Claim: $-\varepsilon \dfrac{\nabla f(x_k)}{\|\nabla f(x_k)\|} = \underset{s}{\operatorname{argmin}} \{ \nabla f(x_k)^T s \mid \|s\|_2 \le \varepsilon \}$

Proof. Suppose for contradiction that s' does better

$$\nabla f(x_k)^T s' < \nabla f(x_k)^T \left(\frac{-\varepsilon \nabla f(x_k)}{\|\nabla f(x_k)\|_2} \right)$$
$$= -\varepsilon \|\nabla f(x_k)\|_2$$

By Cauchy-Schwartz,

$$-\|\nabla f(x_k)\|_2 \|s'\|_2 < -\varepsilon \|\nabla f(x_k)\|_2$$

$\Rightarrow \|s'\|_2 > \varepsilon$.

$\Rightarrow s'$ is not feasible. $\square$

$\Rightarrow x_{k+1} = x_k - \alpha_k \nabla f(x_k)$ for some α_k.

"Gradient Descent".

Starting 9/9

HW1 due before the next lecture.

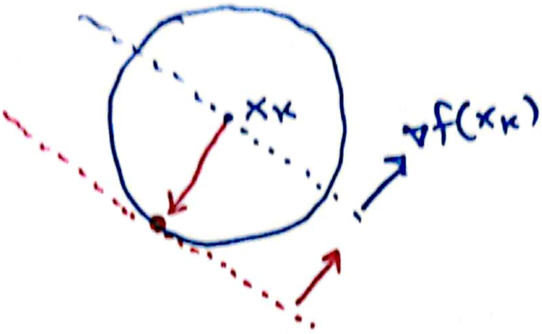
Submit on time. I will release my solutions and then cannot accept yours...

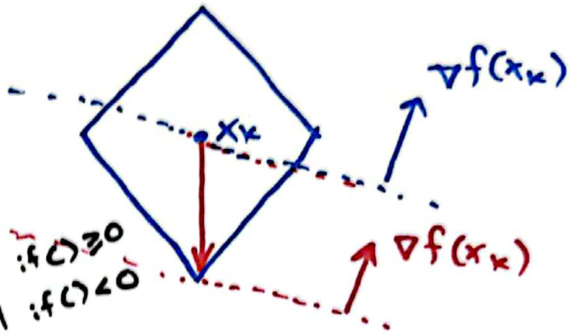
Approach 3 We could consider other norms.

We could look for best improvement with

$$\|s\|_1 \le \varepsilon$$

$$\sum_{i=1}^{d} |s_i|$$





$e_j = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}$ 1 at the j^{th} position

$\operatorname{sign}(\cdot) = \begin{cases} 1 & \text{if } (\,) \ge 0 \\ -1 & \text{if } (\,) < 0 \end{cases}$

Claim $s = -\varepsilon e_j \operatorname{sign}\left(\frac{\partial f}{\partial x_j}(x_k)\right) \in \operatorname{argmin}\{\nabla f(x_k)^T s \mid \|s\|_1 \le \varepsilon\}$.

where $j \in \operatorname{argmax}\{|\frac{\partial f}{\partial x_j}(x_k)|\}$

Proof. Suppose s' does better:

$$\begin{aligned} \nabla f(x_k)^T s' < \nabla f(x_k)^T s &= -\varepsilon \frac{\partial f}{\partial x_j}(x_k) \operatorname{sign}\left(\frac{\partial f}{\partial x_j}(x_k)\right) \\ &= -\varepsilon \left|\frac{\partial f}{\partial x_j}(x_k)\right| \\ &= -\varepsilon \|\nabla f(x_k)\|_\infty \qquad \left(\|y\|_\infty = \max |y_i|\right) \end{aligned}$$

By Hölder's Inequality

$$-\|\nabla f(x_k)\|_\infty \|s'\|_1 < -\varepsilon \|\nabla f(x_k)\|_\infty$$

$$\Rightarrow \|s'\|_1 > \varepsilon.$$

$\Rightarrow$ s' is not feasible. $\square$

Scanned with CamScanner

This motivates "Coordinate Descent"

$$x_{k+1} = x_k - \alpha_k e_j \ \text{sign}\left(\frac{\partial f}{\partial x_j}(x_k) \right)$$

the coordinate with largest partial derivative.

If $x_0 = 0$, $\quad x_k$ is k-sparse ("k nonzeros").

