Smoothness and steepest descent

Administrative/Meta:

- Homework due Wednesday 10/5.
- Office hours in Siebel 3212 from now on.

Intro: steepest descent method

The following method is sometimes called **steepest descent**, although terminology is not standard. We'll analyze this method for (unconstrained!) minimization of smooth, differentiable functions.

1. Let $w_0 \in \mathbb{R}^d$ be given.
2. For $i = 1, \dots, t$:
    1. Choose any $v_i := \operatorname{argmax} \left\{ \langle \nabla f(w_{i-1}), v \rangle : \|v\| \leq 1 \right\}$.
    2. Set $w_i := w_{i-1} - \eta_i v_i$.

The choice of v_i is similar to the choice of v_i from the Frank-Wolfe method (albeit max in place of min) a 1-ball for some norm. However, the second step is different, moving to $w_{i-1} - \eta_i v_i$ rather than $w_{i-1} + \eta_i(v_i - w_i)$. In particular, this method is not attempting to maintain a convex constraint.

Examples.

- When $\|\cdot\| = \|\cdot\|_2$, then $v_i := \nabla f(w_{i-1}) / \|\nabla f(w_{i-1})\|_2$, meaning **gradient descent**.
- When $\|\cdot\| = \|\cdot\|_1$, then it's always possible to make v_i supported on a single coordinate, meaning $v_i := s_i \mathbf{e}_j$ for $j := \operatorname{argmax}_{k \in [d]} |(\nabla f(w_{i-1})_k|$ and $s_i \in \{-1, +1\}$. For this reason, this choice is called **coordinate descent**. Unfortunately this terminology is also not standard, the method has three variants:
    - The above procedure is sometimes called **greedy coordinate descent** or **forward greedy selection**, as a search is made in each iteration to choose the best coordinate.
    - **Cyclic coordinate descent** does not do the argmax, instead it has an inner loop over over all coordinates, taking an appropriate step along that coordinate. (Sometimes this is called "coordinate descent"...)
    - **Random coordinate descent** avoids the argmax and chooses a coordinate at random. This approach has been popular lately, and sometimes it can perform similarly to greedy coordinate descent, despite the reduced computation. Papers to look at here are *[note to future matus: references are nice...]* SDCA, "random kitchen sinks", kaczmarz.

Dual norms

In Fenchel duality we had a primal problem/space, and a dual problem/space; the primal was for parameters, and the dual was for gradients (roughly speaking). It's thus natural for the dual to also have norms whenever

the primal does: given a norm $\|\cdot\|$, define the **dual norm** $\|\cdot\|_*$ as

$$\|v\|_* := \sup\left\{\langle u, v\rangle : u \in \mathbb{R}^d, \|u\| \le 1\right\}.$$

This definition makes more sense with a bit of functional analysis or topology. For now, note that the definition gives a sort of generalized Hölder's for free: given any w, s,

$$\|s\|_* \ge \left\langle w/\|w\|, s\right\rangle = \langle w, s\rangle / \|w\| \qquad \Longrightarrow \qquad \langle w, s\rangle \le \|w\|\|s\|_*.$$

In homework, we *almost* proved Hölder's inequality; the proof is only completed with the "optional" problem which gives the equality case and the exact form of dual norms for conjugate exponents.

Rule of thumb / sanity check. When doing derivations, parameters and other elements of the primal space should be measured with $\|\cdot\|$, whereas gradients and other dual elements should be measured with $\|\cdot\|_*$.

Remark. From the definition of dual norms, we immediately get a propery about steepest descent. The choice of v_i is the maximizing element in the definition of the dual norm (where we always have max rather than just sup since we have only finitely many dimensions), thus

$$\left\langle \nabla f(w_{i-1}), v_i\right\rangle = \|\nabla f(w_{i-1})\|_*.$$

A key property is how norms and conjugates interact.

Proposition. If $f(w) = \|\cdot\|$, then

$$f^*(s) = \begin{cases} 0 & \text{when } \|s\|_* \le 1, \\ \infty & \text{otherwise.} \end{cases}$$

(Since f is closed and convex, then $f^{**} = f$.)

