Gradient descent

Daniel Hsu (COMS 4771)

Smooth functions

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.

For any $\beta > 0$, we say a twice-differentiable function $f \colon \mathbb{R}^d \to \mathbb{R}$ is *β-smooth* if the eigenvalues of its Hessian matrix at any point in $\mathbb{R}^d$ are at most β.

Example: logistic regression

Consider the empirical logistic loss risk on a training data set $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n) \in \mathbb{R}^d \times \{-1, +1\}$:

$$\widehat{\mathcal{R}}(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^n \ln(1 + \exp(-y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w})).$$

The Hessian of $\widehat{\mathcal{R}}$ at $\boldsymbol{w}$ is

$$\nabla^2 \widehat{\mathcal{R}}(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^n \sigma(y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}) \sigma(-y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}) \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T}$$

where $\sigma(t) = 1/(1 + \exp(-t))$ is the sigmoid function. For any unit vector $\boldsymbol{u} \in \mathbb{R}^d$, we have

$$\begin{aligned}
\boldsymbol{u}^\mathsf{T} \nabla^2 \widehat{\mathcal{R}}(\boldsymbol{w}) \boldsymbol{u} &= \frac{1}{n} \sum_{i=1}^n \sigma(y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}) \sigma(-y_i \boldsymbol{x}_i^\mathsf{T} \boldsymbol{w}) (\boldsymbol{x}_i^\mathsf{T} \boldsymbol{u})^2 \\
&\leq \frac{1}{4n} \sum_{i=1}^n (\boldsymbol{x}_i^\mathsf{T} \boldsymbol{u})^2 \\
&= \frac{1}{4} \boldsymbol{u}^\mathsf{T} \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \right) \boldsymbol{u} \\
&\leq \frac{1}{4} \lambda_{\max} \left(\frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T} \right),
\end{aligned}$$

where $\lambda_{\max}(\boldsymbol{M})$ is used to denote the largest eigenvalue of a symmetric matrix $\boldsymbol{M}$. So if λ_1 is the largest eigenvalue of the empirical second moment matrix $\frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i \boldsymbol{x}_i^\mathsf{T}$, then $\widehat{\mathcal{R}}$ is β-smooth for $\beta = \lambda_1/4$.

Quadratic upper bound for smooth functions

A consequence of β-smoothness is the following. Recall that by Taylor's theorem, for any $\boldsymbol{w}, \boldsymbol{\delta} \in \mathbb{R}^d$, there exists $\tilde{\boldsymbol{w}} \in \mathbb{R}^d$ on the line segment between $\boldsymbol{w}$ and $\boldsymbol{w} + \boldsymbol{\delta}$ such that

$$f(\boldsymbol{w} + \boldsymbol{\delta}) = f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\mathsf{T} \boldsymbol{\delta} + \frac{1}{2} \boldsymbol{\delta}^\mathsf{T} \nabla^2 f(\tilde{\boldsymbol{w}}) \boldsymbol{\delta}.$$

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

$$\begin{aligned}\frac{1}{2}\boldsymbol{\delta}^\mathsf{T}\nabla^2 f(\tilde{\boldsymbol{w}})\boldsymbol{\delta} &\leq \frac{1}{2}\|\boldsymbol{\delta}\|_2^2 \max_{\boldsymbol{u}\in\mathbb{R}^d:\|\boldsymbol{u}\|_2=1} \boldsymbol{u}^\mathsf{T}\nabla^2 f(\tilde{\boldsymbol{w}})\boldsymbol{u} \\ &\leq \frac{1}{2}\|\boldsymbol{\delta}\|_2^2\lambda_{\max}(\nabla^2 f(\tilde{\boldsymbol{w}})) \\ &\leq \frac{1}{2}\|\boldsymbol{\delta}\|_2^2\beta.\end{aligned}$$

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

$$f(\boldsymbol{w}+\boldsymbol{\delta}) \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\mathsf{T}\boldsymbol{\delta} + \frac{\beta}{2}\|\boldsymbol{\delta}\|_2^2.$$

