A Tutorial on Gradient Descend

Richard Xu

August 29, 2022

1 Implicit bias of gradient descend

This section explains [1]. The big picture here is to show the gradient $\nabla_{\mathbf{w}}\mathcal{L}(\mathbf{w}) \neq \mathbf{0}$ (section of 1.3.2), the loss function $\mathcal{L}(\mathbf{w})$ will continue to decrease using gradient descent. This makes $\|\mathbf{w}(t)\| \to \infty$ as $t \to \infty$. As a result, the weights of the few dominant linear combination terms correspond to the weights associated with the support vectors.

1.1 classifier without max-margin

looking at support vector machine term below:

$$\begin{aligned} &\min \left(\frac{1}{2}\|\mathbf{w}\|^2 \right) \\ &\text{subject to:} \quad 1 - y_i(\mathbf{w}^\top x_i + w_0) \leq 0 \quad \forall i \end{aligned} \tag{1}$$

If we were not trying to solve a max-margin problem: if we were just trying to express the problem as a linear classier. Then, the objective (for a single $\mathbf{x}_i, y_i$ pair can be written as):

$$y_i(\mathbf{w}^\top \mathbf{x}_i + w_0) > 0 \tag{2}$$

to make things even simpler, drop the w_0:

$$y_i(\mathbf{w}^\top \mathbf{x}_i) > 0 \tag{3}$$

1.1.1 smooth loss

smooth loss function used to penalize incorrect classification, for example:

$$\begin{aligned} &\ell(u) = \exp^{-u} \\ \implies &\ell(\mathbf{w}^\top \mathbf{x}_i y_i) = \exp^{\left(-\mathbf{w}^\top \mathbf{x}_i y_i\right)} \end{aligned} \tag{4}$$

in words, we must "push" value of $\mathbf{w}^\top \mathbf{x}_i y_i$ to be large $+$ve value (for correctly classified data/label pairs) when smooth loss function is assigned to

1.2 use gradient descend

when gradient descend is used to minimize the objective below (note analytical solution available for svm):

$$\begin{aligned}
&\min \mathcal{L}(\mathbf{w}) \\
&= \min \sum_{i=1}^{n} \ell(\mathbf{w}^\top \mathbf{x}_i y_i) \\
&= \min \sum_{i=1}^{n} \ell(\mathbf{w}^\top \tilde{\mathbf{x}}_i) \quad \text{let } \tilde{\mathbf{x}}_i = \mathbf{x}_i y_i
\end{aligned} \tag{5}$$

1.2.1 gradient for generic loss $\mathcal{L}$

$$\begin{aligned}
\nabla_{\mathbf{w}} \mathcal{L}(\mathbf{w}) &= \nabla_{\mathbf{w}} \sum_{i=1}^{n} \ell\big(\mathbf{w}^\top \tilde{\mathbf{x}}_i\big) \\
&= \sum_{i=1}^{n} \ell'\big(\mathbf{w}^\top \tilde{\mathbf{x}}_i\big) \tilde{\mathbf{x}}_i
\end{aligned} \tag{6}$$

substitute into gradient descend:

$$\begin{aligned}
\mathbf{w}(t+1) &= \mathbf{w}(t) - \eta \nabla_{\mathbf{w}(t)} \mathcal{L}\big(\mathbf{w}(t)\big) \\
&= \mathbf{w}(t) - \eta \sum_{i=1}^{n} \ell'\big(\mathbf{w}(t)^\top \tilde{\mathbf{x}}_i\big) \tilde{\mathbf{x}}_i
\end{aligned} \tag{7}$$

we are interested in the behavior of $\mathbf{w}(t) \to \infty$

1.3 magnitude: $\|\mathbf{w}(t)\| \to \infty$

1.3.1 no finite critical points $\nabla_{\mathbf{w}(t)} \mathcal{L}\big(\mathbf{w}(t)\big) = 0$

It's difficult to show from the gradient directly why the expression $\sum_{i=1}^{n} \ell'\big(\mathbf{w}(t)^\top \tilde{\mathbf{x}}_i\big) \tilde{\mathbf{x}}_i$ never reach 0, i.e.,

$$\text{to show why} \quad \lim_{t \to \infty} \sum_{i=1}^{n} \ell'\big(\mathbf{w}(t)^\top \tilde{\mathbf{x}}_i\big) \tilde{\mathbf{x}}_i \neq 0 \tag{8}$$

Note that people may be confused to think if we let $\ell(u) = \exp^{-u}$, then $\ell'(u) \neq 0$ anyway. right? However, since we have a sum and not just a term. Making the gradient zero may still seems "possible". To illustrate, when we let $n = 2$, we may obtain a situation where:

$$\ell'\big(\mathbf{w}^\top \tilde{\mathbf{x}}_1\big) \tilde{\mathbf{x}}_1 = -\ell'\big(\mathbf{w}^\top \tilde{\mathbf{x}}_2\big) \tilde{\mathbf{x}}_2 \qquad \text{for some } \mathbf{w} \tag{9}$$

1.3.2 show $\nabla_{\mathbf{w}}\mathcal{L}\big(\mathbf{w}(t)\big)$ won't be a zero vector

Let's assume $\exists \mathbf{w}^\star \neq \mathbf{0}$ making data separable (if data is separable). looking at the following expression:

$$\begin{aligned}
\mathbf{w}^{\star\top}\eta\nabla_{\mathbf{w}(t)}\mathcal{L}\big(\mathbf{w}(t)\big) &= \mathbf{w}^{*\top}\sum_{i=1}^{n}\ell'\big(\mathbf{w}(t)^\top\tilde{\mathbf{x}}_i\big)\tilde{\mathbf{x}}_i \\
&= \sum_{i=1}^{n}\underbrace{\underbrace{\ell'\big(\mathbf{w}(t)^\top\tilde{\mathbf{x}}_i\big)}_{<0}\underbrace{\tilde{\mathbf{x}}_i^\top\mathbf{w}^*}_{>0}}_{<0}
\end{aligned} \tag{10}$$

