# Smoothness and steepest descent (part 2), intro to AdaBoost

Administrative/Meta:

- Reference solutions distributed.
- References include (Bubeck 2014) and (Boyd and Vandenberghe 2004).

## Steepest descent convergence rates

Recall our definition of steepest descent from last class, where we've specialized the step size for the case of $\beta$-smooth $f$.

1. Let $w_0$ be given.
2. for $i = 1, \ldots, t$ :
    1. Choose $v_i := \operatorname{argmax}\{\langle \nabla f(w_{i-1}), v \rangle : \|v\| \leq 1\}$.
    2. Update $w_i := w_{i-1} - \eta_i v_i$ where $\eta_i := \|\nabla f(w_{i-1})\|_*/\beta$.

Last class, we showed that smoothness (in the form of a second order Taylor upper bound) immediately gives the bound

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

and that this implies $\nabla f(w_i) \to 0$ without convexity, but rather assuming $\inf_w f(w) > -\infty$.

To start this lecture, we'll give basic rates for this method. The proofs all start from (), and then aim to lower bound $\|\nabla f(w_{i-1})\|_*$ in terms of $f(w) - f(\bar{w})$, giving a recurrence relation that gives a rate.

**Theorem.** Suppose a minimizer $\bar{w}$ exists. If $f$ is $\lambda$-strongly-convex, then

$$f(w_t) - f(\bar{w}) \leq (f(w_0) - f(\bar{w})) \exp\left(-\frac{\lambda}{\beta} t\right).$$

Otherwise, if $\|\cdot\| = \|\cdot\|_2$, then

$$f(w_t) - f(\bar{w}) \leq \frac{2\beta \|w_0 - \bar{w}\|_2^2}{t}.$$

**Remarks.**

- The second bound can be made to work for other norms; there is a step of the proof that is only obvious for $\|\cdot\|_2$, but then a relationship like $\|\cdot\|_1 \leq \sqrt{d}\|\cdot\|_2$ can be used to handle the $\ell_1$ norm. However, this does not seem to be the "right" bound. This will be discussed more when the crucial step is invoked later.
- The first bound says that $t \geq (\beta/\lambda) \ln((f(w_0) - f(\bar{w}))/\epsilon)$ iterations are needed for accuracy $\epsilon > 0$, a guarantee which has been mentioned many time throughout the course. The quantity $\beta/\lambda$ is often called the **condition number**.

Let's first worry about the strongly convex case, which is less annoying to prove.

**Lemma.** Suppose $f$ is $\lambda$-strongly-convex. For any $w$,

$$\|\nabla f(w)\|_*^2 \geq 2\lambda(f(w) - f(\bar{w})).$$

**Proof.** The proof strategy is to rewrite $f(w) - f(\bar{w})$ in a way that allows us to invoke the definition of strong convexity. Here's one way:

$$\begin{aligned} f(\bar{w}) - f(w) &= \inf_{v \in \mathbb{R}^d} f(\bar{w} + v) - f(w) \\ &\geq \inf_{v \in \mathbb{R}^d} \langle \nabla f(w), v \rangle + \frac{\lambda}{2}\|v\|^2 \\ &\geq \inf_{v \in \mathbb{R}^d} -\|\nabla f(w)\|_* \|v\| + \frac{\lambda}{2}\|v\|^2. \end{aligned}$$

This expression is a convex univariate quadratic in the scalar quantity $\alpha := \|v\|$, and is minimized at $\alpha = \|\nabla f(w)\|_*/\lambda$. Plugging this back in,

$$f(\bar{w}) - f(w) \geq -\|\nabla f(w)\|_*^2/\lambda + \lambda\|\nabla f(w)\|_*^2/(2\lambda^2) = -\|\nabla f(w)\|_*^2/(2\lambda),$$

which rearranges to give the desired expression. □

**Remarks** (on stopping conditions)**.** The methods we've discussed are run for $t$ rounds, and we bound how big $t$ needs to be, often in terms of problem-dependent quantities that are awkward to measure in practice. Instead, what is often useful in practice is to specify an accuracy $\epsilon > 0$, and have the method automatically stop once $f(w_i) - \inf_w f(w) \leq \epsilon$. Here are two ways to do that.

- By the preceding lemma, if we know $f$ is $\lambda$-strongly-convex (for instance, because it is a regularized ERM problem and has the term $\lambda\|w\|_2^2/2$), then we can stop as soon as $\|\nabla f(w_i)\|_*^2/(2\lambda) \leq \epsilon$. Unfortunately, the choice of $\lambda$ made in practice is often incredibly small. Worse, many standard implementations have this stopping rule even when strong convexity is violated. To see that this is not a stopping condition in general, it suffices to consider a function which is piecwise linear (or polyhedral) with small slope near the optimum; it's possible for $\|\nabla f(w_i)\|_*$ to be tiny while still being arbitrarily far from the optimum.
- The gold standard for stopping conditions is still to check the duality gap, for instance (considering differentiable $f$ and $g$ for simplicity)

  $$f(w) + g(Aw) + f^*(-A^\top \nabla g(Aw)) + g^*(\nabla g(Aw)),$$

  which is guaranteed to be an upper bound on the suboptimality (for instance, due to the Fenchel duality theorem). Unfortunately, it can be expensive to compute.

