate Under Strong Convexity of f
Accelerated Gradient (for convex f)	$f(y_T) - f(x^*) \leq O\left(\frac{1}{T^2}\right)$	Faster Linear Rate Under Restart with Strong Convexity.

This rate is optimal. (No 1st order method can improve this worstcase)

Proven in HW2, Q3.

Structured Constrained/Nonsmooth Optimization

1. Define Proximal Operator
2. Example Orthogonal Projection on Constraints
3. Projected/Proximal Gradient Methods
4. Acceleration
5. More proximal methods (Alternating projections, ADMM).

1. Proximal Operator

We want a local improvement without on the existence of a gradient.

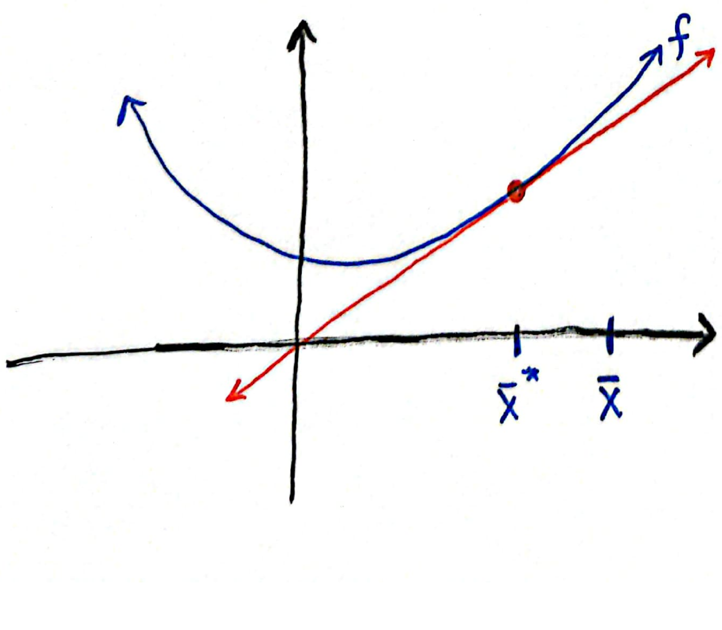
$$\text{prox}_f(\bar{x}) := \underset{x \in \mathbb{R}^d}{\text{argmin}} \left\{ f(x) + \tfrac{1}{2}\|x - \bar{x}\|_2^2 \right\}$$

To allow a stepsize, we can rescale by $\alpha > 0$

$$\text{prox}_{\alpha \cdot f}(\bar{x}) := \text{argmin}\left\{ \alpha f(x) + \tfrac{1}{2}\|x - \bar{x}\|_2^2 \right\}$$
$$= \text{argmin}\left\{ \underbrace{f(x) + \tfrac{1}{2\alpha}\|x - \bar{x}\|_2^2}_{\text{looks like HW2, Q2.}} \right\}$$

In particular, $\bar{x}^* = \text{prox}_{\alpha f}(\bar{x})$, then $g = \frac{1}{\alpha}(\bar{x} - \bar{x}^*)$ gives a linearization

$$f(x) \geq f(\bar{x}^*) + g^T(x - \bar{x}^*).$$



"g is a 'subgradient' of f at $\bar{x}^*$"

$$\partial f(x) = \left\{ g \in \mathbb{R}^d \mid f(y) \geq f(x) + g^T(y - x) \ \forall y \in \mathbb{R}^d \right\}$$

"Intro to Convexity" covers these in more depth.