Lecture 12. (Sketch.)

- Reminder: hwk1 due Wednesday, 3pm. No late homework accepted; answers discussed at start of class.

1. Smoothness recap.

- Definition: \( f \) is \( \beta \)-smooth (has \( \beta \)-Lipschitz gradients) when
  \[
  \| \nabla f(x) - \nabla f(y) \| \leq \beta \| x - y \| \quad \forall x, y.
  \]

- Key inequality/consequence:
  \[
  \left| f(x) - f(y) - \langle \nabla f(x), y - x \rangle \right| \leq \frac{\beta}{2} \| x - y \|^2 \quad \forall x, y.
  \]

- Interpretation / usefulness:
  - The key inequality tells us that at any point we can form convex and concave quadratics which respectively upper and lower bound the function.
  - Smoothness means we can take large gradient descent steps and still expect to decrease in function value.

GD Rates: with \( t \) iterations,

1. \( f \) \( \beta \)-smooth implies
   \[
   \min_{i < t} \| \nabla f(w_i) \|^2 \leq \frac{2\beta}{t} (f(w_0) - f(w_t)).
   \]

2. \( f \) \( \beta \)-smooth and convex implies
   \[
   \forall z, f(w_t) - f(z) \leq \frac{\beta}{2t} \left( \| w_t - z \|^2 - \| w_0 - z \|^2 \right).
   \]

Remark (large steps).

Consider the gradient flow (GF) iteration: \( w(0) \in \mathbb{R}^d \) is given, and \( w'(t) := \dot{w}(t) := -\nabla f(w(t)) \). (Treat these as identities; to be rigorous, we would need to argue that this differential equation has a solution.)

Using the fundamental theorem of calculus, chain rule, and definition,

\[
 f(w(t)) - f(w(0)) = \int_0^t \langle \nabla f(w(s)), \dot{w}(s) \rangle \, ds
 = -\int_0^t \| \nabla f(w(s)) \| \, ds
 \leq -t \inf_{s \in [0,t]} \| \nabla f(w(s)) \|^2,
\]

which rearranges to give

\[
 \inf_{s \in [0,t]} \| \nabla f(w(s)) \|^2 \leq \frac{1}{t} (f(w(0)) - f(w(t))).
\]
Remark (continued).

Therefore, gradient flow (small steps) avoids a factor $\beta$ which appears with gradient descent. Notice however that gradient descent uses a step size $1/\beta$, thus after $t$ steps, a distance $t/\beta$ has been covered “in gradient units”. therefore $\beta/t$ in the GD rates can be related to $1/t$ in the GF rates.

2. Strong convexity.

Here is a sort of companion to Lipschitz gradients; a stronger condition than convexity which will grant much faster convergence rates.

Say that $f$ is $\lambda$-strongly-convex ($\lambda$-sc) when

$$f(y) \geq f(x) - \langle \nabla f(x), y - x \rangle + \frac{\lambda}{2} \|y - x\|^2.$$ 

Some alternative definitions:

- $\nabla^2 f \succeq \lambda I$ ($\beta$-smooth implies $\nabla^2 f \preceq \beta I$).
- $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \lambda \|x - y\|^2$ ($\beta$-smooth gives $\leq \beta \|x - y\|^2$).
- $f$ is $\lambda$-sc iff $f - \frac{\|\cdot\|^2}{2}$ is convex.
- Definitions in terms of subgradients and function values also exist.

Example (least squares).

Last lecture, we derived

$$\frac{1}{2} \|Xw' - y\|^2 =: f(w') = f(w) + \langle \nabla f(w), w' - w \rangle + \frac{1}{2} \|Xw' - Xw\|^2$$

and

$$\sigma_{\min}(X) \|w' - w\|^2 \leq \|Xw' - Xw\|^2 \leq \sigma_{\max}(X) \|w' - w\|^2.$$ 

The latter implies smoothness, now we know the former implies strong convexity. (We can also say that both hold with equality using the special seminorm $\|v\|_X = \|Xv\|$.) We can also verify these properties by noting $\nabla^2 f = X^T X$.

Example (regularization).

Often in ML, $f$ is some risk we care about, but we train

$g(w) := f(w) + \lambda \|w\|^2/2$.

If $f$ is convex, then $g$ is $\lambda$-sc:

- A quick check is that if $f$ is twice-differentiable, then $\nabla^2 g = \nabla^2 f + \lambda I \succeq 0 + \lambda I$.
- Alternatively, it also follows by summing the inequalities

$$f(w') \geq f(w) + \langle \nabla f(w), w' - w \rangle,$$

$$\lambda \|w'\|^2/2 = \lambda \|w\|^2/2 + \langle \lambda w, w' - w \rangle + \lambda \|w' - w\|^2/2.$$
Another very useful property is that λ-sc gives a way to convert gradient norms to suboptimality.

