Summary

Global minimization: $O\left(\left(\frac{1}{\varepsilon}\right)^n\right) \approx \exp(n)$

Goal: Find a stationary point:
$$f'(\bar{x}) \approx 0.$$

$$\min_{x \in \mathbb{R}^n} f(x)$$

The gradient:
$$f(x+h) = f(x) + \langle f'(x), h \rangle + \bar{\bar{o}}(\|h\|)$$

$f'(x) = \nabla f(x) \in \mathbb{R}^n$

The Hessian:
$$f'(x+h) = f'(x) + f''(x)h + \bar{\bar{o}}(\|h\|)$$

$$f''(x) = f''(x)^T \in \mathbb{R}^{n \times n} \text{ – the Hessian matrix.}$$

Idea of gradient descent.

Assume $f'(x) \neq 0$. $\qquad h = -\alpha f'(x), \ \alpha > 0$

$$f(x - \alpha f'(x)) = f(x) - \alpha \|f'(x)\|_2^2 + \bar{\bar{o}}(\alpha) \leq$$

For suff. small $\alpha > 0$

$$\leq f(x) - \alpha \|f'(x)\|_2^2 + \frac{\alpha}{2}\|f'(x)\|_2^2 = f(x) - \frac{\alpha}{2}\|f'(x)\|_2^2$$
$$< f(x)$$

Step of Gradient Descent:

$$\boxed{x^+ = x - \alpha f'(x)}$$

How to choose $\alpha > 0$? $\Rightarrow$ <u>Smoothness</u>

<u>Outline</u>:

1. Dual norms
2. Smoothness
3. Gradient Method for General Norms.

Dual Norm

Let $\|\cdot\|$ be fixed in $\mathbb{R}^n$ (an arbitrary).

Then, the dual norm

$$\|s\|_* = \max_{\|x\|\leq 1} \langle s, x\rangle = \max_{\|x\|=1} \langle s,x\rangle, \quad s\in\mathbb{R}^n$$

The Cauchy-Schwartz: $\forall x, s\in\mathbb{R}^n$

$$|\langle s,x\rangle| \leq \|s\|_* \cdot \|x\|.$$

Ex. 1 $\|x\| := \|x\|_2 = \sqrt{\langle x,x\rangle}$ Euclidean norm.

Then, $\|s\|_* = \|s\|_2$ (classic Cauchy-Schwartz)

Ex. 2. $\|x\| := \sqrt{\langle Bx,x\rangle}$, $B = B^T \succ 0$.

$$\|s\|_* = \langle B^{-1}s, s\rangle^{1/2}.$$

Ex. 3. $\|x\| := \|x\|_p = \left(\sum_{i=1}^n |x^{(i)}|^p\right)^{\frac{1}{p}}$, $p\geq 1$

$$\|x\|_\infty = \max_{1\leq i\leq n} |x^{(i)}|.$$

$\|s\|_* = \|x\|_q$, where $q\geq 1$ s.t. $\frac{1}{p}+\frac{1}{q}=1$.

$\|\cdot\|_\infty \longleftrightarrow \|\cdot\|_1$

$$Df(x)[\cdot] = \langle f'(x), \cdot \rangle$$

$$|\langle f'(x), y-x \rangle| \leq \|f'(x)\|_* \cdot \|y-x\|.$$

Smooth Functions

We call f "smooth" $\Leftrightarrow$ ∇f is Lipschitz:

$\forall x, y \in \mathbb{R}^n$:

$$\|f'(y) - f'(x)\|_* \leq L\|y-x\|.$$

$L > 0$ - Lipschitz constant.

If $x \approx y$ then $f'(y) \approx f'(x)$.

<u>Theorem</u> Let $f: \mathbb{R}^n \to \mathbb{R}$ is twice continuosly diff.

Then, the following cond. are equivalent:

(1) $\|f'(y) - f'(x)\|_* \leq L\|y-x\| \quad \forall x, y \in \mathbb{R}^n$

(2) $\|f''(x)\| \leq L \quad \forall x \in \mathbb{R}^n$.

<u>Remark.</u> For matrix $A \in \mathbb{R}^{n \times n}$: $(x,y) \mapsto \langle Ax, y \rangle$

$$\|A\| := \max_{\|x\| \leq 1} \|Ax\|_* = \max_{\substack{\|x\| \leq 1 \\ \|y\| \leq 1}} \langle Ax, y \rangle.$$

Remark If the norm is Euclidean:

$$\|f''(x)\| \leq L \iff -L \leq \lambda_i(f''(x)) \leq L.$$

Proof. (1) => (2).