Returning to task, the lemma gives the desired bound fairly directly.

**Proof** of first part of theorem**.** Combining the upper bound from () with the preceding lemma,

$$\begin{aligned} f(w_{i+1}) - f(\bar{w}) &\leq f(w_i) - f(\bar{w}) - \frac{\|\nabla f(w_i)\|_*}{2\beta} \\ &\leq f(w_i) - f(\bar{w}) - \frac{\lambda(f(w_i) - f(\bar{w}))}{2\beta} \\ &\leq \big(f(w_i) - f(\bar{w})\big)\,(1 - \lambda/\beta). \end{aligned}$$

Using the inequality $1 - \lambda/\beta \leq \exp(-\lambda/\beta)$ and unrolling this recurrence $t$ times gives the bound. □.

Proving the rate with strong convexity is much more awkward. First, note how $\|\nabla f(w_i)\|_*$ may be lower bounded substantially more weakly.

**Lemma.** Given convex $f$ and any $w, w'$, $\|\nabla f(w)\|_* \geq (f(w) - f(w'))/\|w - w'\|$.

**Proof.** By the subgradient inequality and definition of dual norm,

$$f(w') - f(w) \geq \langle \nabla f(w), w' - w \rangle \geq -\|\nabla f(w)\|_* \|w' - w\|,$$

which rearranges to give the bound. □

The proof of the second part of the theorem will replace $\|w_i - \bar{w}\|$ with $\|w_0 - \bar{w}\|$ by arguing that every gradient step does not increase the distance to $\bar{w}$, established as follows.

**Lemma.** Consider the setting of the second part of the theorem. Then $\|w_{i+1} - \bar{w}\|_2 \leq \|w_i - \bar{w}\|_2$.

**Proof.** In these sorts of situations, is always useful to expand the quantity $\|w_{i+1} - \bar{w}\|_2^2$, since this can be done purely with equalities:

$$\begin{aligned}\|w_{i+1} - \bar{w}\|_2^2 &= \|w_i - \bar{w} - \nabla f(w_i)/\beta\|_2^2 \\ &= \|w_i - \bar{w}\|_2^2 - 2\langle \nabla f(w_i), w_i - \bar{w} \rangle / \beta + \|\nabla f(w_i)\|_2^2/\beta^2.\end{aligned}$$

The proof is thus complete if we can show $\|\nabla f(w_i)\|_2^2 \leq 2\beta \langle \nabla f(w_i), w_i - \bar{w} \rangle$. Setting $w' := \bar{w} + \nabla f(w_i)/\beta$,

$$\begin{aligned}0 \leq f(w_i) - f(\bar{w}) &= f(w_i) - f(w') + f(w') - f(\bar{w}) \\ &\leq \left( \langle \nabla f(w_i), w_i - w' \rangle \right) + \left( \langle \nabla f(\bar{w}), w' - \bar{w} \rangle + \frac{\beta}{2} \|\bar{w} - w'\|_2^2 \right) \\ &\leq \left( \langle \nabla f(w_i), w_i - \bar{w} \rangle - \|\nabla f(w_i)\|_2^2/\beta \right) + \left( 0 + \|\nabla f(w_i)\|_2^2/(2\beta) \right),\end{aligned}$$

which rearranges to the desired bound. □

**Remarks.**

1. Note that this property depends on the choice of step size; if something much larger were chosen, it's easy for the distance to $\bar{w}$ to increase with each step.
2. This is the step which relies upon $\|\cdot\|$. Namely, everything in the preceding bound goes through if $\eta_i$ and $v_i$ are chosen for steepest descent with general $\|\cdot\|$, but it is still shown that $\|\cdot\|_2$ shrinks. Thus, to invoke this bound, other norms must be converted into $\|\cdot\|_2$, which introduces a dimension-dependent constant. There should be another analysis that avoids this, perhaps introducing some better quantity which can be dimension-dependent in the worst case, but often much better. *[future matus: do the lemma for general steepest descent?]*

**Proof** of part 2 of the earlier theorem**.** Combining the pieces,

$$\begin{aligned}f(w_{i+1}) - f(\bar{w}) &\leq f(w_i) - f(\bar{w}) - \frac{\|\nabla f(w_i)\|_*^2}{2\beta} \\ &\leq f(w_i) - f(\bar{w}) - \frac{(f(w_i) - f(\bar{w}))^2}{2\beta \|w_i - \bar{w}\|_2^2} \\ &\leq f(w_i) - f(\bar{w}) - \frac{(f(w_i) - f(\bar{w}))^2}{2\beta \|w_0 - \bar{w}\|_2^2}\end{aligned}$$

