Lecture Notes 12: Regularization and Optimization

Instructor: Ashok Cutkosky

One of the most common strategies in building machine learning systems is *regularization*. For example, it is typical to encourage model parameter $\mathbf{w}$ to be small by augmenting the loss with a penalty for large $\mathbf{w}$ as in the following:

$$\mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

where the λ value is sometimes called the weight-decay parameter, or the l2-regularization constant. Typically, this is justified in machine learning courses by claiming that small $\mathbf{w}$ values are in some sense "simpler" and so avoid overfitting. Making this idea concrete is deceptively difficult, involving a foray into statistical learning theory. However, it turns out that we can also justify regularization from an optimization point of view: adding regularization makes the function *easier to optimize*. This type of justification is in many ways simpler than the statistical learning theory viewpoint - the standard SLT approach would not really suggest why one should use a penalty proportional to $\|\mathbf{w}\|^2$ rather than, say $\|\mathbf{w}\|$ or $\|\mathbf{w}\|^4$ or $\sqrt{\|\mathbf{w}\|}$, all of which still encourage $\mathbf{w}$ to be small and so therefore "simpler".

Let's try to get some intuition behind why adding $\|\mathbf{w}\|^2$ to the objective might make it easier to optimizer. First, let us define

$$\mathbf{w}_\star = \operatorname{argmin} \mathcal{L}(\mathbf{w})$$
$$\mathbf{w}_\star^\lambda = \operatorname{argmin} \mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$$

Intuitively, we might expect that $\|\mathbf{w}_\star^\lambda\| \le \|\mathbf{w}_\star\|$. This is indeed the case:

Lemma 1. *Suppose $\lambda_1 \ge \lambda_2$. Then $\|\mathbf{w}_\star^{\lambda_1}\| \le \|\mathbf{w}_\star^{\lambda_2}\|$.*

Proof.

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_\star^{\lambda_1}) + \frac{\lambda_1}{2}\|\mathbf{w}_\star^{\lambda_1}\|^2 &\le \mathcal{L}(\mathbf{w}_\star^{\lambda_2}) + \frac{\lambda_1}{2}\|\mathbf{w}_\star^{\lambda_2}\|^2\\
&\le \mathcal{L}(\mathbf{w}_\star^{\lambda_2}) + \frac{\lambda_2}{2}\|\mathbf{w}_\star^{\lambda_2}\|^2 + \frac{\lambda_1-\lambda_2}{2}\|\mathbf{w}_\star^{\lambda_2}\|^2\\
&\le \mathcal{L}(\mathbf{w}_\star^{\lambda_1}) + \frac{\lambda_2}{2}\|\mathbf{w}_\star^{\lambda_1}\|^2 + \frac{\lambda_1-\lambda_2}{2}\|\mathbf{w}_\star^{\lambda_2}\|^2\\
\frac{\lambda_1-\lambda_2}{2}\|\mathbf{w}_\star^{\lambda_1}\|^2 &\le \frac{\lambda_1-\lambda_2}{2}\|\mathbf{w}_\star^{\lambda_2}\|^2
\end{aligned}$$

now divide by $\frac{\lambda_1-\lambda_2}{2} \ge 0$ to conclude the result. □

Further, we have the following:

Lemma 2. *Suppose that $\mathcal{L}$ is G-Lipschitz and differentiable. Then*

$$\|\mathbf{w}_\star^\lambda\| \le \frac{G}{\lambda}$$

Proof. Since $\mathcal{L}$ is differentiable, so is $\mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$, and the derivative at $\mathbf{w}_\star^\lambda$ should be zero. Thus:

$$\begin{aligned}
\nabla\mathcal{L}(\mathbf{w}_\star^\lambda) + \lambda\mathbf{w}_\star^\lambda &= 0\\
\|\mathbf{w}_\star^\lambda\| &= \frac{\|\nabla\mathcal{L}(\mathbf{w}_\star^\lambda)\|}{\lambda}\\
&\le \frac{G}{\lambda}
\end{aligned}$$

□

This Lemma tells us that as λ increases, there is ball of radius G/λ for which we can certify that the solution lies in the ball. Therefore in some sense the "search space" is decreasing as λ increases, thus making the task of finding $\mathbf{w}_\star^\lambda$ easier.