Remark. In the homework, we saw that $(\|\cdot\|_p^p/p)^* = \|\cdot\|_q^q/q$ for *conjugate exponents* $p, q \in [1, \infty]$. The preceding proposition tells us what happens if we don't mess with the powers. *[note to future matus. in class I discussed the example of lasso; can compare this to ridge regression. I should have presented OLS, lasso, and ridge clearly in an early lecture. Oh well.]*

Proof. Let $s \in \mathbb{R}^d$ be given, and consider two cases.

- If $\|s\|_* \le 1$, then by definition of dual norm

  $$f^*(s) = \sup_v v^\top s - \|v\| \le \sup_v \|v\|(\|s\|_* - 1) = 0.$$

  This case is complete by noting $f^*(s) \ge 0^\top s - \|0\| = 0$.

- Otherwise $\|s\|_* > 1$; set $\epsilon := (\|s\|_* - 1)/2$, and note by definition of dual norm that there exists u with $\|u\| = 1$ and $\langle s, u\rangle \ge 1 + \epsilon$, whereby

  $$f^*(s) = \sup\left\{\langle v, s\rangle - \|v\| : v \in \operatorname{span}(u)\right\} \ge \sup\left\{c\langle u, s\rangle - c\|u\| : c \in \mathbb{R}\right\} \ge \sup\{c\epsilon : c \in \mathbb{R}\} = \infty.$$

Smoothness

Now that we have a notion of norms and duals, we can correspondingly define smoothness with sensitivity to both norms. From there we will be able to analyze steepest descent.

A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth with respect to $\|\cdot\|$ (or simply $(\|\cdot\|, \beta)$-smooth when

$$\|\nabla f(w) - \nabla f(w')\|_* \le \beta\|w - w'\|.$$

Following the same derivation as in the last set of notes, this implies

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

(which was the original definition of smoothness we gave).

Remark. Recall the sanity check above: as expected, the definition of smoothness has dual norms on gradients, primal norms on parameters.

Examples. Let's see how smoothness arises in two common machine learning problems. Whenever faced with a prediction problem (and a loss has not yet been specified), it's good to check how how these two examples work out.

- **Least squares.** Set $f(w) := |Xw-Y|_2^2 for $X \in \mathbb{R}^{n\times d}$ and $Y \in \mathbb{R}^n$. Since $X^\top X \in \mathbb{R}^{d\times d}$ is s-psd, it has real eigenvalues $0 \leq \lambda_d \leq \cdots \leq \lambda_1$. Since $\nabla^2 f(w) := X^\top X$, we know from previous rules (which were $\|\cdot\|_2$ specific!) that f is $(\|\cdot\|_2, \lambda_1)$-smooth and $(\|\cdot\|_2, \lambda_d)$-strongly-convex, but let's check these manually. Namely,

$$\begin{aligned} f(w') - f(w) - \langle \nabla f(w), w' - w \rangle &= \|Xw' - Y\|_2^2/2 - \|Xw - Y\|_2^2/2 - \left\langle X^\top(Xw - Y), w' - w \right\rangle \\ &= \|Xw'\|_2^2/2 - \langle Xw', Y \rangle - \|Xw\|_2^2/2 + \langle Xw, Y \rangle \\ &\quad - \left\langle X^\top Xw, w' \right\rangle + \|Xw\|_2^2 + \left\langle X^\top Y, w' \right\rangle - \left\langle X^\top Y, w \right\rangle \\ &= \|Xw - Xw'\|_2^2/2. \end{aligned}$$

  From here, we get

$$\lambda_d \|w - w'\|_2 \leq \|Xw - Xw'\|_2^2 \leq \lambda_1 \|w - w'\|_2.$$

  But perhaps more importantly, consider the case $\lambda_d > 0$, meaning $X^\top X$ is s-pd. In that case, we can define a norm $\|v\|_X := \|Xv\|_2$, and we have that this f is exactly $(\|\cdot\|_X, 1)$ smooth and strongly convex, meaning with equality and no between the bounds.

  This is one reason why it's good to check prediction problems on least squares: many things can be worked out with equality, the worry of slop is gone.