Choose h: $\|h\| = 1$. $\varepsilon > 0$.

$$f'(x+\varepsilon h) = f'(x) + \varepsilon f''(x)h + \bar{\bar{o}}(\varepsilon)$$

$$\|f''(x)h\|_* = \left\|\frac{1}{\varepsilon}\left[f'(x+\varepsilon h) - f'(x) + \bar{\bar{o}}(\varepsilon)\right]\right\|_* \leq$$

$$\leq \frac{1}{\varepsilon}\|f'(x+\varepsilon h) - f'(x)\|_* + \bar{\bar{o}}(1)$$

$$\leq \frac{1}{\varepsilon}L\|\varepsilon h\| + \bar{\bar{o}}(1) = L + \bar{\bar{o}}(1) \xrightarrow[\varepsilon \to 0.]{} L$$

$$\Rightarrow \|f''(x)\| \leq L.$$

(2) => (1).

$$\|f'(y) - f'(x)\|_* = \left\|\int_0^1 f''(x+\tau(y-x))(y-x)\,d\tau\right\|_* \leq$$

$$\leq \int_0^1 \|f''(x+\tau(y-x))(y-x)\|_*\,d\tau$$

$$\leq \int_0^1 \underbrace{\|f''(x+\tau(y-x))\|}_{\leq L} \cdot \|y-x\|\,d\tau$$

$$\leq L \cdot \|y-x\|. \qquad \square$$

Ex. (Univariate Functions)

Smooth:

1. $f(x) = a + bx + cx^2$
2. $f(x) = \sin x$
3. $f(x) = \ln(1+e^x)$

Non-smooth:

1. $f(x) = |x|^3$
2. $f(x) = e^x$.

Ex. Quadratic Funct.

$$f(x) = \frac{1}{2}\langle Ax, x\rangle - \langle b, x\rangle, \qquad A = A^T \in \mathbb{R}^{n\times n}, \quad b \in \mathbb{R}^n.$$

$$f'(x) = Ax - b$$

$$f''(x) = A$$

This function is smooth, $L = \max\{\lambda_{max}(A), -\lambda_{min}(A)\}$
(for Euclidean norm)

Theorem (Global Model of Function).

Let $f:\mathbb{R}^n \to \mathbb{R}$ have Lipschitz gradient with $L>0$.
Then:

$$|f(y)-f(x)-\langle f'(x), y-x\rangle| \le \frac{L}{2}\|y-x\|^2 \quad \forall x,y\in\mathbb{R}^n.$$

Proof.

