COMS 4773: Convex optimization

Daniel Hsu

March 28, 2024

1 Smooth functions

In the context of convex optimization, *smooth functions* are functions whose derivatives (gradients) do not change too quickly. The change in the derivative is the second-derivative, so smoothness is a constraint on the second-derivatives of a function (assuming twice-differentiability).

A twice-differentiable function $J \colon \mathbb{R}^d \to \mathbb{R}$ is β-*smooth* if the eigenvalues of its Hessian matrix at any point in $\mathbb{R}^d$ are all at most β. A consequence of β-smoothness is the following. Recall that by Taylor's theorem, for any $w, \delta \in \mathbb{R}^d$, there exists $\tilde{w} \in \mathbb{R}^d$ on the line segment between w and $w + \delta$ such that

$$J(w+\delta) = J(w) + \nabla J(w)^\intercal \delta + \frac{1}{2}\delta^\intercal \nabla^2 J(\tilde{w})\delta.$$

If J is β-smooth, then we can bound the third term from above as

$$\begin{aligned} \frac{1}{2}\delta^\intercal \nabla^2 J(\tilde{w})\delta &\leq \frac{1}{2}\|\delta\|^2 \max_{u \in \mathbb{R}^d : \|u\|=1} u^\intercal \nabla^2 J(\tilde{w}) u \\ &\leq \frac{1}{2}\|\delta\|^2 \lambda_{\max}(\nabla^2 J(\tilde{w})) \leq \frac{1}{2}\|\delta\|^2 \beta. \end{aligned}$$

Therefore, if J is β-smooth, then for any $w, \delta \in \mathbb{R}^d$,

$$J(w+\delta) \leq J(w) + \nabla J(w)^\intercal \delta + \frac{\beta}{2}\|\delta\|^2. \tag{1}$$

A differentiable function $J \colon \mathbb{R}^d \to \mathbb{R}$ (that may not be twice-differentiable) is β-*smooth* if, for all $w, w' \in \mathbb{R}^d$,

$$\|\nabla J(w) - \nabla J(w')\| \leq \beta \|w - w'\|.$$

For such a differentiable function J, we have for any $w, \delta \in \mathbb{R}^d$,

$$\begin{aligned} J(w+\delta) - J(w) - \nabla J(w)^\intercal \delta &= \int_0^1 \nabla J(w + t\delta)^\intercal \delta \, \mathrm{d}t - \nabla J(w)^\intercal \delta \\ &= \int_0^1 (\nabla J(w+t\delta) - \nabla J(w))^\intercal \delta \, \mathrm{d}t \\ &\leq \int_0^1 \|\nabla J(w+t\delta) - \nabla J(w)\| \|\delta\| \, \mathrm{d}t \\ &\leq \int_0^1 \beta t \|\delta\|^2 \, \mathrm{d}t = \frac{\beta}{2}\|\delta\|^2. \end{aligned}$$

The first inequality follows by Cauchy-Schwarz, and the second inequality follows by the definition of β-smoothness. So we again have (1) for all $w, \delta \in \mathbb{R}^d$.

2 Gradient descent on smooth objectives

Gradient descent starts with an initial point $w^{(0)} \in \mathbb{R}^d$, and for a given step size η, iteratively computes a sequence of points $w^{(1)}, w^{(2)}, \ldots$ as follows. For $t = 1, 2, \ldots$:

$$w^{(t)} = w^{(t-1)} - \eta \nabla J(w^{(t-1)}),$$

where $\nabla J \colon \mathbb{R}^d \to \mathbb{R}^d$ is the gradient map for the objective function $J \colon \mathbb{R}^d \to \mathbb{R}$ to be minimized.

2.1 Motivation

The motivation for the gradient descent update is the following. Suppose we have a current point $w \in \mathbb{R}^d$, and we would like to locally change it from w to $w + \delta$ so as to decrease the objective value. How should we choose δ?