  We've also seen another reason: the data covariance directly relates to the convexity properties of f, which we'll see directly relate to the convergence properties of gradient descent.

  [hey future matus: should you have talked about Bregman divergences?]

- **"Margin-based losses" for classification.** Recall the setup of data $((x_i, y_i))_{i=1}^n$ with $x_i \in \mathbb{R}^d$ and $y_i \in \{-1, +1\}$ being collected into a matrix $A \in \mathbb{R}^d$ where row i is $y_i x_i$. With this setup, we can work with single-parameter losses $\ell : \mathbb{R} \to \mathbb{R}_+$.
  - **Logistic regression** has $\ell(z) := \ln(1 + \exp(z))$. Checking derivatives, ℓ is 1/4-smooth, but not strongly convex.
  - **Least squares** at first seems to be in the wrong framework, we seem to need $(x^\top w - y)^2/2$ instead of an expression with xy. But since $y \in \{-1, +1\}$,

$$(x^\top w - y)^2/2 = y^2(x^\top w - y)^2/2 = (yx^\top w - y^2)^2/2 = (1 - yx^\top w)^2/2.$$

    Thus the least squares setup works out with $\ell(z) = (1 + z)^2$.

  Now define $\mathcal{L}(v) := \sum_{i=1}^n \ell(-v_i)/n$; empirical risk minimization thus becomes minimization of $\mathcal{L} \circ (-A)$, meaning

$$\inf_{w\in\mathbb{R}^d} \mathcal{L}(-Aw) = \inf_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell(-(Aw)_i) = \inf_{w\in\mathbb{R}^d} \frac{1}{n}\sum_{i=1}^n \ell(-y_i w^\top x_i).$$

  Homework 2 will have a problem showing how smoothness of ℓ is inherited by $\mathcal{L} \circ (-A)$. For instance, if $\max_{i,j} |A_{i,j}| \leq 1$ and ℓ is β-smooth, then $\mathcal{L} \circ (-A)$ is $(\|\cdot\|_1, \beta)$-smooth.

Smoothness and gradient descent

We'll finish the lecture by putting these pieces together to show an easy convergence result for steepest descent.

Let's see how far we can get just with smoothness. Let's use the step size $\eta_i := \|\nabla f(w_i)_{i-1}\|_*/\beta$ (which can be found by minimizing η_i in the following); by smoothness,

$$\begin{aligned} f(w_{i+1}) &\leq f(w_i) + \langle \nabla f(w_i), -\eta_{i+1} v_{i+1} \rangle + \frac{\beta}{2}\|\eta_{i+1} v_{i+1}\|^2 \\ &= f(w_i) - \frac{1}{\beta}\|\nabla f(w_i)\|_* + \frac{\beta \eta_{i+1}^2}{2} \\ &= f(w_i) - \frac{\|\nabla f(w_i)\|_*^2}{2\beta}. \end{aligned}$$

This is good news: "intuitively", since first order conditions say $\nabla f(w) = 0$ iff w is optimal, we can hope for $\|\nabla f(w)\|_*$ to be large when we are far from optimal, and the above expression says that error reduces quickly in that cases.

We will use this expression to produce convergence *rates* in the next lecture, but for now we get the following.

Proposition. If $\inf_w f(w) > -\infty$, then $\nabla f(w_i) \to 0$.

Remark. No convexity assumption is made!

Proof. Rearranging the earlier bound gives

$$\|\nabla f(w_i)\|_*^2 \leq 2\beta(f(w_i) - f(w_{i+1}))$$

Applying $\sum_{i=0}^t$ gives

$$\sum_{i=0}^t \|\nabla f(w_i)\|_*^2 \leq 2\beta \sum_{i=0}^t (f(w_i) - f(w_{i+1})) = 2\beta(f(w_0) - f(w_{t+1})) \leq 2\beta(f(w_0) - \inf_w f(w)),$$

and applying $\lim_{t\to\infty}$ to both sides gives

$$\sum_{i=0}^\infty \|\nabla f(w_i)\|_*^2 \leq 2\beta(f(w_i) - \inf_w f(w)) < \infty.$$

The result follows by properties of norms and series.

References