Muon Appendix

A. Some Norms are Induced by Others

A.1. Operator Norm (Induced Norm)

- Consider a linear transformation $x \xmapsto{A} Ax$. How do we measure the 'size' of A?
- Here, it's important to note that A is an operator on x.
- In other words, A is defined on what it does to x, so the norm of A should be defined by how much it changes x.
- Given that input x is measured by $||\cdot||_\alpha$ and output Ax is measured by $||\cdot||_\beta$, the norm of A is defined by the maximum change of norm between the input and output (in their own norms):

$$||A||_{\alpha\to\beta} \triangleq \sup_{x\neq 0} \frac{||Ax||_\beta}{||x||_\alpha}$$

- We call the norm $||\cdot||_{\alpha\to\beta}$ the "**operator norm**" or the "**α-to-β induced norm**".
- For example, if both input and output use ℓ_2 norm, then the induced norm is the Spectral norm (i.e., the largest singular value of A). Recall what eigenvalues/singular values meant!

$$||A||_{\ell_2\to\ell_2} \triangleq \sup_{x\neq 0} \frac{||Ax||_{\ell_2}}{||x||_{\ell_2}} = \sigma_{\max}$$

A.2. Dual Norm

- Dual norm is a special case of operator norm where A is a vector rather than a matrix.
- Consider a linear transformation $x \xmapsto{a} a^T x$. How do we measure the size of a?
- Again, note that a is an operator on x. So even though a has the same shape as x, **they live in a completely different space**. Specifically, we call the space of x the primal space (input space), and a the dual space (operator space).
- And following the same flow as above, the norm of a is defined by how much it changes x.
- Given that input x is measured by $||\cdot||_\alpha$, the norm of a is defined by:

$$||a||_\alpha^\dagger \triangleq \sup_{x\neq 0} \frac{a^T x}{||x||_\alpha}$$

- We call the norm $||a||_\alpha^\dagger$ a "**dual norm to α**".
- For example: a dual norm of ℓ_p norm is the ℓ_q norm, where $1/p + 1/q = 1$.
    - Notably, the dual of ℓ_2 is also ℓ_2. This means that if we use ℓ_2 norm, the dual space is identical to the primal space. This is exactly the case when we use gradient descent!

B. Steepest Descent Derivations

B.1. Steepest Descent Problem

- A steepest descent problem has three inputs:
  - gradient g
  - sharpness parameter λ
  - norm $|| \cdot ||$
- And outputs the steepest descent direction (and step size) under these conditions.

Proposition 1 (Steepest descent) *For any $\boldsymbol{g} \in \mathbb{R}^n$ thought of as "the gradient" and any $\lambda \geqslant 0$ thought of as "the sharpness", and for any norm $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ with dual norm $\|\cdot\|^\dagger$:*

$$\underset{\Delta \boldsymbol{w} \in \mathbb{R}^n}{\arg\min} \left[\boldsymbol{g}^\top \Delta \boldsymbol{w} + \frac{\lambda}{2} \|\Delta \boldsymbol{w}\|^2 \right] = -\frac{\|\boldsymbol{g}\|^\dagger}{\lambda} \cdot \underset{\|\boldsymbol{t}\|=1}{\arg\max}\, \boldsymbol{g}^\top \boldsymbol{t}. \tag{1}$$

- We can actually derive this problem using a Lipschitz condition (which connects to sharpness).
- Say we want to minimize $F(x)$, a closed convex function with Lipschitz condition on its gradient:

$$||\nabla F(x) - \nabla F(y)||_q \leq L_p ||x - y||_p$$

- ...where p and q are dual norms (see Appendix A).
- Let's start by drawing a line between arbitrary points x, y.