However, this is far from a rigorous proof. To make things rigorous, we will introduce the notion of *strong convexity*.

Definition 3. *A twice-differentiable function $\mathcal{L}$ is μ-strongly convex if $\nabla^2 L(x) \succeq \mu I$ (recall that $A \succeq B$ if $v\top Av \geq v^\top Bv$ for all v).*

Strong convexity can also be characterized by the following:

Lemma 4. *A twice differentiable convex function $\mathcal{L}$ is μ-strongly convex if and only if for all x and y,*

$$\mathcal{L}(y) \geq \mathcal{L}(x) + \rangle \nabla \mathcal{L}(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$$

Proof. First, suppose that $\mathcal{L}$ is strongly-convex. Then we proceed in a manner very analogous to the smoothness lemma. Apply the fundamental theorem of calculus twice to obtain:

$$\begin{aligned}
\mathcal{L}(y) &= \mathcal{L}(x) + \int_0^1 \langle \nabla \mathcal{L}(x + t(y - x)), (y - x)\rangle \, dt \\
&= \mathcal{L}(x) + \langle \nabla \mathcal{L}(x), (y - x)\rangle + \int_0^1 \int_0^1 t(y - x)^\top \nabla^2 \mathcal{L}(x + kt(y - x))(y - x)\rangle \, dk \, dt \\
&\geq \mathcal{L}(x) + \langle \nabla \mathcal{L}(x), (y - x)\rangle + \mu\|x - y\|^2 \int_0^1 \int_0^1 t \, dk \, dt \\
&= \mathcal{L}(x) + \langle \nabla \mathcal{L}(x), (y - x)\rangle + \frac{\mu\|x - y\|^2}{2}
\end{aligned}$$

For the other direction, suppose $\mathcal{L}(y) \geq \mathcal{L}(x) + \rangle \nabla \mathcal{L}(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$ for all x and y. Then consider the function $f(t) = \mathcal{L}(x + t\delta)$ for some arbitrary x and vector δ. We have:

$$\begin{aligned}
f'(t) &= \langle \nabla \mathcal{L}(x + t\delta), \delta\rangle \\
f''(t) &= \delta^\top \nabla^2 \mathcal{L}(x + t\delta)\delta
\end{aligned}$$

Further, the condition $\mathcal{L}(y) \geq \mathcal{L}(x) + \rangle \nabla \mathcal{L}(x), y - x\rangle + \frac{\mu}{2}\|y - x\|^2$ applied to both the pair $x, x + t\delta$ and $x + t\delta, x$ implies:

$$\begin{aligned}
f(h) &\geq f(0) + f'(0)h + \frac{\mu}{2}h^2\|\delta\|^2 \\
f(0) &\geq f(h) - f'(h)h + \frac{\mu}{2}h^2\|\delta\|^2 \\
\frac{f'(h) - f(0)}{h} &\geq \mu\|\delta\|^2
\end{aligned}$$

Now, be definition of derivative:

$$\begin{aligned}
\delta^\top \nabla^2 \mathcal{L}(x)\delta &= f''(0) \\
&= \lim_{h \to 0} \frac{f'(h) - f'(0)}{h} \\
&\geq \mu\|\delta\|^2
\end{aligned}$$

□

This lemma provides a kind of opposite to the standard smoothness lemma: now the inequality is in the other direction. This will enable faster convergence by having the gradient norm be an *upper bound* for the suptimality $\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w})_\star$. Specifically, we have the following result:

Lemma 5. *Suppose that $\mathcal{L}$ is μ-strongly convex. Then for all $\mathbf{w}$,*

$$\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star) \le \frac{\|\nabla \mathcal{L}(\mathbf{w})\|^2}{2\mu}$$

Proof. By Lemma 4, we have:

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_\star) &\ge \mathcal{L}(\mathbf{w}) + \langle \nabla \mathcal{L}(\mathbf{w}), \mathbf{w}_\star - \mathbf{w}\rangle + \frac{\mu}{2}\|\mathbf{w}_\star - \mathbf{w}\|^2\\
&\ge \inf_\delta \mathcal{L}(\mathbf{w}) + \langle \nabla \mathcal{L}(\mathbf{w}), \delta\rangle + \frac{\mu}{2}\|\delta\|^2\\
&= \mathcal{L}(\mathbf{w}) - \frac{\|\nabla \mathcal{L}(\mathbf{w})\|^2}{2\mu}
\end{aligned}$$

rearranging proves the result □

Thus, using our standard smoothness result we have the gradient descent with learning rate $\eta = \frac{1}{H}$ on an H-smooth loss satisfies:

$$\mathcal{L}(\mathbf{w}_{t+1}) \le \mathcal{L}(\mathbf{w}_t) - \frac{\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2}{2H}$$

Plugging in the result of Lemma 5 yields:

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_\star) &\le \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) - \frac{\|\nabla \mathcal{L}(\mathbf{w}_t)\|^2}{2H}\\
&\le \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) - \frac{\mu}{H}(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star))\\
&\le (1 - \frac{\mu}{H})(\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star))
\end{aligned}$$

unrolling this for t iterations:

$$\begin{aligned}
&\le \left(1 - \frac{\mu}{H}\right)^t (\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))\\
&= \exp(t \log \left(1 - \frac{\mu}{H}\right) (\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))\\
&\le \exp\left(-\frac{\mu}{H}t\right) (\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))
\end{aligned}$$

Thus, we have shown:

Theorem 6. *Suppose $\mathcal{L}$ is μ-strongly convex and H-smooth. Then gradient descent with learning rate $\eta = \frac{1}{H}$ guarantees:*

$$\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_\star) \le \exp\left(-\frac{\mu}{H}t\right) (\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))$$

This is somewhat remarkable: the error is decreasing *exponentially fast*! All of our previous rates have only had a polynomial dependence on t, so this is truly much much faster than was previously possible. Of course, there is a catch: the exponent depends on the ratio $\frac{\mu}{H}$. This quantity is often called the *condition number* of the problem, and in practice on machine learning problems it can be very very small. In fact, on a dataset of size N it is not unusual to have $\mu \propto \frac{1}{\sqrt{N}}$, so that the condition number is also $\propto 1/\sqrt{N}$. This is so small that in fact the exponential rate may look only polynomial!

Strong convexity and Regularization

The study of strongly-convex functions is linked to regularization via the following important result:

Lemma 7. *Suppose that $\mathcal{L}$ is a convex function. Then $\mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$ is λ-strongly convex.*

Proof. Let $F(\mathbf{w}) = \mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$. Then $\nabla^2 F(\mathbf{w}) = \nabla^2 \mathcal{L}(\mathbf{w}) + \lambda I$. Thus for all v, $v^\top \nabla^2 F(\mathbf{w}) v = v^\top \nabla^2 \mathcal{L}(\mathbf{w}) v + \lambda \|v\|^2 \geq \lambda \|v\|^2$, where the final inequality follows since $\mathcal{L}$ is convex. □

This means that we can *force a convex loss to be strongly-convex by adding regularization*. However, there is an inherent tradeoff in doing so: by adding regularization, we bias the objective away from the true optimum. Thus, we want to add as little regularization as possible while still allowing for fast optimization. This is what results in very small condition numbers - the condition number will be $\frac{\lambda}{H}$, and we want to make λ as small as is feasible.

Strong-convexity and distance from optimum

Strong convexity also implies another important property:

Lemma 8. *Suppose $\mathcal{L}$ is μ-strongly convex. Then for all $\mathbf{w}$:*

$$\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star) \geq \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}_\star\|^2$$

Proof. From Lemma 4:

$$\mathcal{L}(\mathbf{w}) \geq \mathcal{L}(\mathbf{w}_\star) + \langle \nabla \mathcal{L}(\mathbf{w}_\star), \mathbf{w} - \mathbf{w}_\star \rangle + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}_\star\|^2$$

Now note that $\nabla \mathcal{L}(\mathbf{w}_\star) = 0$ and rearrange to see the result. □

This is a very powerful statement: it allows us show not only convergence in objective value $\mathcal{L}(\mathbf{w}) \to \mathcal{L}(\mathbf{w}_\star)$, but also convergence in parameter value $\mathbf{w} \to \mathbf{w}_\star$.