In gradient descent, we consider the quadratic upper-bound from (1) granted by smoothness:

$$J(w + \delta) \leq J(w) + \nabla J(w)^\intercal \delta + \frac{\beta}{2} \|\delta\|^2,$$

and then choose δ to minimize this upper-bound. The upper-bound is a convex quadratic function of δ, so its minimizer can be written in closed-form. The minimizer is the value of δ such that

$$\nabla J(w) + \beta \delta = 0.$$

In other words, it is $\delta^\star(w)$, defined by

$$\delta^\star(w) = -\frac{1}{\beta} \nabla J(w).$$

Plugging in $\delta^\star(w)$ for δ in the quadratic upper-bound gives

$$\begin{aligned} J(w + \delta^\star(w)) &\leq J(w) + \nabla J(w)^\intercal \delta^\star(w) + \frac{\beta}{2} \|\delta^\star(w)\|^2 \\ &= J(w) - \frac{1}{\beta} \nabla J(w)^\intercal \nabla J(w) + \frac{1}{2\beta} \|\nabla J(w)\|^2 \\ &= J(w) - \frac{1}{2\beta} \|\nabla J(w)\|^2. \end{aligned}$$

This inequality tells us that this local change to w will decrease the objective value as long as the gradient at w is non-zero. It turns out that if the function J is convex (in addition to β-smooth), then repeatedly making such local changes is sufficient to approximately minimize the function.

2.2 Analysis for smooth convex objectives

One of the simplest ways to mathematically analyze the behavior of gradient descent on smooth functions (with step size $\eta = 1/\beta$) is to monitor the change in a "potential function" during the execution of gradient descent. The potential function we will use is the squared Euclidean distance to a fixed vector $w^\star \in \mathbb{R}^d$, which could be a minimizer of J (but need not be):

$$\Phi(w) = \frac{1}{2\eta} \|w - w^\star\|^2.$$

The scaling by $\frac{1}{2\eta}$ is used just for notational convenience.

Let us examine the "drop" in the potential when we change a point w to $w + \delta^\star(w)$ (as in gradient descent):

$$\begin{aligned}
\Phi(w) - \Phi(w + \delta^\star(w)) &= \frac{1}{2\eta}\|w - w^\star\|^2 - \frac{1}{2\eta}\|w + \delta^\star(w) - w^\star\|^2 \\
&= \frac{\beta}{2}\|w - w^\star\|^2 - \frac{\beta}{2}\left(\|w - w^\star\|^2 + 2\delta^\star(w)^\mathsf{T}(w - w^\star) + \|\delta^\star(w)\|^2\right) \\
&= -\beta\delta^\star(w)^\mathsf{T}(w^\star - w) - \frac{\beta}{2}\|\delta^\star(w)\|^2 \\
&= \nabla J(w)^\mathsf{T}(w - w^\star) - \frac{1}{2\beta}\|\nabla J(w)\|^2.
\end{aligned}$$

In the last step, we have plugged in $\delta^\star(w) = -\frac{1}{\beta}\nabla J(w)$. Now we use two key facts. The first is the inequality we derived above based on the smoothness of J:

$$J(w + \delta^\star(w)) \leq J(w) - \frac{1}{2\beta}\|\nabla J(w)\|^2,$$

which rearranges to

$$-\frac{1}{2\beta}\|\nabla J(w)\|^2 \geq J(w + \delta^\star(w)) - J(w).$$

The second comes from the first-order definition of convexity:

$$J(w^\star) \geq J(w) + \nabla J(w)^\mathsf{T}(w^\star - w),$$

which rearranges to

$$\nabla J(w)^\mathsf{T}(w - w^\star) \geq J(w) - J(w^\star).$$

So, we can bound the drop in potential as follows:

$$\begin{aligned}
\Phi(w) - \Phi(w + \delta^\star(w)) &= \nabla J(w)^\mathsf{T}(w - w^\star) - \frac{1}{2\beta}\|\nabla J(w)\|^2 \\
&\geq (J(w) - J(w^\star)) + (J(w + \delta^\star(w)) - J(w)) \\
&= J(w + \delta^\star(w)) - J(w^\star).
\end{aligned}$$

Let us write this inequality in terms of the iterates of gradient descent with $\eta = 1/\beta$:

$$\Phi(w^{(t-1)}) - \Phi(w^{(t)}) \geq J(w^{(t)}) - J(w^\star).$$

Summing this inequality from $t = 1, 2, \ldots, T$:

$$\sum_{t=1}^{T}\Big(\Phi(w^{(t-1)}) - \Phi(w^{(t)})\Big) \geq \sum_{t=1}^{T}\Big(J(w^{(t)}) - J(w^\star)\Big).$$

The left-hand side simplifies to $\Phi(w^{(0)}) - \Phi(w^{(T)})$. Furthermore, since $J(w^{(t)}) \geq J(w^{(T)})$ for all $t = 1, \ldots, T$, the right-hand side can be bounded from below by

$$T\Big(J(w^{(T)}) - J(w^\star)\Big).$$

So we are left with the inequality

$$J(w^{(T)}) - J(w^\star) \leq \frac{1}{T}\Big(\Phi(w^{(0)}) - \Phi(w^{(T)})\Big) = \frac{\beta}{2T}\Big(\|w^{(0)} - w^\star\|^2 - \|w^{(T)} - w^\star\|^2\Big).$$

3 Gradient descent on non-smooth objectives

Gradient descent can also be used for non-smooth convex functions as long as the function itself does not change too quickly.

