A Tutorial on Gradient Descend

Richard Xu

August 5, 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,

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

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!

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.