Obviously, since:

$$\begin{aligned}
&\ell'\big(\mathbf{w}(t)^\top\tilde{\mathbf{x}}_i\big)\tilde{\mathbf{x}}_i^\top\mathbf{w}^* < 0 \text{ and } \mathbf{w}^* \neq \mathbf{0} \\
\implies &\ell'\big(\mathbf{w}(t)^\top\tilde{\mathbf{x}}_i\big)\tilde{\mathbf{x}}_i \neq \mathbf{0} \quad \forall i \\
\implies &\nabla_{\mathbf{w}}\mathcal{L}\big(\mathbf{w}(t)\big) \neq \mathbf{0}
\end{aligned} \tag{11}$$

Explain each two terms:

1. $\mathbf{w}^{*\top}\tilde{\mathbf{x}}_i > 0 \quad \forall i$ if all data are all correctly classified/linearly separable:

$$y_i(\mathbf{w}^{*\top}\mathbf{x}_i) > 0 \tag{12}$$

   note that up to here, we made **no** reference with max-margin

2. $l'(.) < 0$ as long as we choose a monotonically decreasing l which means its gradient < 0

3. also note that in here, we merely assumed $\exists \mathbf{w}^*$. Don't get confused, it is not where $\mathbf{w}(t)$ converges to!

4. also note if it's possible for $\nabla_{\mathbf{w}}\mathcal{L}\big(\mathbf{w}(t)\big) = \mathbf{0}$, it means the gradient descend will not run indefinitely.

1.3.3 why $\|\mathbf{w}(t)\| \to \infty$?

We know gradient descend on a smooth loss will converge to a minimum. This will be illustrated in the β-smooth section. Since ℓ is a smooth function, so is $\mathcal{L}\big(\mathbf{w}(t)\big) = \sum_{i=1}^{n}\ell(\mathbf{w}(t)^\top\tilde{\mathbf{x}}_i)$:

$$
\begin{aligned}
\|\nabla\mathcal{L}(x)-\nabla\mathcal{L}(y)\| &= \Big\|\frac{1}{n}\sum_i \nabla\ell_i(x) - \frac{1}{n}\sum_i \nabla\ell_i(y)\Big\| \\
&= \frac{1}{n}\Big\|\sum_i (\nabla\ell_i(x)-\nabla\ell_i(y))\Big\| \\
&\leq \frac{1}{n}\sum_i \Big\|\nabla\ell_i(x)-\nabla\ell_i(y)\Big\| \quad \text{triangle inequality} \\
&\leq \frac{1}{n}\sum_i (\beta_i\|x-y\|) \\
&= \Big(\frac{1}{n}\sum_i \beta_i\Big)\|x-y\|
\end{aligned} \tag{13}
$$

However, the above says there is no critical points. Putting above two arguments together, and look at the objective $\sum_{i=1}^n \ell(\mathbf{w}^\top\tilde{\mathbf{x}}_i)$, we can see that, since the gradient descend algorithm continues to run (and the loss will continuously becoming smaller):

$$
\Big(\mathcal{L}\big(\mathbf{w}(t)\big) = \sum_{i=1}^n \ell(\mathbf{w}(t)^\top\tilde{\mathbf{x}}_i)\Big) \to 0 \implies \mathbf{w}(t)^\top\tilde{\mathbf{x}}_i \to \infty \qquad \text{think } \exp(-u) \tag{14}
$$

Since $\tilde{\mathbf{x}}_i$ is fixed, then $\|\mathbf{w}(t)\| \to \infty$. Note that this is why we need to show there is **no** critical points first.

The norm is needed as $y_i \in \{1, -1\}$, it means:

$$
\begin{aligned}
\lim_{t\to\infty} \|\mathbf{w}(t)\| &= \infty \\
\text{or equivalently} \quad \|\mathbf{w}(t)\| &\to \infty
\end{aligned} \tag{15}
$$

1.4 what about direction of $\mathbf{w}(t)$?

To characterize direction, we look at normalized $\lim_{t\to\infty} \frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}$

Theorem 1 *under assumption as* $t \to \infty$*, Gradient descend behaves as:*

$$
\mathbf{w}(t) \approx \frac{\mathbf{w}_{svm}}{\|\mathbf{w}_{svm}\|} \tag{16}
$$

1.4.1 explanation

when $\mathbf{w}(t) \to \infty$, it has the same direction of the SVM solution, i.e., its normalized version $\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}$ becomes that of the $\mathbf{w}_{\text{svm}}$

$\mathbf{w}_{\text{svm}}$ gives max-margin classifier which has better generalization!

1.5 proof of theorem

consider exponential loss $\mathcal{L}(u) = \exp(-u)$, gradient descend in asymptotic regime in shown in Eq.(14):

$$
\mathbf{w}(t)^\top\tilde{\mathbf{x}}_i \to \infty \quad \forall i \tag{17}
$$

1.5.1 what is asymptotic "simplification" convergence?

The definition of the notation $a_n \to b_n$ is designed to mean that $a_n \approx b_n$ for large n, where the fit gets better and better as n gets larger, for example:

$$\lim_{x\to\infty} x^2 + x + 1 = x^2 \tag{18}$$

and,

$$\begin{aligned} u(x+h) &= u(x) + u'(x)h + \frac{u''(x)}{2}h^2 + \dots \\ u(x+h) - u(x) + u'(x)h &= \frac{u''(x)}{2}h^2 + \dots \\ \Big|\frac{u(x+h)-u(x)}{h} - u'(x)\Big| &= \Big|\frac{u''(x)}{2}h + \frac{u'''(x)}{3!}h^2 \dots\Big| \quad \text{divided by } h \\ \implies \lim_{h\to 0}\Big|\frac{u(x+h)-u(x)}{h} - u'(x)\Big| &= \Big|\frac{u''(x)}{2}h\Big| \end{aligned} \tag{19}$$

1.5.2 asymptotic convergence of $\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|}$

we express $\mathbf{w}(t)$ as a linear function in terms of $\mathbf{w}_\infty$ (the remaining work is to find what $\mathbf{w}_\infty$ is):

$$\mathbf{w}(t) = \underbrace{m(t)}_{\text{magnitude}}\ \mathbf{w}_\infty + \underbrace{\mathbf{b}(t)}_{\text{residual}} \tag{20}$$

assume $\exists \mathbf{w}_\infty$ (which is a unit vector), the limit of the normalization $\frac{\mathbf{w}(t)}{\|\mathbf{w}(t)\|} \to \mathbf{w}_\infty$, under assumptions of both already stated, and new ones:

1. $\lim_{t\to\infty} \frac{\mathbf{b}(t)}{m(t)} = 0$ as $\|\mathbf{b}(t)\|$ is relatively smaller compare with $\|\mathbf{w}(t)\|$, as $t \to \infty$
2. $m(t) \to \infty$ makes sense as $\|\mathbf{w}(t)\| \to \infty$

since $m(t)$ is the magnitude, then $m(t) \geq 0$. Looking at the gradient again:

$$
\begin{aligned}
\nabla_{\mathbf{w}(t)}\mathcal{L}\big(\mathbf{w}(t)\big) &= \sum_{i=1}^{n} \ell'\big(\mathbf{w}(t)^\top \tilde{\mathbf{x}}_i\big)\tilde{\mathbf{x}}_i \\
&= -\sum_{i=1}^{n} \exp^{\left(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_i\right)}\,\tilde{\mathbf{x}}_i \quad \text{substitute } \ell'(u) = -\exp(-u) \\
\implies -\nabla_{\mathbf{w}(t)}\mathcal{L}\big(\mathbf{w}(t)\big) &= \sum_{i=1}^{n} \exp^{\left(-\mathbf{w}(t)^\top \tilde{\mathbf{x}}_i\right)}\,\tilde{\mathbf{x}}_i \\
&= \sum_{i=1}^{n} \exp^{\left(-\left(m(t)\mathbf{w}_\infty + \mathbf{b}(t)\right)^\top \tilde{\mathbf{x}}_i\right)}\,\tilde{\mathbf{x}}_i \quad \text{substitute } \mathbf{w}(t) = m(t)\mathbf{w}_\infty + \mathbf{b}(t) \\
&= \sum_{i=1}^{n} \exp^{-m(t)\mathbf{w}_\infty^\top \tilde{\mathbf{x}}_i}\,\tilde{\mathbf{x}}_i \times \exp^{-\mathbf{b}(t)^\top \tilde{\mathbf{x}}_i}\,\tilde{\mathbf{x}}_i \\
&\approx \sum_{i=1}^{n} \exp^{-m(t)\mathbf{w}_\infty^\top \tilde{\mathbf{x}}_i}\,\tilde{\mathbf{x}}_i \qquad \because \lim_{t\to\infty}\frac{\mathbf{b}(t)}{m(t)} = 0 \\
&= \sum_{i=1}^{n} \underbrace{\exp^{\left(-m(t)\mathbf{w}_\infty^\top \tilde{\mathbf{x}}_i\right)}}_{\alpha_i}\tilde{\mathbf{x}}_i
\end{aligned}
\tag{21}
$$

so gradient step would be some non-negative linear combination of $\tilde{\mathbf{x}}_i$, i.e.,:

$$
-\nabla_{\mathbf{w}(t)}\mathcal{L}\big(\mathbf{w}(t)\big) = \sum_{i=1}^{n} \alpha_i \tilde{\mathbf{x}}_i \tag{22}
$$

1.5.3 dominate terms

assumes $\mathbf{w}_\infty$ classifies the linearly separable data correctly, then:

$$
\mathbf{w}_\infty^\top \tilde{\mathbf{x}}_i > 0 \tag{23}
$$

since $m(t) \to \infty$, we have only a few dominate terms in $\{\tilde{\mathbf{x}}_i\}$ (multiply by ∞ makes they dominate!). since these $\tilde{\mathbf{x}}_i$ are closest to (and on the) decision boundary, then they are precisely support vectors! So the set is support vector set s.v.!

Note that if mulitple $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$ are the closest, i.e., $\mathbf{w}_\infty^\top \tilde{\mathbf{x}}_i = \mathbf{w}_\infty^\top \tilde{\mathbf{x}}_j$, then, they both are part of the support vector set!

$$
\begin{aligned}
-\nabla_{\mathbf{w}(t)}\mathcal{L}\big(\mathbf{w}(t)\big) &\approx \sum_{\tilde{\mathbf{x}}_i \in \text{s. v.}} \alpha_i \tilde{\mathbf{x}}_i \\
&= \sum_{\mathbf{x}_i \in \text{s. v.}} \alpha_i \mathbf{x}_i y_i
\end{aligned}
\tag{24}
$$

As each of the gradient step is a linear combination of $x_i \in$ s.v., then, so is $\mathbf{w}_\infty$, i.e.,

$$
\mathbf{w}_\infty = \sum_{\mathbf{x}_i \in \text{s.v.}} \alpha_i' \tilde{\mathbf{x}}_i \qquad \text{for some } \alpha_i' \neq \alpha_i \tag{25}
$$

since $\|\mathbf{w}(t)\| \to \infty$, then the initial $\mathbf{w}(0)$ value won't matter any more. There is one remaining issue though: $\mathbf{w}_\infty^\top \tilde{\mathbf{x}}_i \neq 1 \quad \forall \mathbf{x}_i \in$ s.v. so look at the next section:

1.5.4 from $\mathbf{w}_\infty$ to obtain $\mathbf{w}_{\text{svm}}$

lastly, we need to scale $\mathbf{w}_\infty$ to become $\mathbf{w}_{\text{svm}}$. let's see what if we perform $\frac{\mathbf{w}_\infty}{\text{some constant}}$. Now let's have $\tilde{\mathbf{x}}_{\text{s.v.}}$ such that:

$$\mathbf{w}_\infty^\top \tilde{\mathbf{x}}_{\text{s.v.}} = \min_i \{\mathbf{w}_\infty^\top \tilde{\mathbf{x}}_i\} \tag{26}$$

although the picking of the "some constant" is arbitrary, but we pick $\min_i\{\mathbf{w}_\infty^\top \tilde{\mathbf{x}}_i\}$ to reflect the SVM solution:

$$\begin{aligned} \widehat{\mathbf{w}} &= \frac{\mathbf{w}_\infty}{\text{some constant}} \\ &= \frac{\mathbf{w}_\infty}{\mathbf{w}_\infty^\top \tilde{\mathbf{x}}_{\text{s.v.}}} \end{aligned} \tag{27}$$

note that $\|\mathbf{w}_\infty\| = 1$, but $\|\hat{\mathbf{w}}\| \neq 1$! By this process, it scales $\widehat{\mathbf{w}}$ such that when applying to $\tilde{\mathbf{x}}_{\text{s.v.}}$:

$$\begin{aligned} \widehat{\mathbf{w}}^\top \tilde{\mathbf{x}}_{\text{s.v.}} &= \frac{\mathbf{w}_\infty}{\mathbf{w}_\infty^\top \tilde{\mathbf{x}}_{\text{s.v.}}}^\top \tilde{\mathbf{x}}_{\text{s.v.}} \\ &= 1 \end{aligned} \tag{28}$$

and when it applies to other $\tilde{\mathbf{x}} \notin \{\tilde{\mathbf{x}}_{\text{s.v.}}\}$:

$$\begin{aligned} \widehat{\mathbf{w}}^\top \tilde{\mathbf{x}} &= \frac{\mathbf{w}_\infty}{\mathbf{w}_\infty^\top \tilde{\mathbf{x}}_{\text{s.v.}}}^\top \tilde{\mathbf{x}} \\ &> 1 \end{aligned} \tag{29}$$

Does $\widehat{\mathbf{w}}$ look familiar? Remember KKT condition is:

$$\widehat{\mathbf{w}} = \sum_{i=1}^N \lambda_i \tilde{\mathbf{x}}_i \tag{30}$$

with complementary duality:

$$\begin{cases} \lambda_i > 0 & \widehat{\mathbf{w}}^\top \tilde{\mathbf{x}}_i = 1 \quad \text{support vectors} \\ \lambda_i = 0 & \widehat{\mathbf{w}}^\top \tilde{\mathbf{x}}_i > 1 \quad \text{non support vector} \end{cases} \tag{31}$$

compare with equation in SVM section, $\widehat{\mathbf{w}} = \mathbf{w}_{\text{svm}}$

Since we already prove $\widehat{\mathbf{w}}$ is proportional to $\mathbf{w}_\infty$. Therefore, $\mathbf{w}_\infty$ is the SVM solution up to some constant!

2 Fenchel dual function

$$f^{\star}(\mathbf{y}) = \sup_{\mathbf{x}} \left[\mathbf{x}^{\top}\mathbf{y} - f(\mathbf{x})\right] \tag{32}$$

2.1 property of Fenchel dual

2.1.1 visualization

visualization is achieved similarly to what was done previously in the genera dual function section: Consider $f^{*}(\mathbf{y})$ is a function on the "gradient space", except we now have parameter:

$$\lambda \to \mathbf{y} \tag{33}$$

Just like the general dual function section, we can visualize by generating $f^{*}(\mathbf{y})$ from maximization of finite lines (in $\mathbf{y}$) defined by a finite set $\{\mathbf{x}\}$, and $\mathbf{x}$ and $f(\mathbf{x})$ are treated like "constant line parameters". the alternative way to consider this is to rewrite Eq.(32) as:

$$f^{\star}(\mathbf{y}) = \sup_{\mathbf{x}} \left[- f(\mathbf{x}) + \mathbf{y}^{\top}\mathbf{x}\right] \tag{34}$$

where the dual "constraint" $g(\mathbf{x}) \equiv \mathbf{y}^{\top}\mathbf{x}$. Note $f^{*}(\mathbf{y})$ is always convex.

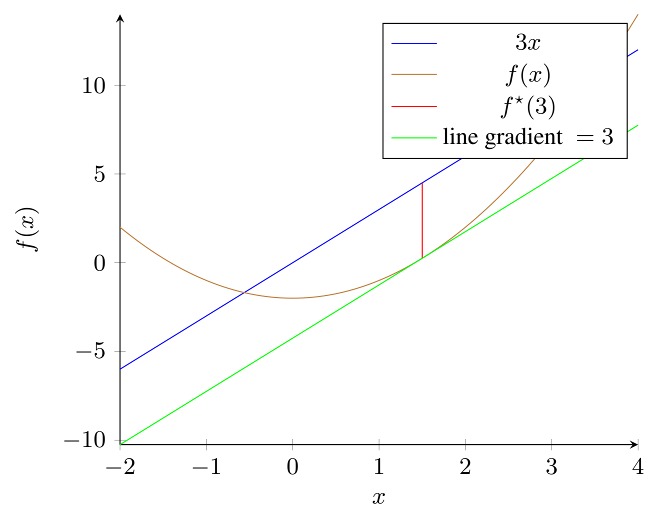


Figure 1: simple demonstration of Fenchel dual function

imagine we picked a particular value of $\mathbf{y} = 3$ in order to compute the value of $f^{\star}(3)$, then we worked out:

$$\begin{aligned} f^{\star}(3) &= \sup_{x} \left[3x - f(x)\right] \\ \hat{x}_3 &= \arg\max_{x} \left[3x - f(x)\right] \end{aligned} \tag{35}$$

which is the red line segment, i.e., the line segment/difference between the line $3x$ (gradient 3, passing through origin), i.e., the blue line and the function $f(\mathbf{x})$, i.e., the brown line.

2.1.2 important observation

introducing $\arg\max$ variable $\hat{\mathbf{x}}_{\mathbf{y}}$, such that by given $\mathbf{y}$:

$$\begin{aligned}\hat{\mathbf{x}}_{\mathbf{y}} &= \arg\max_{\mathbf{x}} \left[\mathbf{y}^\top \mathbf{x} - f(\mathbf{x})\right] \\ \implies \nabla_{\mathbf{x}}\left(\hat{\mathbf{x}}_{\mathbf{y}}^\top \mathbf{y} - f(\hat{\mathbf{x}}_{\mathbf{y}})\right) &= 0 \\ \implies \nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_{\mathbf{y}}) &= \mathbf{y}\end{aligned} \tag{36}$$

It says that given $\mathbf{y}$, maximum of the line segment length between $\mathbf{y}^\top \mathbf{x}$ and $f(\mathbf{x})$ occurs at particular $\hat{\mathbf{x}}_{\mathbf{y}}$, where its gradient $\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_{\mathbf{y}}) = \mathbf{y}$. In our case, we picked $y = 3$, therefore $\nabla_x f(\hat{x}_3) = 3$.