We say that a differentiable function $J\colon \mathbb{R}^d \to \mathbb{R}$ is *L-Lipschitz* if its gradient at any point in $\mathbb{R}^d$ is bounded in Euclidean norm by L.

The motivation for gradient descent based on minimizing quadratic upper-bounds no longer applies. Indeed, the gradient at w could be very different from the gradient at a nearby w', so the function value at $w - \eta\nabla J(w)$ could be worse than the function value at w. Therefore, we cannot expect to have the same convergence guarantee for non-smooth functions that we had for smooth functions.

Gradient descent, nevertheless, will produce a sequence $w^{(1)}, w^{(2)}, \ldots$ such that the function value at these points is approximately minimal "on average".

3.1 Motivation

A basic motivation for gradient descent for convex functions, that does not assume smoothness, comes from the first-order condition for convexity:

$$J(w^\star) \geq J(w) + \nabla J(w)^\intercal (w^\star - w),$$

which rearranges to

$$(-\nabla J(w))^\intercal (w^\star - w) \geq J(w) - J(w^\star).$$

Suppose $J(w) > J(w^\star)$, so that moving from w to $w^\star$ would improve the function value. Then, the inequality implies that the negative gradient $-\nabla J(w)$ at w makes a positive inner product with the direction from w to $w^\star$. This is the crucial property that makes gradient descent work.

3.2 Analysis

We again monitor the change in the potential function

$$\Phi(w) = \frac{1}{2\eta}\|w - w^\star\|^2,$$

for a fixed vector $w^\star \in \mathbb{R}^d$.

Again, let us examine the "drop" in the potential when we change a point w to $w - \eta\nabla J(w)$ (as in gradient descent):

$$\begin{aligned}
\Phi(w) - \Phi(w - \eta\nabla J(w)) &= \frac{1}{2\eta}\|w - w^\star\|^2 - \frac{1}{2\eta}\|w - \eta\nabla J(w) - w^\star\|^2 \\
&= (-\nabla J(w))^\intercal (w - w^\star) - \frac{\eta}{2}\|\nabla J(w)\|^2 \\
&\geq J(w) - J(w^\star) - \frac{L^2\eta}{2},
\end{aligned}$$

where the inequality uses the convexity and Lipschitzness of J. In terms of the iterates of gradient descent, this reads

$$\Phi(w^{(t-1)}) - \Phi(w^{(t)}) \geq J(w^{(t-1)}) - J(w^\star) - \frac{L^2\eta}{2}.$$

Summing this inequality from $t = 1, 2, \ldots, T$:

$$\Phi(w^{(0)}) - \Phi(w^{(T)}) \geq \sum_{t=1}^{T} \Big(J(w^{(t-1)}) - J(w^\star) \Big) - \frac{L^2 \eta T}{2}.$$

Rearranging and dividing through by T (and dropping a term):

$$\frac{1}{T} \sum_{t=1}^{T} \Big(J(w^{(t-1)}) - J(w^\star) \Big) \leq \frac{\|w^{(0)} - w^\star\|^2}{2\eta T} + \frac{L^2 \eta}{2}.$$

The left-hand side is the average sub-optimality relative to $J(w^\star)$. Therefore, there exists some $t^* \in \{0, 1, \ldots, T-1\}$ such that

$$J(w^{(t^*)}) - J(w^\star) \leq \frac{1}{T} \sum_{t=1}^{T} \Big(J(w^{(t-1)}) - J(w^\star) \Big) \leq \frac{\|w^{(0)} - w^\star\|^2}{2\eta T} + \frac{L^2 \eta}{2}.$$

The right-hand side is $O(1/\sqrt{T})$ when we choose $\eta = 1/\sqrt{T}$. Alternatively, we can take

$$\bar{w} = \frac{1}{T} \sum_{t=0}^{T-1} w^{(t)},$$

so that by convexity, we have

$$J(\bar{w}) \leq \frac{1}{T} \sum_{t=0}^{T-1} J(w^{(t)}) \leq J(w^\star) + \frac{\|w^{(0)} - w^\star\|^2}{2\eta T} + \frac{L^2 \eta}{2}.$$

4 Constrained optimization

In a constrained convex optimization problem, one seeks to minimize a convex objective function $J \colon \mathbb{R}^d \to \mathbb{R}$ over a convex set called the feasible region.