$$\text{Let } g(t) = F(x + t(y - x)), \text{where } t \in [0, 1]$$

- Note that $g(0) = F(x), g(1) = F(y)$.
- Then:

$$\begin{gathered}
F(y) - F(x) \\
= \int_0^1 \langle \nabla F(x + t(y - x)), y - x \rangle dt \\
= \int_0^1 \langle \nabla F(x + t(y - x)) + \nabla F(x) - \nabla F(x), y - x \rangle dt \\
= \int_0^1 \langle \nabla F(x), y - x \rangle dt + \int_0^1 \langle \nabla F(x + t(y - x)) - \nabla F(x), y - x \rangle dt \\
= \langle \nabla F(x), y - x \rangle + \int_0^1 \langle \nabla F(x + t(y - x)) - \nabla F(x), y - x \rangle dt \\
\leq \langle \nabla F(x), y - x \rangle + \int_0^1 |\langle \nabla F(x + t(y - x)) - \nabla F(x), y - x \rangle| dt
\end{gathered}$$

▼ First equation can be understood easier with single variable case

$f(y) = f(x) + \int_x^y f'(t)dt$

$f(y) = f(x) + \int_0^1 f'(x + t(y-x))dt \cdot (y-x)$

$f(y) - f(x) = \int_0^1 f'(x + t(y-x))dt \cdot (y-x)$

- Also, $\langle a, b \rangle = a^T b$ for vectors, $\langle A, B \rangle = tr(AB)$ for matrices (which is also just elementwise multiplication and sum).
- By Holder's inequality ($|\langle a, b \rangle| \leq ||a||_p ||b||_q$, where p, q are dual):

$$\begin{gathered} |\langle \nabla F(x + t(y-x)) - \nabla F(x), y - x \rangle| \\ \leq ||\nabla F(x + t(y-x)) - \nabla F(x)||_q \cdot ||y - x||_p \\ \leq L_p ||x + t(y-x) - x||_p \cdot ||y - x||_p \\ = L_p \cdot t \cdot ||y - x||_p^2 \end{gathered}$$

- Where second to third was from the Lipschitz conditioned we started with.
- Plugging this back, we get:

$$\begin{gathered} F(y) - F(x) \\ \leq \langle \nabla F(x), y - x \rangle + \int_0^1 |\langle \nabla F(x + t(y-x)) - \nabla F(x), y - x \rangle| dt \\ \leq \langle \nabla F(x), y - x \rangle + \int_0^1 L_p \cdot t \cdot ||y - x||_p^2 dt \\ \leq \langle \nabla F(x), y - x \rangle + L_p ||y - x||_p^2 \cdot \int_0^1 t dt \\ \leq \langle \nabla F(x), y - x \rangle + \frac{L_p}{2} ||y - x||_p^2 \end{gathered}$$

- We thus end up with the steepest descent problem:

$$F(y) \leq F(x) + \langle \nabla F(x), y - x \rangle + \frac{L_p}{2} ||y - x||_p^2$$

- Or alternatively by letting $y = x + \Delta x$,

$$F(x + \Delta x) \leq F(x) + \langle \nabla F(x), \Delta x \rangle + \frac{L_p}{2} ||\Delta x||_p^2$$

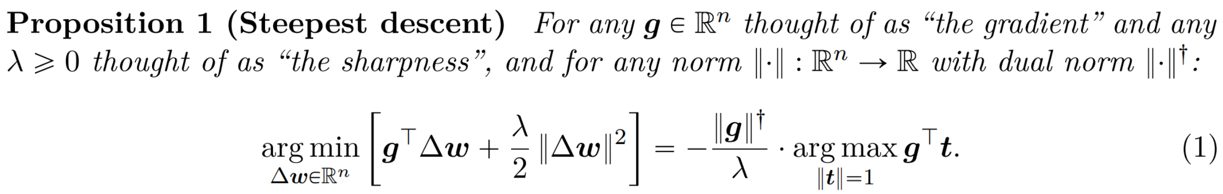
- Which becomes a minimization problem:

$$\Delta x = -\arg\max_{\Delta x} \left[\langle \nabla F(x), \Delta x \rangle + \frac{L_p}{2} ||\Delta x||_p^2 \right]$$

- The constant L_p connects to λ, which is natural since L_p tells us how fast the gradient changes (i.e., sharpness)! We will later see in B.3. that L_p actually represents the norm of the Hessian: $||\nabla^2 F(x)||_{p \to q} \leq L_p, \forall x$.

