Class 6
Early Stopping

Carlo Ciliberto
Department of Computer Science, UCL

November 9, 2017

Characterization of Convexity

Lemma. $F : \mathcal{H} \to \mathbb{R}$ differentiable. Then the following statements are equivalent

(i) F is convex

(ii) $F(w) - F(v) \geq \langle \nabla F(v), w - v \rangle_{\mathcal{H}} \qquad \forall w, v \in \mathcal{H}$

(iii) $\langle \nabla F(w) - \nabla F(v), w - v \rangle_{\mathcal{H}} \geq 0 \qquad \forall w, v \in \mathcal{H}$

Characterization of Convexity (i) $\Rightarrow$ (ii)

Assume (i). By convexity, for every $\theta \in (0,1]$

$$F(v+\theta(w-v)) \leq F(\theta w + (1-\theta)v) \leq \theta F(w) + (1-\theta)F(v)$$

Therefore, by bringing $F(v)$ on the left side and dividing with respect to θ, we have

$$\frac{F(v+\theta(w-v)) - F(v)}{\theta} \leq F(w) - F(v)$$

By sending $\theta \to 0$ we have

$$\lim_{\theta\to 0} \frac{F(v+\theta(w-v)) - F(v)}{\theta} = \langle \nabla F(v), w-v\rangle_{\mathcal{H}} \leq F(w) - F(v)$$

as desired.

Characterization of Convexity (ii) $\Rightarrow$ (iii)

By (ii), for any $w, v \in \mathcal{H}$ we have

- $F(w) - F(v) \geq \langle \nabla F(v), w - v \rangle_{\mathcal{H}}$
- $F(v) - F(w) \geq \langle \nabla F(w), v - w \rangle_{\mathcal{H}} = -\langle \nabla F(w), w - v \rangle_{\mathcal{H}}$

By summing the two inequalities we have

$$0 \geq \langle \nabla F(v) - F(w), w - v \rangle_{\mathcal{H}}$$

Or equivalently

$$\langle \nabla F(v) - F(w), v - w \rangle_{\mathcal{H}} \geq 0$$

as desired.

Characterization of Convexity (iii) $\Rightarrow$ (i)

Assume (iii). Define $\phi : [0, 1] \to \mathbb{R}$, $\phi(\theta) = F(v + \theta(w - v))$. Then

$$\phi'(\theta) = \langle \nabla F(v + \theta(w - v)), w - v \rangle_{\mathcal{H}}.$$

For any $0 \leq \alpha < \beta \leq 1$, let $v_\alpha = v + \alpha(w - v)$ and $v_\beta = v + \beta(w - v)$. Then,

$$\begin{aligned} \phi'(\beta) - \phi'(\alpha) &= \langle \nabla F(v_\beta) - \nabla(v_\alpha), w - v \rangle_{\mathcal{H}} \\ &= \frac{1}{\beta - \alpha} \langle \nabla F(v_\beta) - \nabla(v_\alpha), v_\beta - v_\alpha \rangle_{\mathcal{H}} \geq 0 \end{aligned}$$

where the last inequality is a consequence of (iii).

This implies that ϕ' is a non-decreasing function on $[0, 1]$.

Characterization of Convexity (iii) $\Rightarrow$ (i) (Continued)

Lemma. Let $\phi : [0, 1] \to \mathbb{R}$ differentiable with ϕ' non-decreasing. Then ϕ is convex.

Proof. Let $x, z \in [0, 1]$. For $\theta \in [0, 1]$ define

$$\psi(z) = \theta\phi(x) + (1 - \theta)\psi(z) - \psi(\theta x + (1 - \theta)z)$$

Then

$$\psi'(z) = (1 - \theta)(\psi(z) - \psi(\theta x + (1 - \theta)z)$$

Since ϕ' is non-decreasing we have that for $z \leq x$, $\psi'(z) \leq 0$ while for $z \geq x$, $\psi'(z) \geq 0$. Therefore the function ψ has a minimum in $z = x$.