4.1 Projected gradient descent

The projected gradient descent algorithm is a variant of gradient descent for such problems. It requires a subroutine—called a "projection oracle"—for computing orthogonal projections to the convex feasible region $K \subseteq \mathbb{R}^d$. The projection oracle Π_K should, on input $w \in \mathbb{R}^d$, return the (unique) point in K closest to w in Euclidean distance:

$$\Pi_K(w) = \operatorname*{arg\,min}_{u \in K} \|u - w\|^2.$$

For example, if $K = \{w \in \mathbb{R}^d : \|w\| \leq 1\}$ is the the unit ball in $\mathbb{R}^d$, then the projection oracle is as follows:

$$\Pi_K(w) = \begin{cases} w & \text{if } \|w\| \leq 1, \\ \frac{w}{\|w\|} & \text{otherwise.} \end{cases}$$

The output of the projection oracle Π_K is required to satisfy, for all $w \in \mathbb{R}^d$ and $u \in K$,

$$(w - \Pi_K(w))^\intercal (u - \Pi_K(w)) \leq 0.$$

This is equivalent to the following:

$$\begin{aligned}
\|u - w\|^2 &= \|u - \Pi_K(w)\|^2 + 2(u - \Pi_K(w))^\intercal(\Pi_K(w) - w) + \|\Pi_K(w) - w\|^2 \\
&\geq \|u - \Pi_K(w)\|^2 + \|\Pi_K(w) - w\|^2,
\end{aligned}$$

which can be viewed as a generalization of the Pythagorean theorem. Indeed, if K is an affine subspace, then the inequality above holds with equality.

The update rule for projected gradient descent is as follows. For $t = 1, 2, \ldots$:

$$w^{(t)} = \Pi_K(w^{(t-1)} - \eta \nabla J(w^{(t-1)})).$$

The potential-based analysis of gradient descent (both for smooth and non-smooth objectives J) extends to projected gradient descent. The only modifications to the argument needed are: (i) to restrict $w^\star$ to be in K, and (ii) to lower-bound the change in potential Φ (which implicitly depends on $w^\star \in K$)

$$\Phi(w) - \Phi(\Pi_K(w - \eta \nabla J(w)))$$

by the change in potential without the projection step:

$$\Phi(w) - \Phi(w - \eta \nabla J(w)).$$

Such a lower-bound is a direct consequence of the generalized Pythagorean theorem: for any $w \in \mathbb{R}^d$,

$$\begin{aligned}
\Phi(\Pi_K(w - \eta \nabla J(w))) &= \frac{1}{2\eta} \|\Pi_K(w - \eta \nabla J(w)) - w^\star\|^2 \\
&\leq \frac{1}{2\eta} \left(\|w - \eta \nabla J(w) - w^\star\|^2 - \|\Pi_K(w - \eta \nabla J(w)) - (w - \eta \nabla J(w))\|^2 \right) \\
&\leq \frac{1}{2\eta} \|w - \eta \nabla J(w) - w^\star\|^2 \\
&= \Phi(w - \eta \nabla J(w)),
\end{aligned}$$

and therefore

$$\Phi(w) - \Phi(\Pi_K(w - \eta \nabla J(w))) \geq \Phi(w) - \Phi(w - \eta \nabla J(w)).$$

4.2 Convex feasibility problems

In some convex optimization problems, it may not be obvious how to implement a projection oracle for the feasible region. Let us consider a convex feasibility problem, defined by a (simple) convex set $S \subseteq \mathbb{R}^d$, as well as n convex functions $f_1, \ldots, f_n \colon \mathbb{R}^d \to \mathbb{R}$, where the goal is to find $w \in S$ satisfying

$$f_i(w) \leq 0 \quad \text{for all } i = 1, \ldots, n,$$

or determine if no such w exists. The set S is regarded as a constraint that is "easy" to enforce, while the f_i's are regarded as constraints that are more "difficult" to enforce. The overall feasible region is $K = S \cap \{w \in \mathbb{R}^d : f_i(w) \leq 0 \text{ for all } i = 1, \ldots, n\}$.