Visually, we see two parallel lines in blue and green as both have gradient $y = 3$. So far, we express everything in tersm of gradients. However, they will be replaced by subgradients.

2.1.3 Fenchel's inequality

for any $\mathbf{x}$ and y:

$$f(\mathbf{x}) + f^\star(\mathbf{y}) \geq \mathbf{x}^\top \mathbf{y} \tag{37}$$

the reason is because:

$$\begin{aligned}f^\star(\mathbf{y}) &= \sup_{\mathbf{x}} \left[\mathbf{x}^\top \mathbf{y} - f(\mathbf{x})\right] \\ &\geq \mathbf{x}^\top \mathbf{y} - f(\mathbf{x}) \\ \implies f(\mathbf{x}) + f^\star(\mathbf{y}) &\geq \mathbf{x}^\top \mathbf{y}\end{aligned} \tag{38}$$

2.1.4 example

$$\begin{aligned}f(\mathbf{x}) &= \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} + \mathbf{q}^\top \mathbf{x} + c \\ f^\star(\mathbf{y}) &= \sup_{\mathbf{x}} \left(\mathbf{y}^\top \mathbf{x} - \frac{1}{2}\mathbf{x}^\top \mathbf{Q}\mathbf{x} - \mathbf{q}^\top \mathbf{x} - c\right)\end{aligned} \tag{39}$$

to find optimal $\mathbf{x}$:

$$\begin{aligned}\hat{\mathbf{x}}_{\mathbf{y}} - \mathbf{Q}\mathbf{x} - \mathbf{q} &= \mathbf{0} \\ \hat{\mathbf{x}}_{\mathbf{y}} &= \mathbf{Q}^{-1}(\mathbf{y} - \mathbf{q})\end{aligned} \tag{40}$$

substitute it back to $f^\star(\mathbf{y})$:

$$
\begin{aligned}
f^\star(\mathbf{y}) &= \mathbf{y}^\top \big(\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{q})\big) - \frac{1}{2}\big(\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{q})\big)^\top \mathbf{Q}\big(\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{q})\big) - \mathbf{q}^\top \big(\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{q})\big) - c \\
&= (\mathbf{y}-\mathbf{q})^\top \big(\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{q})\big) - \frac{1}{2}\big(\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{q})\big)^\top \mathbf{Q}\big(\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{q})\big) - c \\
&= (\mathbf{y}-\mathbf{q})^\top \mathbf{Q}^{-1}(\mathbf{y}-\mathbf{q}) - \frac{1}{2}(\mathbf{y}-\mathbf{q})^\top \mathbf{Q}^{-\top}\mathbf{Q}\mathbf{Q}^{-1}(\mathbf{y}-\mathbf{q}) - c \\
&= (\mathbf{y}-\mathbf{q})^\top \mathbf{Q}^{-1}(\mathbf{y}-\mathbf{q}) - \frac{1}{2}(\mathbf{y}-\mathbf{q})^\top \mathbf{Q}^{-1}(\mathbf{y}-\mathbf{q}) - c \qquad \because \mathbf{Q}^\top = \mathbf{Q} \\
&= \frac{1}{2}(\mathbf{y}-\mathbf{q})^\top \mathbf{Q}^{-1}(\mathbf{y}-\mathbf{q}) - c
\end{aligned}
\tag{41}
$$

2.1.5 conjugate of conjugate function in general

In general, for **any** arbitary function $f(x)$, conjugate of conjugate function is no greater than the original function:

$$
f^{**}(\mathbf{x}) \le f(\mathbf{x}) \tag{42}
$$

this can be easily proven:

$$
\begin{aligned}
f^{**}(\mathbf{x}) &= \sup_{\mathbf{y}} \big[\mathbf{x}^\top \mathbf{y} - f^*(\mathbf{y})\big] \\
&= \sup_{\mathbf{y}} \big[\mathbf{x}^\top \mathbf{y} - \sup_{\mathbf{z}} \big[\mathbf{y}^\top \mathbf{z} - f(\mathbf{z})\big]\big] \qquad \text{need to use variable } \mathbf{z} \\
&= \sup_{\mathbf{y}} \big[\mathbf{x}^\top \mathbf{y} + \inf_{\mathbf{z}} \big[-\mathbf{y}^\top \mathbf{z} + f(\mathbf{z})\big]\big] \qquad -\sup\{\mathbf{x}\} \Leftrightarrow +\inf\{-\mathbf{x}\} \\
&= \sup_{\mathbf{y}} \big[\inf_{\mathbf{z}} \big[\mathbf{y}^\top (\mathbf{z}-\mathbf{x}) + f(\mathbf{z})\big]\big] \\
&\le \inf_{\mathbf{z}} \big[\sup_{\mathbf{y}} \big[\mathbf{y}^\top (\mathbf{z}-\mathbf{x}) + f(\mathbf{z})\big]\big] \\
&= f(\mathbf{x})
\end{aligned}
\tag{43}
$$

the last line can be seen as $\mathbf{y}$ is not bounded, so it can always be choosen that the inner term $\sup_{\mathbf{y}} \big[\mathbf{y}^\top (\mathbf{z}-\mathbf{x}) + f(\mathbf{z})\big] \to \infty$, therefore, we can prevent this by having $\mathbf{z} = \mathbf{x}$. Note that this is used often in the context of $\min \max$ or $\inf \sup$.