By construction, $\psi(x) = 0$ and therefore, for any $y \in [0, 1]$ we have $\psi(y) \geq \psi(x) = 0$, which implies the convexity of ϕ as desired.

Characterization of Convexity (iii) $\Rightarrow$ (i) (Continued)

We have shown that (iii) implies $\phi(\theta) = F(v + \theta(w - v))$ convex. In particular, by writing $\theta = \theta \cdot 1 + (1 - \theta) \cdot 0$, we have

$$F(v + \theta(w - v)) = \phi(\theta) \leq \theta\phi(1) + (1 - \theta)\phi(0) = \theta F(w) + (1 - \theta)F(v)$$

which proves the convexity of F as desired

Quadratic Upper Bound

Lemma. $F : \mathcal{H} \to \mathbb{R}$ convex M-smooth. The function $G : \mathcal{H} \to \mathbb{R}$

$$G(w) = \frac{M}{2} \|w\|^2 - F(w)$$

is convex.

Proof. By the M-smoothness of F combined with Cauchy-Swartz inequality, we have

$$\langle \nabla F(w) - \nabla F(v), w - v \rangle_{\mathcal{H}} \leq \|\nabla F(w) - \nabla F(v)\|_{\mathcal{H}} \|w - v\|_{\mathcal{H}} \leq M \|w - v\|_{\mathcal{H}}^2$$

Then, since $\nabla G(w) = Mw - \nabla F(w)$, we have $\forall w, v \in \mathcal{H}$

$$\begin{aligned} \langle \nabla G(w) - \nabla G(v), w - v \rangle_{\mathcal{H}} &= \langle M(w - v) - \nabla F(v) + \nabla F(w), w - v \rangle_{\mathcal{H}} \\ &= M \|w - v\|_{\mathcal{H}}^2 - \langle \nabla F(w) - \nabla F(v), w - v \rangle_{\mathcal{H}} \geq 0 \end{aligned}$$

which implies the convexity of G as desired.

Consequence of Quadratic Upper Bound

Lemma. $F : \mathcal{H} \to \mathbb{R}$ convex M-smooth with minimizer $w_* \in \mathcal{H}$. Then

$$F(w) - F(w_*) \geq \frac{1}{2}\|\nabla F(w)\|_{\mathcal{H}}^2 \qquad \forall w \in \mathcal{H}$$

Proof. From a previous class (Lec 4) we know that for any $v, w \in \mathcal{H}$

$$F(v) \leq F(w) + \langle \nabla F(w), v - w\rangle_{\mathcal{H}} + \frac{L}{2}\|w - v\|_{\mathcal{H}}^2$$

By minimizing the left and right sides w.r.t. $v \in \mathcal{H}$, we have

$$\begin{aligned} F(w_*) &\leq \inf_{v \in \mathcal{H}} F(w) + \langle \nabla F(w), v - w\rangle_{\mathcal{H}} + \frac{L}{2}\|w - v\|_{\mathcal{H}}^2 \\ &= F(w) - \frac{1}{2L}\|\nabla F(w)\|_{\mathcal{H}}^2 \end{aligned}$$

Which yields the desired result. (Note that the minimizer of the quadratic upper bound is indeed given by $v = w - \frac{1}{L}\nabla F(w)$).

Co-coercivity of the Gradient

Proposition. $F : \mathcal{H} \to \mathbb{R}$ convex M-smooth. Then $\forall v, w \in \mathcal{H}$

$$\langle \nabla F(w) - \nabla F(v), w - v \rangle_{\mathcal{H}} \ \geq \ \frac{1}{M} \|\nabla F(w) - \nabla F(v)\|_{\mathcal{H}}^2$$

Proof. Define

$$F_w(z) = F(z) - \langle \nabla F(w), z \rangle_{\mathcal{H}} \qquad \text{and} \qquad F_v(z) = F(z) - \langle \nabla F(v), z \rangle_{\mathcal{H}}.$$

It is trivial to verify that F_w and F_v are M-smooth as well.