Our goal is to approximately solve the feasibility problem, where we allow some slack in the "difficult" constraints. Specifically, for a given $\epsilon > 0$, we either find a proof that the problem is infeasible, or we return $\hat{w} \in S$ satisfying

$$f_i(\hat{w}) \leq \epsilon \quad \text{for all } i = 1, \ldots, n.$$

For ϵ to be a meaningful parameter, we assume that $|f_i(w)| \leq 1$ for all $w \in S$ and $i \in [n]$.

One approach to solving this problem is to formulate the objective function

$$J(w) = \max_{i \in [n]} f_i(w),$$

and then to attempt to optimize J over S. If we have a projection oracle for S, then we can use projected gradient descent to solve the problem. This is because the maximum of convex functions is also convex.[1]

We consider a second approach to solving the problem that is related to a more general optimization scheme. Define $L\colon S \times \Delta^{n-1} \to \mathbb{R}$ by

$$L(w, p) = \sum_{i=1}^{n} p_i f_i(w).$$

Observe that, for any $w \in S$, we have

$$J(w) = \max_{p \in \Delta^{n-1}} L(w, p).$$

This second approach to solving the the problem requires an "optimization oracle" for approximately minimizing $L(w, p)$ over $w \in S$. Specifically, given $p \in \Delta^{n-1}$, the oracle should return $\hat{w} \in S$ satisfying

$$L(\hat{w}, p) \leq \min_{w \in S} L(w, p) + \epsilon/2.$$

The algorithm is based on the HEDGE algorithm for the online allocation problem.

- Let $p_1 \in \Delta^{n-1}$ be the initial allocation vector used by HEDGE (which, by default, is the uniform distribution).
- For $t = 1, 2, \ldots, T$:
  - Invoke the optimization oracle to obtain $w_t \in S$ satisfying

    $$L(w_t, p_t) \leq \min_{w \in S} L(w, p_t) + \epsilon/2.$$
  - If $L(w_t, p_t) > \epsilon/2$, then abort and return "infeasible".
  - Otherwise, provide loss vector $\ell_t = -(f_1(w_t), \ldots, f_n(w_t)) \in [-1, 1]^n$ to HEDGE to obtain updated allocation vector $p_{t+1} \in \Delta^{n-1}$.
- Return $\hat{w} = \frac{1}{T} \sum_{t=1}^{T} w_t$.

[1] The function $\max_{i \in [n]} f_i(w)$ is convex but not differentiable. Nevertheless, (projected) gradient descent works just as well with subgradients, which may be easy to obtain. In this case, a subgradient of $\max_{i \in [n]} f_i(w)$ is the gradient of any f_i at w for which $f_i(w)$ attains the max.

If the algorithm aborts in iteration t and returns "infeasible", then we have found $p_t \in \Delta^{n-1}$ such that

$$\epsilon/2 < L(w_t, p_t) \leq \min_{w \in S} L(w, p_t) + \epsilon/2,$$

which implies that

$$\min_{w \in S} L(w, p_t) > 0.$$

This proves that the problem is infeasible.

Now suppose instead the algorithm does not abort prematurely, so $L(w_t, p_t) \leq \epsilon/2$ for all $t = 1, 2, \ldots, T$. Notice that

$$\langle e_i, \ell_t \rangle = -f_i(w_t) \quad \text{for all } i \in [n],$$

$$\text{and} \qquad \langle p_t, \ell_t \rangle = -\sum_{i=1}^{n} p_{t,i} f_i(w_t) = -L(w_t, p_t).$$

Therefore, the guarantee from HEDGE (with a suitable choice of hyperparameter $\eta > 0$) is

$$\sum_{t=1}^{T} -L(w_t, p_t) \leq \min_{i \in [n]} \sum_{t=1}^{T} -f_i(w_t) + O\left(\sqrt{T \log n}\right)$$

Dividing both sides by T, re-arranging, and using $T = O((\log n)/\epsilon^2)$, we have

$$\frac{1}{T} \sum_{t=1}^{T} f_i(w_t) \leq \frac{1}{T} \sum_{t=1}^{T} L(w_t, p_t) + \epsilon/2 \quad \text{for all } i \in [n].$$

By Jensen's inequality and the assumption that the algorithm does not abort prematurely, we have

$$f_i(\hat{w}) \leq \frac{1}{T} \sum_{t=1}^{T} f_i(w_t) \leq \frac{1}{T} \sum_{t=1}^{T} L(w_t, p_t) + \epsilon/2 \leq \epsilon$$

for all $i \in [n]$.