2.1.6 conjugate of conjugate function when $f(\mathbf{x})$ is convex

however, when $f(\mathbf{x})$ is convex:

$$
f^{**}(\mathbf{x}) = f(\mathbf{x}) \tag{44}
$$

we can exploit the above property:

$$
\begin{aligned}
{\color{red}f^*(\mathbf{y})} &= \max_{\mathbf{x}} \big[\mathbf{x}^\top \mathbf{y} - {\color{blue}f(\mathbf{x})}\big] \quad \forall \mathbf{y} \quad \text{by definition} \\
\Longrightarrow \ {\color{blue}f(\mathbf{x})} = f^{**}(\mathbf{x}) &= \max_{\mathbf{y}} \big[\mathbf{x}^\top \mathbf{y} - {\color{red}f^*(\mathbf{y})}\big] \quad \forall \mathbf{x} \quad \text{by } f^{**}(\mathbf{x}) = f(\mathbf{x})
\end{aligned}
\tag{45}
$$

2.1.7 derivatives of conjugate function

when $f(\mathbf{x})$ is convex and differentiable, then Eq.(36) can be extended to both ways:

$$\begin{aligned} f(\mathbf{x}) + f^\star(\mathbf{y}) = \mathbf{x}^\top \mathbf{y} \quad \forall \mathbf{x}, \mathbf{y} \\ \implies \nabla_{\mathbf{x}} f(\mathbf{x}) = \mathbf{y} \quad \text{and} \quad \nabla_{\mathbf{y}} f^\star(\mathbf{y}) = \mathbf{x} \end{aligned} \tag{46}$$

note that we make the two equations of the last line having the same $\mathbf{x}$ and $\mathbf{y}$, we call them $\hat{\mathbf{x}}_{\mathbf{y}}$ and $\hat{\mathbf{y}}_{\mathbf{x}}$:

$$\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_{\mathbf{y}}) = \hat{\mathbf{y}}_{\mathbf{x}} \quad \text{and} \quad \nabla_{\mathbf{y}} f^\star(\hat{\mathbf{y}}_{\mathbf{x}}) = \hat{\mathbf{x}}_{\mathbf{y}} \tag{47}$$

$$\nabla_{\mathbf{x}} f\big(\underbrace{\nabla_{\mathbf{y}} f^\star(\hat{\mathbf{y}}_{\mathbf{x}})}_{\hat{\mathbf{x}}_{\mathbf{y}}}\big) = \hat{\mathbf{y}}_{\mathbf{x}} \tag{48}$$

therefore, for a corresponding pair $(\mathbf{x}_{\mathbf{y}}, \mathbf{y}_{\mathbf{x}})$, obtained from computing the conjugate dual:

$$\nabla_{\mathbf{x}} f(\hat{\mathbf{x}}_{\mathbf{y}}) = \hat{\mathbf{y}}_{\mathbf{x}} \quad \iff \quad \nabla_{\mathbf{y}} f^\star(\hat{\mathbf{y}}_{\mathbf{x}}) = \hat{\mathbf{x}}_{\mathbf{y}} \tag{49}$$

This property says that the function argument $\hat{\mathbf{x}}_{\mathbf{y}}$ and its gradient through f, i.e., $\hat{\mathbf{y}}_{\mathbf{x}}$ have their roles switched if applied to the conjugate function $f^\star$.

2.1.8 A practical implication

A practical implication of the above is when you need to find stationary point for an dual function $f^\star(\cdot)$:

$$\hat{\mathbf{y}}_{\mathbf{x}} = \arg\max_{\mathbf{y}} \left[\mathbf{x}^\top \mathbf{y} - f^\star(\mathbf{y})\right] \tag{50}$$

and you know by definition, the following is true:

$$\begin{aligned} \nabla_{\mathbf{y}} f^\star(\hat{\mathbf{y}}_{\mathbf{x}}) - \mathbf{x} &= 0 \\ \nabla_{\mathbf{y}} f^\star(\hat{\mathbf{y}}_{\mathbf{x}}) &= \mathbf{x} \\ &= \hat{\mathbf{x}}_{\mathbf{y}} \end{aligned} \tag{51}$$

one way to solve for $\hat{\mathbf{y}}_{\mathbf{x}}$ is by taking the inverse of the Jacobian matrix:

$$\hat{\mathbf{y}}_{\mathbf{x}} = (\nabla_{\mathbf{y}} f^\star)^{-1}\big(\hat{\mathbf{x}}_{\mathbf{y}}\big) \tag{52}$$

However, using the property in Eq.(47), we can compute $\hat{\mathbf{y}}_{\mathbf{x}}$ simply by:

$$\hat{\mathbf{y}}_{\mathbf{x}} = \nabla_{\mathbf{x}} f\big(\hat{\mathbf{x}}_{\mathbf{y}}\big) \tag{53}$$

2.1.9 derivatives of conjugate function when non-differentiable

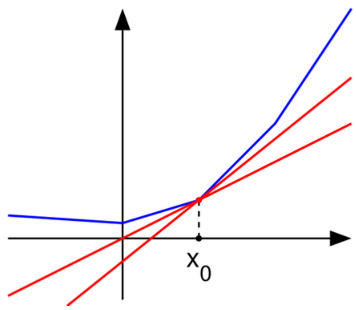
what if $f(\mathbf{x})$ or $f^*(\mathbf{y})$ is **not** differentiable everywhere then the last two lines become:

$$\Longrightarrow \ \mathbf{y} \in \partial f(\mathbf{x}) \quad \Longleftrightarrow \quad \mathbf{x} \in \partial f^*(\mathbf{y}) \tag{54}$$

in there $\partial f(\mathbf{x})$ is called **sub-differential**

2.1.10 what is sub-differential?

this is used when $f(\mathbf{x})$ is convex, but $f(\mathbf{x})$ is not differentiable at some $\mathbf{x}$



$$\partial f(x_0) := \left\{m \in \mathbb{R} \mid f(x) \geq f(x_0) + m(x - x_0)\ \forall x \in \mathbb{R}\right\} \tag{55}$$

in words, it means for convex, but non-differentiable function, sub-gradients at point x_0 are the "collection of gradients" where the lines they represent touches the function at $f(x_0)$, but is $\leq f(x)$ everywhere else

Contrary to gradient, i.e., $\frac{\mathrm{d}f(x)}{\mathrm{d}x}$, sub-differential returns a set of gradient values.