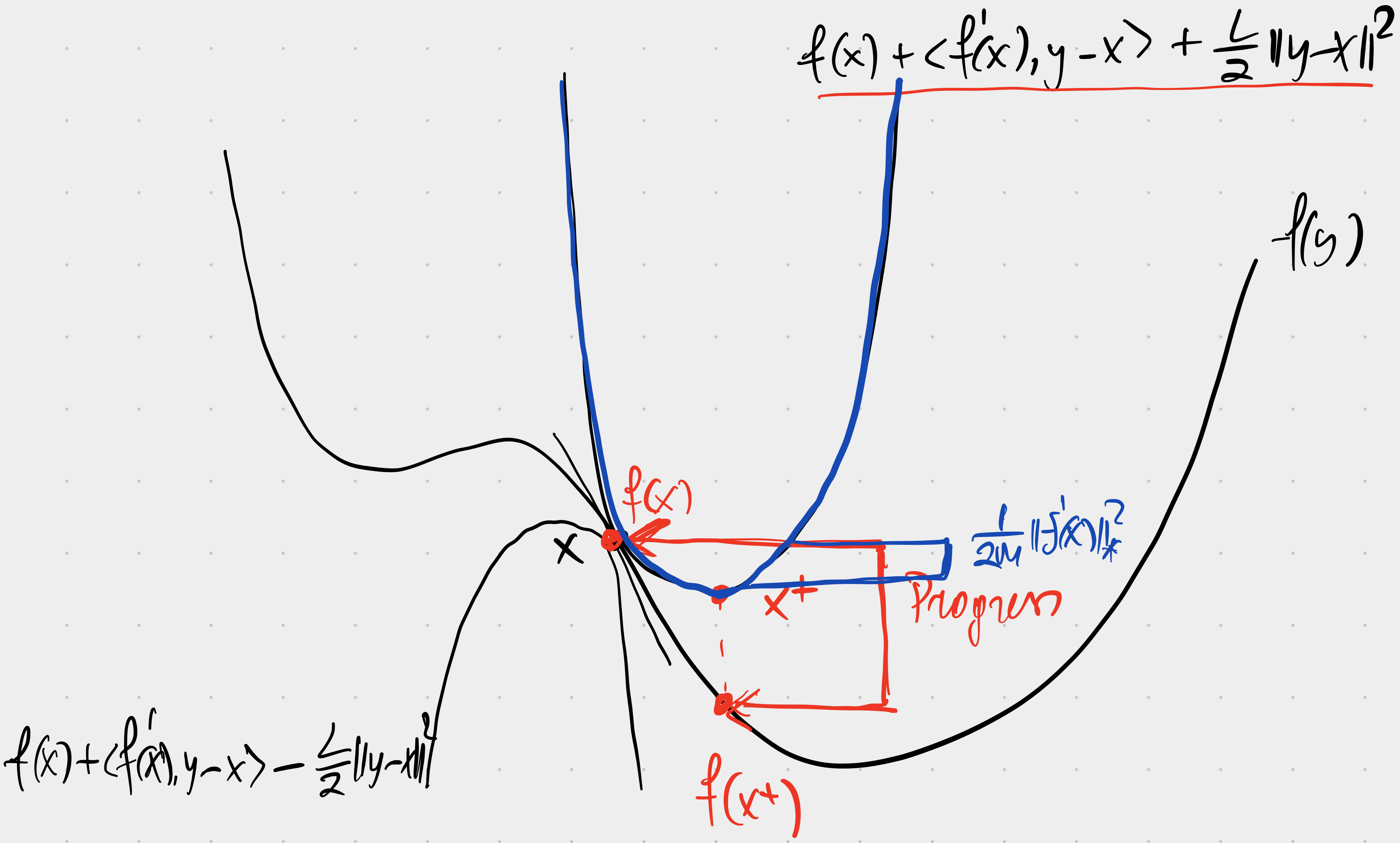
$$|f(y)-f(x)-\langle f'(x), y-x\rangle| = \left|\int_0^1 \langle f'(x+\tau(y-x))-f'(x), y-x\rangle d\tau\right|$$

$$\le \int_0^1 |\langle f'(x+\tau(y-x))-f'(x), y-x\rangle| d\tau$$

$$\le \int_0^1 \|f'(x+\tau(y-x))-f'(x)\|_* \cdot \|y-x\| d\tau$$

$$\le \int_0^1 L\cdot\|\tau(y-x)\|\cdot\|y-x\| d\tau = \int_0^1 \tau d\tau \cdot L\cdot\|y-x\|^2 = \frac{L}{2}\|y-x\|^2.$$

$\square$



Gradient Step

x, fix $M>0$:

$$\Omega_M(x;y) = f(x) + \langle f'(x), y-x\rangle + \frac{M}{2}\|y-x\|^2$$

M - regularization parameter.

Remark: For $M \geq L$: $f(y) \leq \Omega_M(x;y)$ $\forall x,y \in \mathbb{R}^n$.

$$x^+ = x^+(M) = \arg\min_{y\in\mathbb{R}^n} \Omega_M(x;y).$$

Algorithm (Gradient Method)

Initialization: $x_0 \in \mathbb{R}^n$, $\varepsilon > 0$.

For $k \geq 0$ iterate:

1. If $\|f'(x_k)\|_* \leq \varepsilon$ then
   return x_k.

2. Choose $M_k > 0$.

3. Perform the gradient step:

$$x_{k+1} = x_k^+(M_k) = \arg\min_{y\in\mathbb{R}^n} \left\{ f(x_k) + \langle f'(x_k), y - x_k\rangle + \frac{M_k}{2}\|y-x_k\|^2 \right\}$$

Choose M_k

1. Constant choice, $M_k \equiv L$
2. Adaptive search.
3. Normalized directions (later in the course)

Remark x_k^+ is not unique.