B.2. Solution to Steepest Descent Problem

- Here we'll fall back to the problem stated in Old Optimizer, New Norm: An Anthology.
  The derivation explained here is also directly from AppendixB.1. of that paper.



Proposition 1 (Steepest descent) *For any $\boldsymbol{g} \in \mathbb{R}^n$ thought of as "the gradient" and any $\lambda \geqslant 0$ thought of as "the sharpness", and for any norm $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$ with dual norm $\|\cdot\|^\dagger$:*

$$\underset{\Delta \boldsymbol{w} \in \mathbb{R}^n}{\arg\min} \left[\boldsymbol{g}^\top \Delta \boldsymbol{w} + \frac{\lambda}{2} \|\Delta \boldsymbol{w}\|^2 \right] = -\frac{\|\boldsymbol{g}\|^\dagger}{\lambda} \cdot \underset{\|\boldsymbol{t}\|=1}{\arg\max}\, \boldsymbol{g}^\top \boldsymbol{t}. \tag{1}$$

- Decouple Δw to direction and step size: $\Delta w = c \cdot t, c \in \mathbb{R}, t \in \mathbb{R}^n, c \geq 0, ||t|| = 1$

$$\min_{\Delta w \in \mathbb{R}^n} \left[g^T \Delta w + \frac{\lambda}{2} ||\Delta w||^2 \right] = \min_{c \geq 0} \min_{t \in \mathbb{R}^n : ||t|| = 1} \left[c \cdot g^T \cdot t + \frac{\lambda}{2} c^2 ||t||^2 \right]$$

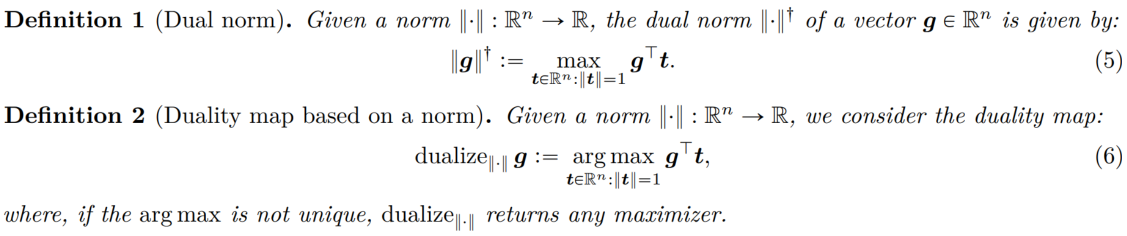
- Because $||t|| = 1$:

$$= \min_{c \geq 0} \left[c \cdot \min_{t \in \mathbb{R}^n : ||t|| = 1} [g^T t] + \frac{\lambda}{2} c^2 \right]$$

- By definition of dual norm:

$$= \min_{c \geq 0} \left[-c \cdot ||g||^\dagger + \frac{\lambda}{2} c^2 \right]$$

- Deriving Δw
  - From (2): $t = \arg\min_{||t||=1} [g^T t] = -\arg\max_{||t||=1} [g^T t]$
  - From (3): $c = \frac{||g||^\dagger}{\lambda}$ (solving quadratic eq)
  - Thus, $\Delta w = c \cdot t = -\frac{||g||^\dagger}{\lambda} \arg\max_{||t||=1} g^T t$
- Interestingly, the step size is made by the dual norm: $||g||^\dagger \;= \min_{||t||=1} [g^T t]$,
  but the direction is actually made identically but in argmax:
  $-\arg\max_{||t||=1} [g^T t]$
- So not only the norm is being dualized, the direction itself is also being dualized!
- In the later paper of Jeremy Bernstein, this operation will be called the dualize function.