If f is differentiable at x, then it can be seen as a special case, $\partial f(x)$ just contains a singleton, $\{\frac{\mathrm{d}f(x)}{\mathrm{d}x}\}$:

$$\partial f(x) := \left\{m \in \mathbb{R} \mid f(x) \geq f(x_0) + m(x - x_0)\ \forall x \in \mathbb{R}\right\} = \left\{\frac{\mathrm{d}f(x)}{\mathrm{d}x}\right\} \tag{56}$$

2.2 proof for sub differential inverse

In addition to what was described in Section 2, let's look at the inverse from the sub differential perspective. Now we simplify our notation $\hat{\mathbf{y}}_{\mathbf{x}} \rightarrow \hat{\mathbf{y}}$ and $\hat{\mathbf{x}}_{\mathbf{y}} \rightarrow \hat{\mathbf{x}}$ instead:

$$f^\star(\hat{\mathbf{y}}) \equiv \sup_{\mathbf{x}}\{\hat{\mathbf{y}}^\top \mathbf{x} - f(\mathbf{x})\} \tag{57}$$

for a particular $\hat{\mathbf{y}}$, assume we find $\hat{\mathbf{x}}$ to be optimum, for which iff:

$$\begin{aligned} &\mathbf{0} \in \hat{\mathbf{y}} - \partial f(\hat{\mathbf{x}}) \\ \Longrightarrow\ &\hat{\mathbf{y}} \in \partial f(\hat{\mathbf{x}}) \end{aligned} \tag{58}$$

by substitution $\mathbf{x} = \hat{\mathbf{x}}$ into Eq.(57), (we did not do anything else) we have:

$$f^\star(\hat{\mathbf{y}}) \equiv \hat{\mathbf{y}}^\top \hat{\mathbf{x}} - f(\hat{\mathbf{x}}) \tag{59}$$

looking at the equation without picking $\mathbf{y} \equiv \hat{\mathbf{y}}$ (we will add it back into it). For a generic $\mathbf{y}$:

$$\begin{aligned} f^\star(\mathbf{y}) &= \sup_{\mathbf{x}} \{\mathbf{y}^\top \mathbf{x} - f(\mathbf{x})\} \\ &\geq \mathbf{y}^\top \hat{\mathbf{x}} - f(\hat{\mathbf{x}}) \qquad \hat{\mathbf{x}} \quad \text{optimized for } \hat{\mathbf{y}}, \text{ not for generic } \mathbf{y} \\ &= (\mathbf{y} - \hat{\mathbf{y}})^\top \hat{\mathbf{x}} - f(\hat{\mathbf{x}}) + \hat{\mathbf{y}}^\top \hat{\mathbf{x}} \qquad \text{add and subtract } \hat{\mathbf{y}}^\top \hat{\mathbf{x}} \\ &= f^\star(\hat{\mathbf{y}}) + (\mathbf{y} - \hat{\mathbf{y}})^\top \hat{\mathbf{x}} \qquad \text{from Eq.(59)} \quad f^\star(\hat{\mathbf{y}}) \equiv \hat{\mathbf{y}}^\top \hat{\mathbf{x}} - f(\hat{\mathbf{x}}) \\ &= f^\star(\hat{\mathbf{y}}) + \langle \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}} \rangle \end{aligned} \tag{60}$$

then, the moral of the story is that by looking at:

$$f^\star(\mathbf{y}) \geq f^\star(\hat{\mathbf{y}}) + \langle \hat{\mathbf{x}}, \mathbf{y} - \hat{\mathbf{y}} \rangle \tag{61}$$

since we know $f^\star(\mathbf{y})$ must be convex and using the definition of convex function, we know that it must be the case where $\hat{\mathbf{x}} \in \partial_{\mathbf{y}} f^\star(\hat{\mathbf{y}})$

we have shown that:

$$\hat{\mathbf{y}} \in \partial f(\hat{\mathbf{x}}) \implies \hat{\mathbf{x}} \in \partial f^\star(\hat{\mathbf{y}}) \tag{62}$$

2.3 example of Fenchel/conjugate for $f(\cdot) \equiv \|\cdot\|$

2.3.1 $f \equiv \|\cdot\|$ is a vector norm

$$\begin{aligned} f(\mathbf{x}) &= \|\mathbf{x}\| \\ f^\star(\mathbf{y}) &= \sup_{\mathbf{x}} \left\{\mathbf{y}^\top \mathbf{x} - \|\mathbf{x}\|\right\} \\ &\equiv \sup_{\mathbf{x}} \left\{\langle \mathbf{y}, \mathbf{x} \rangle - \|\mathbf{x}\|\right\} \end{aligned} \tag{63}$$

what is $f^\star(\mathbf{y})$? It actually depends on the sign of $\langle \mathbf{y}, \mathbf{x} \rangle - \|\mathbf{x}\|$, and interestingly, we can obtain this through dual norm $\|\mathbf{y}\|_\star$ (for its compuatation, $\|\mathbf{x}\| \leq 1$):

1. case $\|\mathbf{y}\|_\star > 1$

   using the definition of dual norm:

$$\begin{aligned} \|\mathbf{y}\|_\star &= \sup_{\|\mathbf{x}\| \leq 1} \langle \mathbf{y}, \mathbf{x} \rangle \\ \|\mathbf{y}\|_\star > 1 &\implies \sup_{\|\mathbf{x}\| \leq 1} \langle \mathbf{y}, \mathbf{x} \rangle > 1 \\ &\implies \exists \mathbf{x} \text{ s.t } \|\mathbf{x}\| \leq 1 : \underbrace{\langle \mathbf{y}, \mathbf{x} \rangle}_{>1} > \underbrace{\|\mathbf{x}\|}_{\leq 1} \end{aligned} \tag{64}$$

therefore, when we apply this $\mathbf{y}$ s.t., $\|\mathbf{y}\|_\star > 1$ to unconstrained $\mathbf{x}$ in $\sup_{\mathbf{x}} \{\langle \mathbf{y}, \mathbf{x}\rangle - \|\mathbf{x}\|\}$:

we can rewrite:

$$\mathbf{x} \rightarrow t\mathbf{x} \tag{65}$$

i.e., now $\mathbf{x}$ can have any norm, which is the case of $f^\star(\mathbf{y}) \equiv \sup_{\mathbf{x}} \{\langle \mathbf{y}, \mathbf{x}\rangle - \|\mathbf{x}\|\}$, i.e., $\mathbf{x}$ is unbounded:

$$\langle \mathbf{y}, t\mathbf{x}\rangle - \|t\mathbf{x}\| = t(\langle \mathbf{y}, \mathbf{x}\rangle - \|\mathbf{x}\|) \tag{66}$$