Second-Order Local Improvements

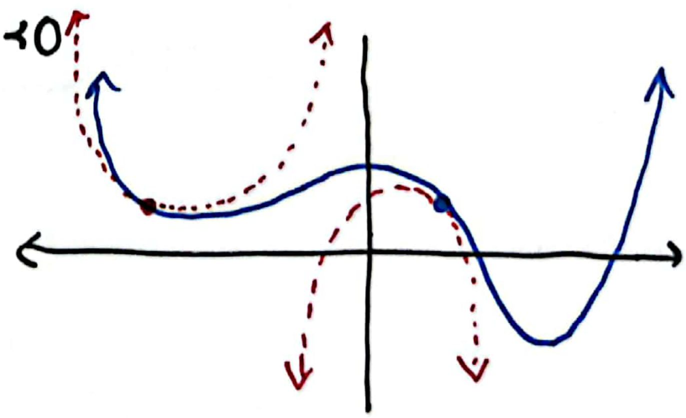
$$f(x_k + s) \approx f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s$$

Approach 4

$$\min_{s \in \mathbb{R}^d} \left(f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s \right)$$

$$= \begin{cases} -\infty & \text{if } \nabla^2 f(x_k) \prec 0 \\ ??? & \end{cases}$$

promising



If $\nabla^2 f(x_k) \succ 0$, then

$$x_{k+1} = x_k - \nabla^2 f(x_k)^{-1} \nabla f(x_k)$$

"Newton's Method" for $\nabla f(x_k) = 0$

Approach 5 Trust-Region Methods

$$x_{k+1} = x_k + \underset{\|s\| \le \varepsilon}{\operatorname{argmin}} \left\{ f(x_k) + \nabla f(x_k)^T s + \frac{1}{2} s^T \nabla^2 f(x_k) s \right\}$$

How can we guarantee good quality?

Taylor Approximation Theorems.

Gradient Descent / Smooth Optimization

4 lectures

1. A Descent Lemma
2. Stepsizes / Linesearches
3. Nonconvex Smooth Opt Guarantees
4. Shape of Smooth Convex Funcs
5. Better Guarantees
6. Complexity Lowerbounds
7. Acceleration

1. A Descent Lemma

We defined gradient descent (GD) as

$$x_{k+1} = x_k - \alpha_k \nabla f(x_k)$$

for some $x_0 \in \mathbb{R}^d$ and $\alpha_k \in \mathbb{R}$.

<u>Lemma</u> For any f with L-Lipschitz gradient, any $k \geq 0$ has

$$f(x_{k+1}) \leq f(x_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) \|\nabla f(x_k)\|_2^2 .$$

<u>Consequences</u>: Decrease whenever $\alpha_k - \frac{L\alpha_k^2}{2} > 0$

$$\Leftrightarrow \alpha_k < 2/L .$$

Better descent when L small
$\|\nabla f(x_k)\|_2$ large.

Best decrease when $\alpha_k = 1/L$
of $\frac{-1}{2L} \|\nabla f(x_k)\|_2^2$

(Taylor Approx Theorem)

Proof. $\left| f(x_{k+1}) - \left(f(x_k) + \nabla f(x_k)^T (x_{k+1} - x_k)\right)\right| \leq \frac{L}{2} \|x_{k+1} - x_k\|_2^2$

- dropping $|\cdot|$
- plug in $x_{k+1} - x_k = -\alpha_k \nabla f(x_k)$

$$\Rightarrow f(x_{k+1}) - f(x_k) - \alpha_k \|\nabla f(x_k)\|_2^2 \leq \frac{L\alpha_k^2}{2} \|\nabla f(x_k)\|_2^2$$

$$\Rightarrow f(x_{k+1}) \leq f(x_k) - \left(\alpha_k - \frac{L\alpha_k^2}{2}\right) \|\nabla f(x_k)\|_2^2 .$$

$\square$

2. Stepsize Choice / Linesearch

Based on our lemma, $\alpha_k = 1/L$ gives

$$\frac{1}{2L} \|\nabla f(x_k)\|_2^2 .$$

Impractical. We do not often know L.

Exact Linesearch Pick the best α_k

$$\alpha_k = \underset{\alpha \in \mathbb{R}}{\operatorname{argmin}} \ f(x_k - \alpha \nabla f(x_k))$$

Trivially outperforms $\alpha_k = 1/L$ since

$$\begin{aligned}
\text{(namely, } f(x_{k+1}) &\le f(x_k - \alpha \nabla f(x_k)) \quad \forall \alpha \\
&\le f(x_k - \tfrac{1}{L} \nabla f(x_k)) \\
&\le f(x_k) - \tfrac{1}{2L} \|\nabla f(x_k)\|_2^2 \quad \text{by Descent Lemma)}
\end{aligned}$$

Impractical. Slowdown our iteration.

Backtracking Linesearch Pick $\alpha \in \mathbb{R}$, $\tau \in (0,1)$

$$\alpha_k = \sup \{ \alpha \tau^n \mid n = 0, 1, 2, \ldots \quad f(x_k - \alpha \tau^n \nabla f(x_k)) \text{ “}<\text{” } f(x_k) \}$$

"Keep backing off exponentially until decrease is found."

Does this terminate? Yes!