Definition 1 (Dual norm). *Given a norm* $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$*, the dual norm* $\|\cdot\|^\dagger$ *of a vector* $\boldsymbol{g} \in \mathbb{R}^n$ *is given by:*

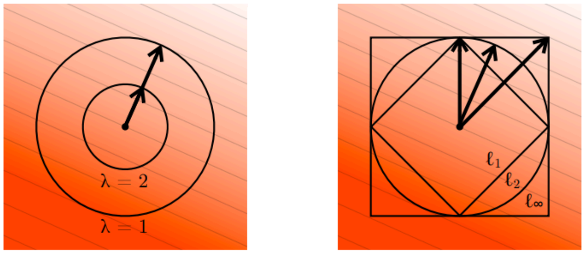
$$\|\boldsymbol{g}\|^\dagger := \max_{\boldsymbol{t}\in\mathbb{R}^n : \|\boldsymbol{t}\|=1} \boldsymbol{g}^\top \boldsymbol{t}. \tag{5}$$

Definition 2 (Duality map based on a norm). *Given a norm* $\|\cdot\| : \mathbb{R}^n \to \mathbb{R}$*, we consider the duality map:*

$$\operatorname{dualize}_{\|\cdot\|} \boldsymbol{g} := \operatorname*{arg\,max}_{\boldsymbol{t}\in\mathbb{R}^n : \|\boldsymbol{t}\|=1} \boldsymbol{g}^\top \boldsymbol{t}, \tag{6}$$

where, if the arg max *is not unique,* $\operatorname{dualize}_{\|\cdot\|}$ *returns any maximizer.*

- Which does not add another knowledge, but makes it clear that gradient live in a dual space, and that they must be translated back to primal space (parameter space) to behave properly.
- Obviously, the direction is mainly controlled by $|| \cdot ||$, and the size mainly by λ.



a) varying sharpness λ

b) varying choice of norm ‖·‖

- For how λ is decided, I have no clear understanding but it may be answered by the sensitivity parameter they define later [1], [2].

B.3. Connection to Hessian

- Start with univariate function $\mathbb{R} \to \mathbb{R}$:

$$f(y) = f(x) + \int_x^y f'(t)dt$$

$$f(y) = f(x) + \int_0^1 f'(x + t(y-x))dt \cdot (y-x)$$

$$f(y) - f(x) = \int_0^1 f'(x + t(y-x))dt \cdot (y-x)$$

- Replace f with multivariate gradient function $\nabla F : \mathbb{R}^n \to \mathbb{R}^n$, of which input is a parameter space with norm $|| \cdot ||_p$ and output is a gradient (dual) space with dual norm $|| \cdot ||_q$.

$$\nabla F(y) - \nabla F(x) = \int_0^1 \nabla^2 F(x + t(y-x))dt \cdot (y-x)$$

- Take their norms:

$$||\nabla F(y) - \nabla F(x)||_q = \int_0^1 ||\nabla^2 F(x + t(y-x))||_{p\to q} dt \cdot ||y-x||_p$$

- Notice here, that this looks quite similar to the Lipschitz condition we stated before!

$$||\nabla F(x) - \nabla F(y)||_q \leq L_p||x-y||_p$$

- So this means that the norm of the Hessian is bounded by the Lipschitz constant:

$$||\nabla^2 F(x)||_{p\to q} \leq L_p, \forall x$$

- For example, if we are using ℓ_2 norm on the parameter, its dual space is also ℓ_2. Then, the induced norm $\ell_2 \to \ell_2$ is a spectral norm, which takes the maximum singular value of the matrix. It's well known that the maximum singular value of an Hessian represents the sharpness of the loss landscape around current point x!
- Additionally, notice that for high L_p (or λ), the step size proportionally decreases, which is natural: you do have to take smaller steps when the landscape is sharp!