Ex. **Euclidean** $\|\cdot\| = \|\cdot\|_2$:

$$x^+ = \arg\min_{y \in \mathbb{R}^n} \left[f(x) + \langle f'(x), y-x \rangle + \frac{M}{2}\|y-x\|_2^2 \equiv g(y) \right]$$

$$g'(y) = f'(x) + M(y-x)$$

$$g'(x^+) = 0 \quad \Rightarrow \quad \boxed{x^+ = x - \frac{1}{M} f'(x).}$$

Ex. General Norm.

$$\min_{y \in \mathbb{R}^n} \Omega_M(x;y) \equiv f(x) + \langle f'(x), y-x \rangle + \frac{M}{2}\|y-x\|^2 =$$

$y - x = \tau h$, where $\|h\| = 1$, $\tau > 0$

$$= \min_{\tau > 0} \min_{\|h\|=1} \left[f(x) + \tau \langle f'(x), h \rangle + \frac{M}{2}\tau^2 \right] =$$

$$\min_{\|h\|=1} \langle f'(x), h \rangle = -\max_{\|h\|=1} -\langle f'(x), h \rangle = \underline{-\max_{\|h\|=1} \langle f'(x), h \rangle = -\|f'(x)\|_*}.$$

$$= \min_{\tau > 0} \left[f(x) - \tau \|f'(x)\|_* + \frac{M}{2}\tau^2 \right] =$$

$$M\tau^+ = \|f'(x)\|_* \quad \Rightarrow \quad \boxed{\tau^+ = \frac{\|f'(x)\|_*}{M}}$$

$$= f(x) - \frac{\|f'(x)\|_*^2}{2M}. \qquad \boxed{h^+ \in \mathbb{R}^n : \langle f'(x), h^+ \rangle = -\|f'(x)\|_*}$$

$$x^+ = x + \tau^+ h^+.$$

Analysis. "Descent Lemma".

<u>Proposition</u> Let $M \geq L$. Then $x^+ = x^+(M)$:

$$f(x) - f(x^+) \geq \frac{1}{2M} \|f'(x)\|_*^2 . \qquad (\star)$$

<u>Proof</u> By smoothness:

$$f(x^+) \leq f(x) + \langle f'(x), x^+ - x \rangle + \frac{L}{2}\|x^+ - x\|^2 \leq$$

$$\leq f(x) + \langle f'(x), x^+ - x \rangle + \frac{M}{2}\|x^+ - x\|^2 =$$

$$= \Omega_M(x; x^+) = \Omega_M^*(x) =$$

$$= f(x) - \frac{1}{2M}\|f'(x)\|_*^2 . \qquad \square$$

<u>Theorem</u> Let f be Bounded Below: $f^* = \inf\limits_{x \in \mathbb{R}^n} f(x) > -\infty$.

Consider iterates of Gradient Method:

$$x_{k+1} = x_k^+(M_k), \qquad k \geq 0$$

Assume, $M_k \leq M_*$ $\forall k \geq 0$ and $(\star)$ is satisfied.

(For example: $M_k \equiv L$). Then,

$$\frac{2M_*(f(x_0) - f^*)}{K} \geq \frac{1}{k}\sum_{i=0}^{k-1} \|f'(x_i)\|_*^2 \geq \min_{0 \leq i \leq k-1} \|f'(x_i)\|_*^2 .$$

<u>Proof.</u>

For every iteration,

$$f(x_k) - f(x_{k+1}) \geq \frac{1}{2M_k}\|f'(x_k)\|_*^2 \geq \frac{1}{2M_*}\|f'(x_k)\|_*^2$$

Telescope:

$$f(x_0) - f^* \geq f(x_0) - f(x_k) \geq \frac{1}{2M_*} \sum_{i=0}^{k-1} \|f'(x_i)\|_*^2 . \quad \square$$

<u>Complexity</u> let $M_k \equiv L$. Fix $\varepsilon > 0$:

$$\|f'(\bar{x})\|_* \leq \varepsilon .$$

If $\|f'(x_i)\|_* > \varepsilon$ $\quad i = 0 \ldots k-1$: then:

$$\frac{2LF_0}{k} > \varepsilon^2 . \quad \Leftrightarrow \quad k < \frac{2LF_0}{\varepsilon^2} .$$

$F_0 = f(x_0) - f^*$

<u>Conclusion</u>: to find $\|f'(\bar{x})\|_* \leq \varepsilon$

It's sufficient to perform

$$K = \left\lfloor \frac{2LF_0}{\varepsilon^2} \right\rfloor$$

first-order oracle calls.

$\varepsilon \approx 10^{-3}$

$\Downarrow$

$K \approx 10^6$

iterations to solve the problem.