1 Recap of Algorithmic Stability

1. A learning algorithm \( \hat{y} \) is said to be Uniform Replace One (URO) stable with rate \( \epsilon_{\text{stable}} \) if

\[
\frac{1}{n} \sum_{t=1}^{n} \left| \ell(\hat{y}(S), z_t') - \ell(\hat{y}(S^{(t)}), z_t') \right| \leq \epsilon_{\text{stable}}(n)
\]

where \( S^{(t)} \) is a sample identical to \( S \) except on the \( t \)'th entry where \( z_t \) is replaced by \( z_t' \).

2. If a learning algorithm \( \hat{y} \) is URO stable with rate \( \epsilon_{\text{stable}} \) then it generalizes at the same rate.

3. If a learning algorithm \( \hat{y} \) is URO stable with rate \( \epsilon_{\text{stable}} \) and is an AERM with rate \( \epsilon_{\text{AERM}} \) then:

\[
\mathbb{E}_S [L(\hat{y}(S))] - \inf_{f \in \mathcal{F}} L(f) \leq \epsilon_{\text{stable}}(n) + \epsilon_{\text{AERM}}(n)
\]

4. If there exists an algorithm \( \hat{y} \) such that for any distribution \( \mathcal{D} \), for sample drawn from this distribution:

\[
\mathbb{E}_S [L(\hat{y}(S))] - \inf_{f \in \mathcal{F}} L(f) \leq \epsilon_{\text{rate}}(n)
\]

then, there exists an algorithm \( \hat{\hat{y}} \) s.t.

(a) \( \hat{\hat{y}} \) is \( \epsilon_{\text{stable}}(n) = \frac{2}{\sqrt{n}} \) URO stable

(b) \( \hat{\hat{y}} \) is an AERM with rate \( \epsilon_{\text{AERM}}(n) = 2\epsilon_{\text{rate}}(n^{1/4}) + O\left(\frac{1}{\sqrt{n}}\right) \)

Thus existence of a stable AERM is both necessary and sufficient condition for statistical Learnability.

2 Stability of ERM for Strongly convex objectives and more

Assumption 1. Assume that for sample \( S \) drawn, it is true that for any \( f \in \mathcal{F} \)

\[
\mathbb{E}_S \left[ \hat{L}_S(f) - \min_{f \in \mathcal{F}} \hat{L}_S(f) \right] \geq \frac{\lambda}{2} \mathbb{E}_S \left[ \| f - \hat{f}_S \|^2 \right]
\]

where \( \hat{f}_S = \arg\min_{f \in \mathcal{F}} \hat{L}_S(f) \)
Note that if our functions were strongly convex then the above assumption would be true
deterministically. This is because, by strong convexity
\[
\hat{L}_S(\hat{f}_S) \leq \hat{L}_S(f) + \nabla \hat{L}_S(\hat{f}_S)^\top (\hat{f}_S - f) - \frac{\lambda}{2} \| \hat{f}_S - f \|^2
\]
Rearranging we get the assumption. However, the assumption we need is milder than strong
convexity. For instance, one point strong convexity empirically would also imply the assumption.

**Theorem 2.** Assume that our loss is \( L \)-Lipschitz and that Assumption 1 holds, then, for any \( t \),
\[
\mathbb{E}_S \left[ \sup_z \left| \ell(\hat{y}(S(t)), z) - \ell(\hat{y}(S), z) \right| \right] \leq \frac{4L^2}{\lambda n}
\]
That is, the ERM algorithm is stable in expectation.

**Proof.**
\[
\hat{L}_S(\hat{y}(S(t))) - \hat{L}_S(\hat{y}(S)) = \frac{1}{n} \left( \ell(\hat{y}(S(t)), z_t) - \ell(\hat{y}(S), z_t) \right) + \frac{1}{n} \sum_{s \in [n] \setminus \{t\}} \left( \ell(\hat{y}(S(t)), z_s) - \ell(\hat{y}(S), z_s) \right)
\]
\[
= \frac{1}{n} \left( \ell(\hat{y}(S(t)), z_t) - \ell(\hat{y}(S), z_t) \right) + \frac{1}{n} \left( \ell(\hat{y}(S), z'_t) - \ell(\hat{y}(S(t)), z'_t) \right)
\]
\[
+ \hat{L}_{S(t)}(\hat{y}(S(t))) - \hat{L}_{S(t)}(\hat{y}(S))
\]
\[
\leq \frac{1}{n} \left( \ell(\hat{y}(S(t)), z_t) - \ell(\hat{y}(S), z_t) \right) + \frac{1}{n} \left( \ell(\hat{y}(S), z'_t) - \ell(\hat{y}(S(t)), z'_t) \right)
\]
\[
\leq \frac{2L}{n} \left\| \hat{y}(S(t)) - \hat{y}(S) \right\|
\]
On the other hand, from our premise,
\[
\mathbb{E}_S \left[ \hat{L}_S(\hat{y}(S(t))) - \hat{L}_S(\hat{y}(S)) \right] \geq \frac{\lambda}{2} \left\| \hat{y}(S(t)) - \hat{y}(S) \right\|^2
\]
Hence we conclude that
\[
\mathbb{E}_S \left[ \left\| \hat{y}(S(t)) - \hat{y}(S) \right\| \right] \leq \frac{4L}{\lambda n}
\]
Hence we conclude that:
\[
\mathbb{E}_S \left[ \sup_z \left| \ell(\hat{y}(S(t)), z) - \ell(\hat{y}(S), z) \right| \right] \leq \frac{4L^2}{\lambda n}
\]
In fact the above proof shows that uniform stability holds for strongly convex objectives.
3 Stability of Stochastic Gradient Descent

Given a sample $S$, let us consider the multi-epoch SGD algorithm that uses a prefixed order over instances. That is: at iteration $t$,

$$\hat{y}_{t+1} = \hat{y}_t - \eta \nabla \ell(\hat{y}_t, z_{t \pmod{n} + 1})$$

In short, we will use $G_t$ to denote the above update. That is $\hat{y}_{t+1} = G_t(\hat{y}_t)$.