C. Modular Steepest Descent

C.1. Generalizing to Modular Norm

- Previous definition of steepest descent was defined on vectors $g, \Delta w \in \mathbb{R}^n$,
  where
  n is the total number of parameters in the model (i.e., all weights flattened).
- This disregards the structure of the architecture: that each *layer* has the parameters.
- In the most general form, each layer can use different types of norm:
  - Each layer has parameters $W_1, \cdots, W_L$
  - with scalar coefficients $s_1, \cdots, s_L > 0$ (honestly not sure why this popped up)
  - and with norms $||\cdot||_1, \cdots, ||\cdot||_L$
- Then the corresponding steepest descent problem is given by:

$$\underset{\Delta W_1,\cdots,\Delta W_L}{\arg\min}\left[\sum_{l=1}^{L}\langle G_l, \Delta W_l\rangle + \frac{\lambda}{2}\max_{l=1}^{L} s_l^2||\Delta W_l||_l^2\right]$$

- Note
  - $\langle\cdot,\cdot\rangle$ is the Frobenius inner product, which is a fancy name for flatten-then-dot-product.
  - This means that we can alternatively use $\mathrm{tr}(G_l^T W_l) = \langle G_l, \Delta W_l\rangle$.
  - W_l and G_l are always the same shape of $\mathbb{R}^{d_{out}\times d_{in}}$
- Same as before, this can be directly solved given $\{W_l, s_l, ||\cdot||_l\}_{l=1\ldots L}$:

$$\Delta W_l = -\frac{\eta}{s_l} \underset{||T_l||_l=1}{\arg\max} \langle G_l, T_l \rangle, \text{where } \eta = \frac{1}{\lambda} \sum_{k=1}^{L} \frac{1}{s_k} ||G_k||_k^\dagger$$

- Funnily enough, the max operation in the descent problem didn't matter; the step size η will be a weighted sum of the dual norms across every layer (see derivation).

C.2. Solution to Modular Steepest Descent Problem

- Similar to the first-order steepest descent.
- Decouple ΔW_l to direction and step size: $\Delta W_l = c_l \cdot T_l, c \geq 0, ||T_l||_l = 1$

$$\underset{\Delta W_1, \cdots, \Delta W_L}{\arg\min} \left[\sum_{l=1}^{L} \langle G_l, \Delta W_l \rangle + \frac{\lambda}{2} \max_{l=1}^{L} s_l^2 ||\Delta W_l||_l^2 \right] = \underset{c_l, \cdots, c_L \geq 0}{\arg\min} \left[\sum_{l=1}^{L} c_l \min_{||T_l||_l=1} \langle G_l, T_l \rangle + \frac{\lambda}{2} \max_{l=1}^{L} s_l^2 ||c_l$$

- Because $||T_l||_l = 1$:

$$\underset{c_l, \cdots, c_L \geq 0}{\arg\min} \left[\sum_{l=1}^{L} c_l \min_{||T_l||_l=1} \langle G_l, T_l \rangle + \frac{\lambda}{2} \max_{l=1}^{L} s_l^2 c_l^2 \right]$$

- By definition of dual norm:

$$\underset{c_l, \cdots, c_L \geq 0}{\arg\min} \left[-\sum_{l=1}^{L} c_l ||G_l||_l^\dagger + \frac{\lambda}{2} \max_{l=1}^{L} s_l^2 c_l^2 \right]$$

- Here, minimum is only reached when $s_1 c_1 = s_2 c_2 = \cdots = s_L c_L$
- To see that, let's call that certain value η.
- If there one i such that $s_i c_i < \eta$, we can simply increase c_i until $s_i c_i = \eta$, which minimizes the left term without changing the max term on the right! As a result, this becomes the problem of deciding η:

$$\underset{c_l, \cdots, c_L \geq 0}{\arg\min} \left[-\sum_{l=1}^{L} c_l ||G_l||_l^\dagger + \frac{\lambda}{2} \eta^2 \right]$$