$t \rightarrow \infty \implies \|\mathbf{y}\|_\star$ to be unbounded, so we have shown half of the dual norm:

$$f^\star(\mathbf{y}) = \mathbb{1}_{\|\mathbf{x}\|_\star \leq 1}(\mathbf{x}) = \begin{cases} ? & \|\mathbf{y}\|_\star \leq 1 \\ +\infty & \|\mathbf{y}\|_\star > 1 \end{cases}. \tag{67}$$

2. case $\|\mathbf{y}\|_\star \leq 1$

   using the definition of dual norm:

$$\begin{aligned} \|\mathbf{y}\|_\star &= \sup_{\|\mathbf{x}\| \leq 1} \langle \mathbf{y}, \mathbf{x}\rangle \\ \|\mathbf{y}\|_\star \leq 1 &\implies \sup_{\|\mathbf{x}\| \leq 1} \langle \mathbf{y}, \mathbf{x}\rangle \leq 1 \\ &\implies \forall \mathbf{x} \text{ s.t } \|\mathbf{x}\| \leq 1: \quad \langle \mathbf{y}, \mathbf{x}\rangle - \|\mathbf{x}\| \leq 0 \end{aligned} \tag{68}$$

   since the term inside $\sup_{\mathbf{x}} \{\langle \mathbf{y}, \mathbf{x}\rangle - \|\mathbf{x}\|\}$ is ≤ 0, and therefore $\sup_{\mathbf{x}}$ makes it 0.

Combined the two, we have:

$$f^\star(\mathbf{y}) = \mathbb{1}_{\|\mathbf{y}\|_\star \leq 1}(\mathbf{y}) = \begin{cases} 0 & \|\mathbf{y}\|_\star \leq 1 \\ +\infty & \|\mathbf{y}\|_\star > 1 \end{cases}. \tag{69}$$

In general, indicator function of a convex set:

$$\mathbb{1}_{\mathcal{X}}(\mathbf{x}) = \begin{cases} 0, & \mathbf{x} \in \mathcal{X} \\ +\infty & \text{otherwise} \end{cases} \tag{70}$$

Also, it's important to note that $f(\mathbf{x}) = \|\mathbf{x}\|$ is un-constrained, but $f^\star(\mathbf{y})$ is constrained to $\mathcal{Y}$, i.e., a unit ball of a norm $\|\cdot\| = \{\mathbf{y} : \|\mathbf{y}\| \leq 1\}$.

2.3.2 $f = \|\cdot\|^2$ vector norm

$$f(\mathbf{x}) = \frac{1}{2}\|\mathbf{x}\|^2 \implies f^\star(\mathbf{y}) = \frac{1}{2}\|\mathbf{y}\|_\star^2 \tag{71}$$

1. show $f^\star(\mathbf{y}) \le \frac{1}{2}\|\mathbf{y}\|_\star^2$:

$$\begin{aligned}
\mathbf{y}^\top\mathbf{x} &\le \|\mathbf{y}\|_\star\|\mathbf{x}\| \qquad \forall\mathbf{x} \\
\implies \mathbf{y}^\top\mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2 &\le \|\mathbf{y}\|_\star\|\mathbf{x}\| - \frac{1}{2}\|\mathbf{x}\|^2 \qquad \text{both sides } -\frac{1}{2}\|\mathbf{x}\|^2
\end{aligned} \tag{72}$$

R.H.S is quadratic function in terms of $\|\mathbf{x}\|$. After solving, it has max occur at $\|\mathbf{x}\| = \|\mathbf{y}\|_\star\|$, substituting into:

$$\begin{aligned}
\max_{\|\mathbf{x}\|}\left(\|\mathbf{y}\|_\star\|\mathbf{x}\| - \frac{1}{2}\|\mathbf{x}\|^2\right) &= \|\mathbf{y}\|_\star^2 - \frac{1}{2}\|\mathbf{y}\|_\star^2 \\
&= \frac{1}{2}\|\mathbf{y}\|_\star^2 \\
\implies f^\star(\mathbf{y}) = \mathbf{y}^\top\mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2 &\le \frac{1}{2}\|\mathbf{y}\|_\star^2 \qquad \forall\mathbf{x}
\end{aligned} \tag{73}$$

2. $f^\star(\mathbf{y}) \ge \frac{1}{2}\|\mathbf{y}\|_\star^2$

Let $\mathbf{x}$ be any chosen vector with $\mathbf{y}^\top\mathbf{x} = \|\mathbf{y}\|_\star\|\mathbf{x}\|$ (think this as when norm on the R.H.S is fixed, then one may choose the value of $\mathbf{y}$ and $\mathbf{x}$ to change the "direction". In $\|\cdot\|_2$ case, they should have $\cos(\theta) = 0$). W.l.o.g, we then scale so that $\|\mathbf{x}\| = \|\mathbf{y}\|_\star$. then we have, for this $\mathbf{x}$:

$$\begin{aligned}
\|\mathbf{x}\| &= \|\mathbf{y}\|_\star \\
\implies \mathbf{y}^\top\mathbf{x} &= \|\mathbf{x}\|^2 \qquad \because \mathbf{y}^\top\mathbf{x} = \|\mathbf{y}\|_\star\|\mathbf{x}\| \\
&= \frac{1}{2}\|\mathbf{x}\|^2 + \frac{1}{2}\|\mathbf{y}\|_\star^2 \\
\implies \mathbf{y}^\top\mathbf{x} - \frac{1}{2}\|\mathbf{x}\|^2 &= \frac{1}{2}\|\mathbf{y}\|_\star^2 \\
\implies f^\star(\mathbf{y}) &\ge \frac{1}{2}\|\mathbf{y}\|_\star^2
\end{aligned} \tag{74}$$

References

[1] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro, "The implicit bias of gradient descent on separable data," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 2822–2878, 2018.