Class #18

Tuesday, April 1, 2025 10:20 AM

Dual norm

Let $\|\cdot\|$ be any norm on $\mathbb{R}^n$

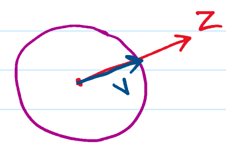
$\|\cdot\|_*$ is a dual norm of $\|\cdot\|$, defined on $\mathbb{R}^n$ as follows:

$$\forall\, z \in \mathbb{R}^n, \quad \|z\|_* = \sup\{ z^T v : \|v\| \le 1 \}$$

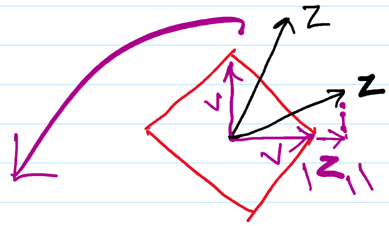
How much can the vector z inflate v when $\|v\| \le 1$?

$$z^T v \le \|z\|_* \|v\|$$

E.g., $\|\cdot\| = \|\cdot\|_2$ (l_2 norm)



$$\sup(z^T v) = \|z\|_2$$



Say, $\|\cdot\| = \|\cdot\|_1$

$z^T v = |z_1|$ $\qquad$ $z^T v = |z_2|$

$\|\cdot\|_*$ when $\|\cdot\| = \|\cdot\|_1 \Rightarrow \|\cdot\|_* = \|\cdot\|_\infty$

Dual norms of $\|\cdot\|_p$ when:

$p = 2 \Rightarrow \|\cdot\|_* = \|\cdot\|_2$

$p = 1 \Rightarrow \|\cdot\|_* = \|\cdot\|_\infty$

$p = \infty \Rightarrow \|\cdot\|_* = \|\cdot\|_1$

Dual norm of l_p is l_q norm

where $\frac{1}{p} + \frac{1}{q} = 1 \iff q = \frac{p}{p-1}$

$$(\|\cdot\|_*)_* = \|\cdot\|$$

Steepest Descent

Assume $f(x)$ attains its minimum

$$x^* \in \underset{x \in \text{dom} f}{\arg\min} \; f(x)$$

Descent method : $x^{(k+1)} = x^{(k)} + t^{(k)} \Delta x^{(k)}$

where $\Delta x^{(k)}$ is a descent direction

i.e. $-\nabla f(x^{(k)})^T \Delta x^{(k)} > 0$

Remember : First-order approx. of f around $x^{(k)}$

$$f(x^{(k+1)}) \simeq f(x^{(k)}) + t^{(k)} \nabla f(x^{(k)})^T \Delta x^{(k)}$$

when $\Delta x^{(k)}$ is a descent direction

$\Longrightarrow f(x^{(k+1)}) < f(x^{(k)})$

The reduction in function value $\propto$ $\boxed{\nabla f(x^{(k)})^T \Delta x^{(k)}}$

$\Longrightarrow f(x^{(k+1)}) < f(x^{(k)})$

The reduction in function value $\propto$ $\boxed{\nabla f(x^{(k)})^T \Delta x^{(k)}}$

Steepest descent is the descent method in which we have the most reduction in the objective function value (i.e., $\nabla f(x)^T \Delta x$ is the smallest)

↳ i.e. $-\nabla f(x)^T \Delta x$ is the largest value.

This is a non-rigorous statement.

we could ask $\Delta x_{sd} = \underset{v \in \mathbb{R}^n}{\arg\min} \left\{ \nabla f(x)^T v \right\}$

$$= \underset{v \in \mathbb{R}^n}{\arg\max} \left\{ -\nabla f(x)^T v \right\}$$

But this would always return $\Delta x_{sd} \to -\begin{bmatrix} \infty \\ \infty \\ \vdots \\ \infty \end{bmatrix}$

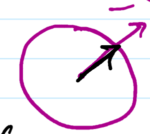
Better idea: Normalized steepest descent direction:

$$\Delta x_{nsd} = \underset{v \in \mathbb{R}^n}{\arg\min} \left\{ \nabla f(x)^T v : \|v\| \le 1 \right\}$$

$\|\cdot\| = \|\cdot\|_2$

$$= \arg \underbrace{\left(\max_{v \in \mathbb{R}^n} \left\{ -\nabla f(x)^T v : \|v\| \le 1 \right\} \right)}$$

$-\nabla f(x)$



Dual norm of $-\nabla f(x)$

$= \nabla f(x)$

$\Delta x_{nsd} = \dfrac{-\nabla f(x)}{\|\nabla f(x)\|}$

$= \|\nabla f(x)\|_*$

$$\Delta x_{nsd} = \frac{-\nabla f(x)}{\|\nabla f(x)\|_2} \qquad = \|\nabla f(x)\|_*$$

$$\Delta x_{sd} = \|\nabla f(x)\|_* \, \Delta x_{nsd} \Rightarrow \text{Steepest descent direction}$$