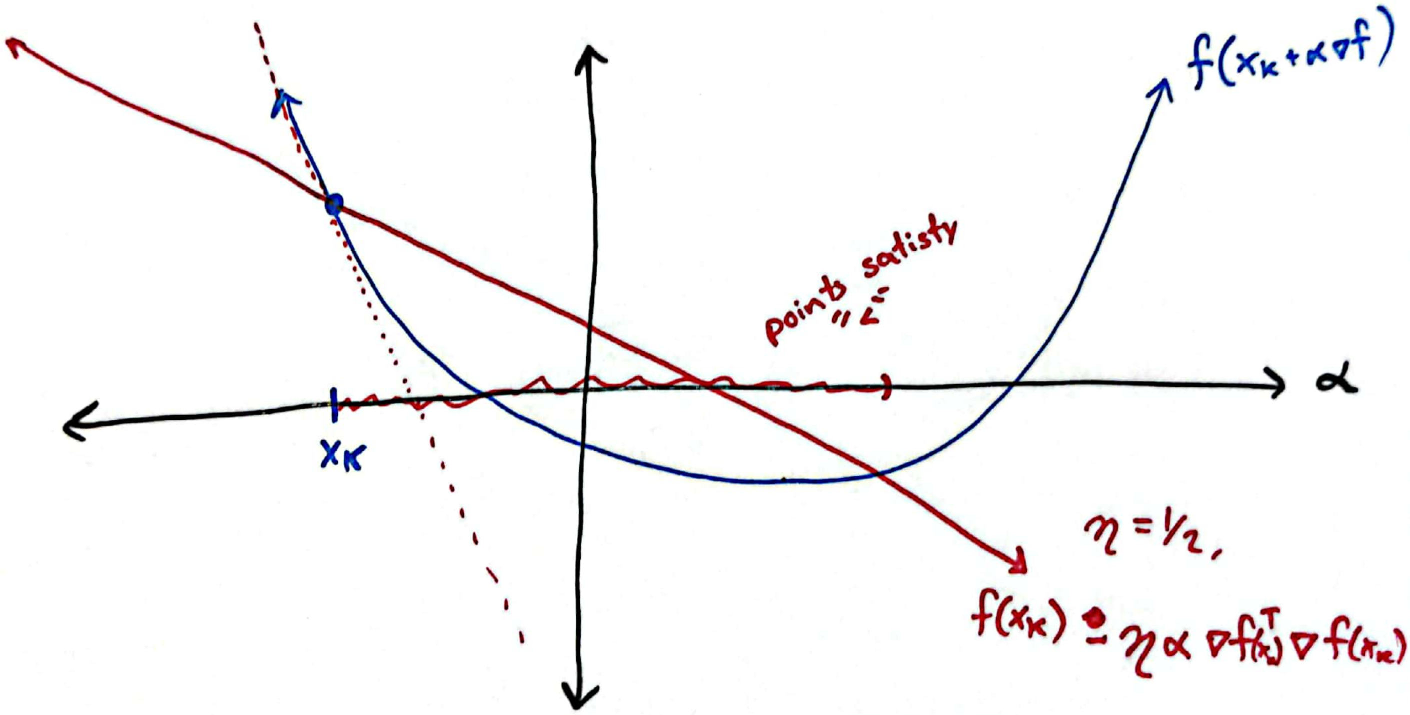
Descent Lemma ensures decrease $\alpha \in (0, \frac{2}{L})$.

Only $\left\lceil \log_{1/2}\left(\frac{\alpha}{2/L} \right) \right\rceil$ backtracking steps.

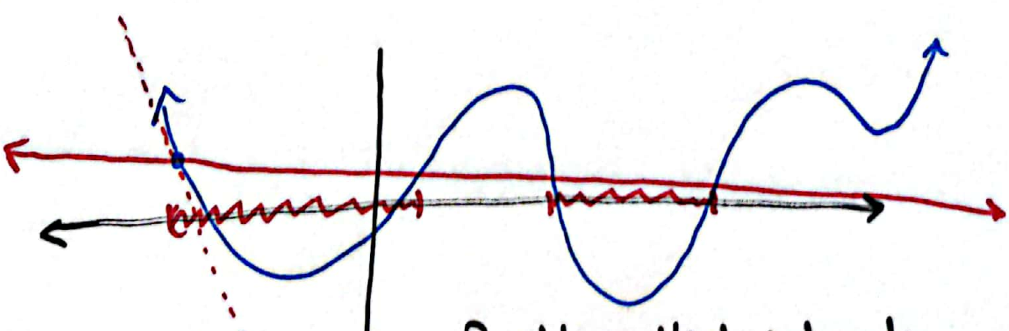
How should we measure descent? What is "<"?

Pick $\eta \in (0,1)$, we want

"Armijo Condition" $\quad f(x_k - \alpha_k \nabla f(x_k)) \leq f(x_k) - \eta \alpha_k \|\nabla f(x_k)\|_2^2$



$\eta = 1$, $f(x_k) - \eta \alpha \nabla f(x_k)^T \nabla f(x_k)$



Possible multiple intervals.

Lemma The Armijo Condition holds for

$$\alpha \in \left[0, \frac{2(1-\eta)}{L}\right].$$