Moreover w and v are the minimizers of respectively F_w and F_v since

$$\nabla F_w(z) = \nabla F(z) - \nabla F(w) = 0 \iff z = w.$$

Therefore we can apply the previous Lemma.

Co-coercivity of the Gradient (Continued)

By applying the Lemma we have

- $\frac{1}{2M}\|\nabla F_w(v)\|_{\mathcal{H}}^2 \leq F_w(v) - F_w(w) = F(v) - F(w) - \langle \nabla F(w), v - w\rangle_{\mathcal{H}}$
- $\frac{1}{2M}\|\nabla F_v(w)\|_{\mathcal{H}}^2 \leq F_v(w) - F_v(v) = F(w) - F(v) - \langle \nabla F(v), w - v\rangle_{\mathcal{H}}$

Since $\|\nabla F_w(v)\|_{\mathcal{H}} = \|\nabla F_v(w)\|_{\mathcal{H}} = \|\nabla F(w) - \nabla F(w)\|_{\mathcal{H}}$, by summing the two inequalities we have

$$\frac{1}{M}\|\nabla F(w) - \nabla F(w)\|_{\mathcal{H}}^2 \leq \langle \nabla F(v) - \nabla F(w), v - w\rangle_{\mathcal{H}}$$

as desired.

Gradient Descent is Non-expansive

Theorem. $\ell : \mathcal{H} \to \mathbb{R}$ convex, differentiable and M-smooth. Let $0 \geq \gamma \geq 2/L$ and $G : \mathcal{H} \to \mathcal{H}$ be the gradient step operator $G(f) = f - \gamma \nabla \ell(f)$ for $f \in \mathcal{H}$. Then

$$\|G(f) - G(g)\|_{\mathcal{H}} \leq \|f - g\|_{\mathcal{H}}$$

Proof. By applying the co-coercivity of a convex M-smooth loss, we have

$$\begin{aligned}
\|G(f) - G(g)\|_{\mathcal{H}}^2 &= \|f - \gamma \nabla \ell(f) - g + \gamma \nabla \ell(g)\|_{\mathcal{H}}^2 \\
&= \|f - g\|_{\mathcal{H}}^2 - 2\gamma \langle \nabla \ell(f) - \nabla \ell(g), f - g \rangle_{\mathcal{H}} + \gamma^2 \|\nabla \ell(f) - \nabla \ell(g)\|_{\mathcal{H}}^2 \\
&\leq \|f - g\|_{\mathcal{H}}^2 - \gamma(\frac{2}{M} - \gamma)\|\nabla \ell(f) - \nabla \ell(g)\|_{\mathcal{H}}^2 \\
&\leq \|f - g\|_{\mathcal{H}}^2
\end{aligned}$$

since $\gamma(\frac{2}{M} - \gamma) \leq 1$ for $\gamma \in [0, 2/L]$. This implies the desired result.

Stability of Gradient Descent

Theorem. Let $\ell(\cdot, y) : \mathcal{H} \to \mathbb{R}$ be convex, L-Lipschitz and M-smooth uniformly for $y \in \mathcal{Y}$. For any $S \in \mathcal{Z}^n$, let f_T be the estimator produced by performing T steps of gradient descent with steps-size $\gamma = 1/M$ on the dataset S with loss ℓ starting from $0 \in \mathcal{H}$. This algorithm is $\beta(n, T)$ stable with

$$\beta(n, T) \leq \frac{2L^2k^2}{M}\frac{T}{n}$$

Proof. For any $S \in \mathcal{Z}^n$, $z \in \mathcal{Z}$ and $i = 1, \dots, n$ let us denote f_T the T-th iterate of gradient descent with step γ on S. Denote with $f_T' \in \mathcal{H}$ the T-th iterate of gradient descent with step γ on $S^{i,z}$. We want to control

$$\sup_{\bar{z} \in \mathcal{Z}} |\ell(f_T, \bar{z}) - \ell(f_T', \bar{z})| \leq Lk\|f_T - f_T'\|_{\mathcal{H}}$$