How much regularization should we add?

Adding regularization makes the objective strongly convex, which makes the optimization process faster, but it also makes our solutions subtly incorrect. Thus there is a natural question of what the "right" amount of regularization is that would optimally tradeoff between having an accurate solution and obtaining the solution quickly.

In particular, we can ask the following question: Given that we are allowed to perform T gradient computations, how much regularization should we add to obtain the smallest possible error?

To address this question, we need to analyze two different sources of error: error from the optimization algorithm, and error from adding regularization. Specifically, we have:

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) &= \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star^\lambda) + \mathcal{L}(\mathbf{w}_\star^\lambda) - \mathcal{L}(\mathbf{w}_\star) \\
&= \mathcal{L}(\mathbf{w}_t) + \frac{\lambda}{2}\|\mathbf{w}_t\|^2 - \mathcal{L}(\mathbf{w}_\star^\lambda) - \frac{\lambda}{2}\|\mathbf{w}_\star^\lambda\| + \mathcal{L}(\mathbf{w}_\star^\lambda) - \mathcal{L}(\mathbf{w}_\star) + \frac{\lambda}{2}(\|\mathbf{w}_\star^\lambda\|^2 - \|\mathbf{w}_t\|^2) \\
&\leq \mathcal{L}(\mathbf{w}_t) + \frac{\lambda}{2}\|\mathbf{w}_t\|^2 - (\mathcal{L}(\mathbf{w}_\star^\lambda) + \frac{\lambda}{2}\|\mathbf{w}_\star^\lambda\|^2) + \mathcal{L}(\mathbf{w}_\star^\lambda) - \mathcal{L}(\mathbf{w}_\star) + \frac{\lambda}{2}\|\mathbf{w}_\star\|^2 \\
&\leq \mathcal{L}(\mathbf{w}_t) + \frac{\lambda}{2}\|\mathbf{w}_t\|^2 - (\mathcal{L}(\mathbf{w}_\star^\lambda) + \frac{\lambda}{2}\|\mathbf{w}_\star^\lambda\|^2) + \frac{\lambda}{2}\|\mathbf{w}_\star\|^2
\end{aligned}$$

where the last inequality holds since $\mathcal{L}(\mathbf{w}_\star) \leq \mathcal{L}(\mathbf{w}_\star^\lambda)$ by definition, and the second-to-last inequality is due to Lemma 1. Now, let us suppose we start at $\mathbf{w}_1 = 0$ and use gradient descent on the loss $\mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}\|^2$. Then we have:

$$\begin{aligned}
\mathcal{L}(\mathbf{w}_1) + \frac{\lambda}{2}\|\mathbf{w}_1\|^2 - (\mathcal{L}(\mathbf{w}_\star^\lambda) + \frac{\lambda}{2}\|\mathbf{w}_\star^\lambda\|^2) &\leq \mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star^\lambda) - \frac{\lambda}{2}\|\mathbf{w}_\star^\lambda\|^2 \\
&\leq \mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star)
\end{aligned}$$

Further, observe that $\mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}_\star\|^2$ is $\lambda + H$-smooth if $\mathcal{L}$ is H smooth. To see this last fact, define $F(\mathbf{w}) = \mathcal{L}(\mathbf{w}) + \frac{\lambda}{2}\|\mathbf{w}_\star\|^2$. Then we have $\nabla^2 F(\mathbf{w}) = \nabla^2 \mathcal{L}(\mathbf{w}) + \lambda I$, and since $\nabla^2 \mathcal{L}(\mathbf{w}) \preceq HI$ since $\mathcal{L}$ is H-smooth, this implies $\nabla^2 F(\mathbf{w}) \leq (H + \lambda) I$. We also have that F is λ-strongly convex. Therefore by Theorem 6:

$$\begin{aligned}
F(\mathbf{w}_t) - F(\mathbf{w}_\star^\lambda) &\leq \exp\left(-\frac{\lambda}{H+\lambda}t\right)(F(\mathbf{w}_1) - F(\mathbf{w}_\star^\lambda)) \\
\mathcal{L}(\mathbf{w}_t) + \frac{\lambda}{2}\|\mathbf{w}_t\|^2 - (\mathcal{L}(\mathbf{w}_\star^\lambda) + \frac{\lambda}{2}\|\mathbf{w}_\star^\lambda\|^2) &\leq \exp\left(-\frac{\lambda}{H+\lambda}t\right)(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))
\end{aligned}$$

Putting this all together:

$$\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \le \exp\left(-\frac{\lambda}{H+\lambda}t\right)(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star)) + \frac{\lambda}{2}\|\mathbf{w}_\star\|^2$$

Now, notice that:

$$\begin{aligned}\mathcal{L}(\mathbf{w}_1) &\le \mathcal{L}(\mathbf{w}_\star) + \langle \nabla\mathcal{L}(\mathbf{w}_\star), \mathbf{w}_1 - \mathbf{w}_\star\rangle + \frac{H}{2}\|\mathbf{w}_1 - \mathbf{w}_\star\|^2\\ &\le \mathcal{L}(\mathbf{w}_\star) + \frac{H}{2}\|\mathbf{w}_\star\|^2\end{aligned}$$

where we have used $\nabla\mathcal{L}(\mathbf{w}_\star) = 0$ and $\mathbf{w}_1 = 0$. Thus our bound becomes:

$$\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \le \frac{\|\mathbf{w}_\star\|^2}{2}\left(H\exp\left(-\frac{\lambda}{H+\lambda}t\right) + \lambda\right)$$

Now to balance the two terms here exactly is somewhat complicated, but we can essentially just guess. To start, notice that $\lambda \ge H$ is unlikely to be the right answer: in this case just the λ term would already be so big that we would be not convergence. So, let us restrict attention to $\lambda \le H$. In this case, we have

$$\exp\left(-\frac{\lambda}{H+\lambda}t\right) \le \exp\left(-\frac{\lambda}{2H}t\right)$$

Next, let us set $c = \exp(\lambda t/2H)$ so that $\lambda = \frac{2H}{t}\log(c)$. Then we have:

$$H\exp\left(-\frac{\lambda t}{2H}\right) \le \frac{H}{c}$$

Thus, if we consider $\log(c)$ to be an unimportant log factor, when setting $\frac{H}{c} = \lambda$ to balance terms, we would get $c = t/2$. So, if t is large enough that $\frac{2H}{t}\log(t/2) \le H$, we can set $\lambda = \frac{2H}{t}\log(t/2)$ to obtain:

$$\begin{aligned}\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) &\le \frac{\|\mathbf{w}_\star\|^2}{2}\left(H\exp\left(-\frac{\lambda}{H+\lambda}t\right) + \lambda\right)\\ &\le \frac{\|\mathbf{w}_\star\|^2}{2}\left(\frac{2H}{t} + \frac{2H\log(t/2)}{t}\right)\\ &= O\left(\frac{H\|\mathbf{w}_\star\|^2\log(t)}{t}\right)\end{aligned}$$

This is almost the same rate we obtained with ordinary gradient descent, but now there is an extra log factor.

So, what was the point of that? It seems that we actually didn't gain anything from adding the regularization. It's actually not so clear: remember that all of our results so far have been *upper bounds*, so it might easily hold that a small amount of regularization (smaller than suggested by this theory), would actually result in a much better output than suggested here. In particular, it might be that $\mathcal{L}$ is in some sense "almost" already strongly convex except at a few locations $\mathbf{w}$. By adding a small amount of regularization, we may encourage gradient descent to somehow quickly bypass those locations and result in an overall faster convergence rate. However, it is very difficult to predict when this will occur in practice, so typically the optimal value for λ is tuned manually.