- Deriving ΔW_l
    - $T_l = \arg\min_{|||T_l|_l=1} \langle G_l, T_l \rangle = -\arg\max_{|||T_l|_l=1} \langle G_l, T_l \rangle$
    - $c = \frac{\eta}{s_l}$, where $\eta = \frac{1}{\lambda} \sum_{k=1}^{L} \frac{1}{s_k} ||G_k||_k^\dagger$
    - Thus, $\Delta W_l = c_l \cdot T_l = -\frac{\eta}{s_l} \arg\max \langle G_l, T_l \rangle$

C.3. Steepest Descent Under Spectral Norm

- The steepest descent under spectral norm:

$$\underset{\Delta W_1,\cdots,\Delta W_L}{\arg\min}\left[\sum_{l=1}^{L}\langle G_l,\Delta W_l\rangle+\frac{\lambda}{2}\max_{l=1}^{L}||\Delta W_l||_{\ell_2\to\ell_2}^2\right]$$

- Then the solution is:

$$\Delta W_l=\eta\cdot U_l V_l^T,\text{where }\eta=\frac{1}{\lambda}\sum_{l=1}^{L}\mathrm{tr}(\Sigma_l)$$

- Let's try deriving this.
- We start with the general solution:

$$\Delta W_l=-\frac{\eta}{s_l}\underset{||T_l||_l=1}{\arg\max}\langle G_l,T_l\rangle,\text{where }\eta=\frac{1}{\lambda}\sum_{k=1}^{L}\frac{1}{s_k}||G_k||_k^{\dagger}$$

- Above problem sets $s_l=1,||\cdot||_l=||\cdot||_{\ell_2\to\ell_2}$, for all l.
- Also, we can use trace to compute the Frobenius dot product:

$$\Delta W_l=-\eta\cdot\underset{||T_l||_l=1}{\arg\max}\,\mathrm{tr}(G_l^T T_l),\text{where }\eta=\frac{1}{\lambda}\sum_{l=1}^{L}||G_l||_{\ell_2\to\ell_2}^{\dagger}$$

- Given the results first, we need to show that:
  1. $||G_l||_{\ell_2\to\ell_2}^{\dagger}=\mathrm{tr}(\Sigma_l)$
  2. $\arg\max_{||T_l||_l=1}\mathrm{tr}(G_l^T T_l)=UV^T$

1. **Dual-norm of spectral norm**

- We'll use the rank-1 SVD decomposition: $G=U\Sigma V^T$

$$\begin{aligned}&||G||_{\ell_2\to\ell_2}^{\dagger}\\=&\max_{||T||_{\ell_2\to\ell_2}=1}\mathrm{tr}(G^TT)\\=&\max_{||T||_{\ell_2\to\ell_2}=1}\mathrm{tr}(V\Sigma U^TT)\end{aligned}$$

- Using the cycle property of trace (i.e. $\mathrm{tr}(ABC)=\mathrm{tr}(BCA)=\mathrm{tr}(CAB)$):

$$=\max_{||T||_{\ell_2\to\ell_2}=1}\mathrm{tr}(\Sigma U^TTV)$$

- Since $||T||_{\ell_2\to\ell_2}=1$, the SVD decomposition of $T=U'V'^T$ (all $\sigma=1$)
- This means that the maximum is obtained with $T=UV^T$, which cancels everything out:

$$= \max_{||T||_{\ell_2 \to \ell_2} = 1} \mathrm{tr}(\Sigma)$$

2. **Argmax of trace**

- It is already shown in #1 that $T = UV^T$ maximizes $tr(G^T T)$ when T is constrained to spectral norm.