Stability of Gradient Descent (Continued)

For any $t \in \mathbb{N}$, by construction $f_{t+1} = f_t - \gamma \nabla \mathcal{E}_S(f_t)$ and $f'_{t+1} = f_t - \gamma \nabla \mathcal{E}_{S^{i,z}}(f'_t)$. Therefore

$$\begin{aligned}
\|f_{t+1} - f'_{t+1}\|_{\mathcal{H}} &= \left\| f_t - f'_t - \frac{\gamma}{n} \sum_{j \neq i} \left[\nabla \ell(f_t, z_j) - \nabla \ell(f'_t, z_j)\right] - \frac{\gamma}{n} \left[\nabla \ell(f_t, z_i) - \nabla \ell(f'_t, z)\right] \right\|_{\mathcal{H}} \\
&\leq \frac{1}{n} \sum_{j \neq i} \left\| f_t - \gamma \nabla \ell(f_t, z_j) - f'_t + \gamma \nabla \ell(f'_t, z_j) \right\|_{\mathcal{H}} \\
&\quad + \frac{1}{n} \|f_t - f'_t\|_{\mathcal{H}} + \frac{\gamma}{n} (\|\nabla \ell(f_t, z_i)\|_{\mathcal{H}} + \|\nabla \ell(f'_t, z)\|_{\mathcal{H}})
\end{aligned}$$

Recall that for $\gamma \in [0, 2/M]$, the gradient descent step $f - \gamma \nabla \ell(f, z)$ is non-expansive for any $f \in \mathcal{H}$ and $z \in \mathcal{Z}$. Therefore, for any $j \neq i$,

$$\left\| f_t - \gamma \nabla \ell(f_t, z_j) - f'_t + \gamma \nabla \ell(f'_t, z_j) \right\|_{\mathcal{H}} \leq \|f_t - f'_t\|_{\mathcal{H}}$$

Stability of Gradient Descent (Continued)

For the remaining terms, note that since ℓ is Lipschitz

$$\|\nabla \ell(f_t, z_j)\| \leq Lk \qquad \text{and} \qquad \|\nabla \ell(f_t', z)\| \leq Lk$$

(Proof.) for any $F : \mathcal{H} \to \mathbb{R}$ convex differentiable L-Lipschitz,

$$\begin{aligned}
\|\nabla F(w)\|_{\mathcal{H}} &= \sup_{\|v\|_{\mathcal{H}} \leq} \langle \nabla F(w), v \rangle_{\mathcal{H}} = \sup_{\|v\|_{\mathcal{H}} \leq 1} \lim_{t \to 0} \frac{F(w + tv) - F(w)}{t} \\
&\leq \sup_{\|v\|_{\mathcal{H}} \leq} \frac{L}{t} \|w + tv - w\|_{\mathcal{H}} = \sup_{\|v\|_{\mathcal{H}} \leq} L\|v\|_{\mathcal{H}} = L
\end{aligned}$$

Stability of Gradient Descent (Continued)

Going back to $\|f_{t+1} - f'_{t+1}\|_{\mathcal{H}}$ we have

$$\|f_{t+1} - f'_{t+1}\|_{\mathcal{H}} \leq \|f_t - f'_t\|_{\mathcal{H}} + \frac{2Lk}{n}\gamma = \frac{2Lk}{M}\frac{t+1}{n}$$

Therefore, iterating on all $t = 1, \ldots, T$, we have

$$\sup_{\bar{z}\in\mathcal{Z}} |\ell(f_T, \bar{z}) - \ell(f'_T, \bar{z})| \leq \frac{2L^2k^2}{M}\frac{T}{n}$$

as desired