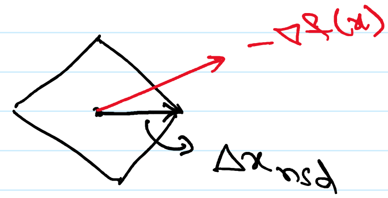
Steepest descent iteration: $x^{(k+1)} = x^{(k)} + t^{(k)} \underbrace{\|\nabla f(x^{(k)})\|_* \Delta x^{(k)}_{nsd}}$

When $\|\cdot\| = \|\cdot\|_2 \Rightarrow \|\cdot\|_* = \|\cdot\|_2$ $\qquad \cancel{\|\nabla f(x^{(k)})\|_2} \cdot \frac{-\nabla f(x^{(k)})}{\cancel{\|\nabla f(x^{(k)})\|_2}}$

$$\Delta x_{nsd} = -\frac{\nabla f(x)}{\|\nabla f(x)\|_2} \Rightarrow$$

$\Rightarrow$ SD = GD when we are using $\|\cdot\|_2$.

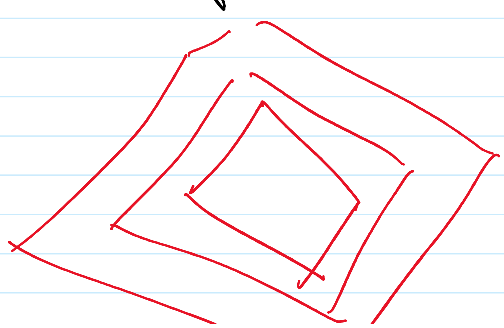
E.g., $\|\cdot\| = \|\cdot\|_1$

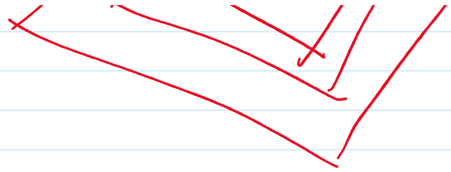


In that case, the direction is coordinate in which $-\frac{\partial f}{\partial x_i}$ is the largest.

The norm that one should use in steepest descent should be such that norm-ball has geometry that matches the geometry of the level sets.

Say:





When the norm is correct: SD converges faster than GD.
When the guess is wrong: SD converges slower than GD.

Special Case of Steepest Descent: Use P-quadratic norm, where $P \succ 0$. $P \in \mathbb{R}^{n \times n}$

Reminder: $\|v\|_P = (v^T P v)^{1/2} = \|P^{1/2} v\|_2$

Also called Gradient descent with Preconditioning.
P is called the preconditioning matrix.

Facts: $\|z\|_*$ when $\|\cdot\| = \|\cdot\|_P$

$$\|z\|_* = \|P^{-1/2} z\|_2 = (z^T P^{-1} z)^{1/2}$$

$$\Delta x_{nsd} = \frac{-P^{-1} \nabla f(x)}{\|\nabla f(x)\|_*}$$

$$= -\left(\nabla f(x)^T P^{-1} \nabla f(x)\right)^{-1/2} P^{-1} \nabla f(x)$$

$$\Delta x_{sd} = \|\nabla f(x)\|_* \Delta x_{nsd} = -P^{-1} \nabla f(x)$$

Bottom line: SD with $\|\cdot\|_P$ is

$$x^{(k+1)} = x^{(k)} - t^{(k)} P^{-1} \nabla f(x) \quad \text{——} \; (*)$$

$GD \Rightarrow P = I$;

Newton's method $\Rightarrow P = \nabla^2 f(x)$

Steepest descent under P-quadratic norm is equivalent to doing a change of coordinates of our problem by $P^{-1/2}$ and then running gradient descent.

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

Let $\bar{x} = P^{1/2} x$; $\bar{f}(\bar{x}) = \boxed{f(P^{-1/2} x)}$

$\Leftrightarrow x = P^{-1/2} \bar{x}$

$$\nabla_{\bar{x}} \bar{f}(\bar{x}) = P^{-1/2} \nabla f(P^{-1/2} x)$$

$$x^{(k+1)} = x^{(k)} + t^{(k)} \underbrace{\nabla f(x^{(k)})^T}_{} \underbrace{\Delta x_{sd}^{(k)}}_{}$$

$$\nabla f^T P^{-1/2} \quad \nabla f(x) = P^{-1/2} \nabla f$$

$$\nabla f^T P^{-1} \nabla f = \|\nabla f(x)\|_*^2$$

Special Case: $f(x) = x^T Q x \Rightarrow \operatorname{cond}(f(x)) = \frac{\lambda_{max}(Q)}{\lambda_{min}(Q)}$

condition number of $f(x)$ is determined by eigenvalues of Q.

Pick $= P = Q$

$$\Rightarrow P^{-1/2} = Q^{-1/2}$$

$$\bar{f}(\bar{x}) = f(P^{-1/2}x) = (P^{-1/2}x)^T Q (P^{-1/2}x)$$

$$= x^T \underbrace{P^{-1/2} Q P^{-1/2}}_{= I} x$$

$$= x^T x$$

$$\text{cond}(\bar{f}(\bar{x})) = 1$$