**Lemma.** Suppose $f$ is λ-sc. Then

$$\forall w . \quad f(w) - \inf_v f(v) \leq \frac{1}{2\lambda} \|\nabla f(w)\|^2.$$  

**Proof.** Let $w$ be given, and define the convex quadratic

$$Q_w(v) := f(w) + \langle \nabla f(w), v - w \rangle + \frac{\lambda}{2} \|v - w\|^2,$$

which attains its minimum at $\bar{v} := w - \nabla f(w)/\lambda$. By definition λ-sc,

$$\inf_v f(v) \geq \inf_v Q_w(v) = Q_w(\bar{v}) = f(w) - \frac{1}{2\lambda} \|\nabla f(w)\|^2.$$  

3. Rates when strongly convex and smooth.

**Theorem.** Suppose $f$ is λ-sc and β-smooth, and GD is run with step size $1/\beta$. Then a minimum $w$ exists, and

$$f(w_t) - f(\bar{w}) \leq (f(w_0) - f(\bar{w})) \exp(-t\lambda/\beta),$$  

$$\|w_t - \bar{w}\|^2 \leq \|w_0 - \bar{w}\|^2 \exp(-t\lambda/\beta).$$

**Proof.** Using previously-proved Lemmas from smoothness and strong convexity,

$$f(w_{i+1}) - f(\bar{w}) \leq f(w_i) - f(\bar{w}) - \frac{\|\nabla f(w_i)\|^2}{2\beta}$$

$$\leq f(w_i) - f(\bar{w}) - \frac{2\lambda(f(w_i) - f(\bar{w}))}{2\beta}$$

$$\leq (f(w_i) - f(\bar{w})) (1 - \lambda/\beta),$$

which gives the first bound by induction since

$$\prod_{i \leq t} (1 - \lambda/\beta) \leq \prod_{i \leq t} \exp(-\lambda/\beta) = \exp(-t\lambda/\beta).$$

**Remark (stopping conditions).**

Say our goal is to find $w$ so that $f(w) - \inf_v f(v) \leq \epsilon$. When do we stop gradient descent? It is a pain in general and black box solvers use lots of heuristics.

- The λ-sc case is easy: by the preceding lemma, we know that we can stop when $\|\nabla f(w)\| \leq \sqrt{2\lambda}\epsilon$.
- Another easy case is when $\inf_v f(v)$ is known, and we just keep recomputing $f(w)$. This is generally the case for neural networks (where we assume $\inf_v f(v) = 0$, which often holds).
- In general though, we don’t have a nice way to do it; the usual library heuristics (checking $\|\nabla f(w)\|$ without strong convexity, checking for $f(w_t) - f(w_{t-1})$, and many other things) all stop prematurely in some cases.

The only gold standard is to use duality gaps, but these can be computationally infeasible.

**Proof (continued).**

For the second guarantee, expanding the square as usual,

$$\|w' - \bar{w}\|^2 = \|w - \bar{w}\|^2 + \frac{2}{\beta} \langle \nabla f(w), \bar{w} - w \rangle + \frac{1}{\beta^2} \|\nabla f(w)\|^2$$

$$\leq \|w - \bar{w}\|^2 + \frac{2}{\beta} \left( f(\bar{w}) - f(w) - \frac{\lambda}{2} \|\bar{w} - w\|^2 \right)$$

$$+ \frac{1}{\beta^2} \left( 2\beta(f(w) - f(w')) \right)$$

$$= (1 - \lambda/\beta) \|w - \bar{w}\|^2 + \frac{2}{\beta} \left( f(\bar{w}) - f(w) + f(w) - f(w') \right)$$

$$\leq (1 - \lambda/\beta) \|w - \bar{w}\|^2,$$

which gives the argument after a similar induction argument as before.
Remarks.

- $\beta/\lambda$ is sometimes called the \textit{condition number}, based on linear system solvers, where it is $\sigma_{\text{max}}(X)/\sigma_{\text{min}}(X)$ as in least squares. Note that $\beta \geq \lambda$ and a good condition numbers improves these bounds.

- Setting the bounds to $\epsilon$, it takes a linear number of iterations to learn a linear number of bits of $\bar{w}$.

- As will be explored in homework, much of the analysis we’ve done goes through if the norm pair $(\| \cdot \|_2, \| \cdot \|_2)$ is replaced with $(\| \cdot \|, \| \cdot \|^*)$ where the latter \textit{dual norm} is defined as

\[
\|s\|^* = \sup \left\{ \langle s, w \rangle : \|w\| \leq 1 \right\};
\]

for instance, we can define $\beta$-smooth wrt $\| \cdot \|$ as

\[
\|\nabla f(x) - \nabla f(y)\|^* \leq \beta \|x - y\|.
\]

Remark (more on gradient flow).

Assuming $f$ is $\lambda$-sc and again the gradient flow $\dot{w}(t) := -\nabla f(w(t))$, the fact $\nabla f(\bar{w}) = 0$ and inequality

\[
\frac{d}{dt} \frac{1}{2} \|w(t) - \bar{w}\|^2 = \langle w(t) - \bar{w}, \dot{w}(t) \rangle
\]

\[
= -\langle w(t) - \bar{w}, \nabla f(w(t)) - \nabla f(\bar{w}) \rangle
\]

\[
\leq -\lambda \|w(t) - \bar{w}\|^2.
\]

By Grönwall’s inequality, this implies

\[
\|w(t) - \bar{w}\|^2 \leq \|w(0) - \bar{w}\|^2 \exp(-\lambda t),
\]

which as before drops $1/\beta$, but $t/\beta$ in gradient descent in a sense has the same “units” as $t$ in gradient flow.

Remark (incomplete).

It is also interesting to replace the potential functions with

\[
\|w_t - u_t\|^2 \quad \text{and} \quad \|w(t) - u(t)\|^2,
\]

where $u_t$ and $u(t)$ are respectively gradient descent and gradient flow initialized at some $u_0 = u(0)$, possible distinct from $w_0 = w(0)$. This gives a “mixing time” style analysis (and things go through, even if we throw in coupled randomness and give a Langevin guarantee).