**Definition 1.** We say that an update rule $G$ is $\alpha$-expansive if:

$$\sup_{f,g \in F} \frac{\|G(f) - G(g)\|}{\|f - g\|} \leq \alpha$$

And we say that an update rule is $\sigma$-bounded if

$$\sup_{f \in F} \|f - G(f)\| \leq \sigma$$

**Lemma 3.** Consider two sequences of updates $G_1, \ldots, G_T$ and $G'_1, \ldots, G'_T$ with $\hat{y}_{t+1} = G_t(\hat{y}_t)$ and $\hat{y}'_{t+1} = G'_t(\hat{y}'_t)$. Let $\delta_t = \|\hat{y}_t - \hat{y}'_t\|$ and assume that $\delta_1 = 0$ (that is both algorithms are initialized at same point). Then we have:

$$\delta_{t+1} \leq \begin{cases} 
\alpha \delta_t & \text{if } G_t = G'_t \text{ is } \alpha\text{-expansive} \\
\delta_t + 2\sigma & \text{if } G_t, G'_t \text{ are } \sigma\text{-bounded}
\end{cases}$$

**Proof.** if $G'_t = G_t$ is $\alpha$-expansive, then

$$\delta_{t+1} = \|G_t(\hat{y}_t) - G'_t(\hat{y}'_t)\| \leq \alpha \delta_t$$

Also note that for the second case,

$$\delta_{t+1} = \|G_t(\hat{y}_t) - G'_t(\hat{y}'_t)\| \\
\leq \|G_t(\hat{y}_t) - \hat{y}_t + \hat{y}_t - G'_t(\hat{y}'_t)\| + \|\hat{y}_t - \hat{y}'_t\| \\
\leq \delta_t + \|G_t(\hat{y}_t) - \hat{y}_t\| + \|\hat{y}_t' - G'_t(\hat{y}'_t)\| \\
\leq \delta_t + 2\sigma$$

**Theorem 4.** Assume that an algorithm uses update of the form $\hat{y}_{t+1} = G_t(\hat{y}_t)$ where $G_t(f) = f - \eta \nabla \ell(f, z_{t \pmod{n} + 1})$. Now if the gradient updates $G_t$’s are $\alpha$-expansive and $\sigma$-bounded, then for any $j \in [n]$,

$$\sup_z \mathbb{E}_S \left[ \ell(\hat{y}_T(S), z) - \ell(\hat{y}_T(S^{(j)}), z) \right] \leq \frac{4T}{n} \sigma$$

**Proof.** We start using the Lipschitz property to note that:

$$\sup_z \mathbb{E}_S \left[ \ell(\hat{y}_T(S), z) - \ell(\hat{y}_T(S^{(j)}), z) \right] \leq L \mathbb{E}_S \left[ \|\hat{y}_T - \hat{y}'_T\| \right]$$

3
Let $G_1, \ldots, G_T$ be the sequence of updates using sample $S$ in SGD and let $G'_1, \ldots, G'_T$ be the updates with $S^{(j)}$. Note that $G_t \neq G'_t$ only when $t \mod n + 1 = j$ and otherwise the updates are identical. Hence, using the previous lemma (and crudely upper bounding),

$$
\mathbb{E} [\delta_T] \leq \eta^{T-T/n} \left( \delta_1 + \frac{2T}{n} \sigma \right) + \frac{2T}{n} \sigma
$$

Now if $\eta \leq 1$ then we conclude that

$$
\mathbb{E} [\delta_T] \leq \frac{4T}{n} \sigma
$$

Hence we get stability of

$$
\sup_z \mathbb{E}_S \left[ |\ell(\hat{y}_T(S), z) - \ell(\hat{y}_T(S^{(j)}), z)| \right] \leq \frac{4T}{n} \sigma
$$

Lemma 5. For any $L$-Lipschitz objective, SGD update is $\eta L$ bounded and if loss function is both convex and $H$-smooth then update is $1$-expansive as long as step size $\eta \leq 2/H$

Proof. First note that for boundedness,

$$
\|G_t(f) - f\| = \|\eta \nabla \ell(f, z)\| \leq \eta L
$$

Next note that

$$
\|G_t(f) - G_t(g)\|^2 = \|g - f\|^2 - 2\eta \langle \nabla \ell(f, z_t) - \nabla \ell(g, z_t), f - g \rangle + \eta^2 \|\nabla \ell(f, z_t) - \nabla \ell(g, z_t)\|^2
\leq \|g - f\|^2 - \left( \frac{2\eta}{H} + \eta^2 \right) \|\nabla \ell(f, z_t) - \nabla \ell(g, z_t)\|^2
\leq \|g - f\|^2
$$

where the second inequality is a consequence of smoothness + convexity. 

Putting all this together, the stability of SGD for smooth convex loss is given by

$$
\sup_z \mathbb{E}_S \left[ |\ell(\hat{y}_T(S), z) - \ell(\hat{y}_T(S^{(j)}), z)| \right] \leq \frac{4L \eta T}{n}
$$