This inequality holds for every $i$. There are many ways to solve this recurrence; the following approach is based on one due to Bubeck (2014). Setting $\epsilon_i := f(w_i) - f(\bar{w})$ and $\alpha := 1/(2\beta\|w_0 - \bar{w}\|_2^2)$, this gives recurrence relation

$$\epsilon_{i+1} \leq \epsilon_i - \alpha \epsilon_i^2,$$

which also immediately implies $\epsilon_{i+1} \leq \epsilon_i$, thus

$$\alpha \epsilon_i \epsilon_{i+1} \leq \alpha \epsilon_i^2 \leq \epsilon_i - \epsilon_{i+1}.$$

If $\epsilon_i$ or $\epsilon_{i+1}$ is ever 0, the proof is complete; otherwise, dividing both side by $\epsilon_i \epsilon_{i+1}$ gives

$$\alpha \leq \frac{1}{\epsilon_{i+1}} - \frac{1}{\epsilon_i}.$$

Applying $\sum_{i=0}^{t-1}$ to both sides gives

$$t\alpha = \sum_{i=0}^{t-1} \alpha \leq \sum_{i=0}^{t-1} \left( \frac{1}{\epsilon_{i+1}} - \frac{1}{\epsilon_i} \right) \leq \frac{1}{\epsilon_t},$$

which implies $\epsilon_t \leq 1/(t\alpha)$. $\square$

## AdaBoost

To close the discussion of convex optimization algorithms and their analysis, we'll begin the discussion of the classical method of AdaBoost (which will continue in the next lecture), which can be formulated as $\|\cdot\|_1$ steepest descent applied to a convex risk function. Specifically, AdaBoost is concerned with the objective function

$$w \mapsto \frac{1}{n} \sum_{i=1}^{n} \ell(-(Aw)_i).$$

The matrix $A$ in boosting has a special meaning, which we will elaborate in the next lecture. Briefly, one can think of each column of $A$ (each "feature" of the data points) as the output of some prediction function, and boosting is learning a linear combination of these elementary function. For instance, a popular use is to have boosting determine a linear combination of decision trees.

**Remark.** There are many variants of AdaBoost; often a careful inspection is necessary to determine whether the proposed reformulations indeed produce the same iterate sequences. The study here is of the classical AdaBoost algorithm Freund and Schapire (1997); namely, with an appropriate choice of step size, $\|\cdot\|_1$ steepest descent applied to the earlier objective function (with $\ell = \exp$) will produce the same iterate sequence as the original method.

An essential trait of () is that it is not constrained or regularized in any obvious way. Indeed, this situation is worst in the setting of the data being **separable**, which is a standard setting to analyze boosting: suppose there exists a vector $v$ such that $Av > 0$ (coordinate-wise), which means that the classifier corresponding to $v$ classifies each point correctly. Unfortunately, this setting destroys the optimization structure we've been analyzing.

**Proposition.** Consider any convex loss $\ell : \mathbb{R} \to \mathbb{R}_+$ with $\lim_{z\to-\infty} \ell(z) = 0$ and $\lim_{z\to-\infty} \ell''(-z) = 0$ (both of which are true for exponential loss $\ell = \exp$ and logistic loss $\ell = \ln(1 + \exp(\cdot))$. Suppose there exists $v \in \mathbb{R}^d$ such that $Av > 0$. Then $f(w) := \sum_{i=1}^{n} \ell(-(Aw)_i)$ does not attain a minimum, and $f$ is not strongly convex.

**Proof.** Let $v$ be given with $Av > 0$, and define $r := \min_i (Av)_i > 0$. Since $\ell \geq 0$.

$$0 \leq \inf_w f(w) \leq \lim_{c\to\infty} \sum_{i=1}^{n} \ell(-c(Av)_i) \leq \lim_{c\to\infty} \sum_{i=1}^{n} \ell(-cr) = 0,$$

meaning $\inf_w f(w) = 0$. But $f > 0$ everywhere, meaning the infimum is not attained.

For strong convexity, it suffices to prove that $\inf_w \nabla^2 f(w) = 0$, meaning the zero matrix. (Note, we care about $\|\cdot\|_1$ strong convexity, but it can be shown (perhaps it will be a homework problem) that if a finite dimensional optimization problem is not $\|\cdot\|_2$ strongly convex, then it is not $\|\cdot\|_1$ strongly convex.) To this end,

$$\inf_w \nabla^2 f(w) = \inf_w A^\top \begin{bmatrix} \ell''(-(Aw)_1) & & \\ & \ddots & \\ & & \ell''(-(Aw)_n) \end{bmatrix} A.$$

By the same argument as before, the matrix in the middle can be brought arbitrarily close to 0 by considering the ray $\{cv : c > 0\}$. $\square$

The next lecture will describe AdaBoost more precisely, and provide convergence rates in this setting which appears far from standard convex optimization.

## References

Boyd, Stephen P., and Lieven Vandenberghe. 2004. *Convex Optimization*. Cambridge University Press.

Bubeck, Sébastien. 2014. "Theory of Convex Optimization for Machine Learning."

Freund, Yoav, and Robert E. Schapire. 1997. "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting." *J. Comput. Syst. Sci.* 55 (1): 119–39.