Gradient descent on smooth functions

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

$$\boldsymbol{w}^{(t)} := \boldsymbol{w}^{(t-1)} - \eta\nabla f(\boldsymbol{w}^{(t-1)}).$$

Motivation for gradient descent on smooth functions

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

In gradient descent, we consider the quadratic upper-bound that smoothness grants, i.e.,

$$f(\boldsymbol{w}+\boldsymbol{\delta}) \leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\mathsf{T}\boldsymbol{\delta} + \frac{\beta}{2}\|\boldsymbol{\delta}\|_2^2,$$

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

$$\nabla f(\boldsymbol{w}) + \beta\boldsymbol{\delta} = \boldsymbol{0}.$$

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

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

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

$$\begin{aligned}f(\boldsymbol{w}+\boldsymbol{\delta}^\star(\boldsymbol{w})) &\leq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\mathsf{T}\boldsymbol{\delta}^\star(\boldsymbol{w}) + \frac{\beta}{2}\|\boldsymbol{\delta}^\star(\boldsymbol{w})\|_2^2 \\ &= f(\boldsymbol{w}) - \frac{1}{\beta}\nabla f(\boldsymbol{w})^\mathsf{T}\nabla f(\boldsymbol{w}) + \frac{1}{2\beta}\|\nabla f(\boldsymbol{w})\|_2^2 \\ &= f(\boldsymbol{w}) - \frac{1}{2\beta}\|\nabla f(\boldsymbol{w})\|_2^2.\end{aligned}$$

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

Analysis of gradient descent on smooth convex functions

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 $\boldsymbol{w}^\star \in \mathbb{R}^d$, which could be a minimizer of f (but need not be):

$$\Phi(\boldsymbol{w}) := \frac{1}{2\eta}\|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2.$$

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 $\boldsymbol{w}$ to $\boldsymbol{w} + \boldsymbol{\delta}^\star(\boldsymbol{w})$ (as in gradient descent):

$$\begin{aligned}
\Phi(\boldsymbol{w}) - \Phi(\boldsymbol{w} + \boldsymbol{\delta}^\star(\boldsymbol{w})) &= \frac{1}{2\eta}\|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 - \frac{1}{2\eta}\|\boldsymbol{w} + \boldsymbol{\delta}^\star(\boldsymbol{w}) - \boldsymbol{w}^\star\|_2^2 \\
&= \frac{\beta}{2}\|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 - \frac{\beta}{2}\left(\|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 + 2\boldsymbol{\delta}^\star(\boldsymbol{w})^\intercal(\boldsymbol{w} - \boldsymbol{w}^\star) + \|\boldsymbol{\delta}^\star(\boldsymbol{w})\|_2^2\right) \\
&= -\beta\boldsymbol{\delta}^\star(\boldsymbol{w})^\intercal(\boldsymbol{w}^\star - \boldsymbol{w}) - \frac{\beta}{2}\|\boldsymbol{\delta}^\star(\boldsymbol{w})\|_2^2 \\
&= \nabla f(\boldsymbol{w})^\intercal(\boldsymbol{w} - \boldsymbol{w}^\star) - \frac{1}{2\beta}\|\nabla f(\boldsymbol{w})\|_2^2.
\end{aligned}$$

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

$$f(\boldsymbol{w} + \boldsymbol{\delta}^\star(\boldsymbol{w})) \leq f(\boldsymbol{w}) - \frac{1}{2\beta}\|\nabla f(\boldsymbol{w})\|_2^2,$$

which rearranges to

$$-\frac{1}{2\beta}\|\nabla f(\boldsymbol{w})\|_2^2 \geq f(\boldsymbol{w} + \boldsymbol{\delta}^\star(\boldsymbol{w})) - f(\boldsymbol{w}).$$

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

$$f(\boldsymbol{w}^\star) \geq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\intercal(\boldsymbol{w}^\star - \boldsymbol{w}),$$

which rearranges to

$$\nabla f(\boldsymbol{w})^\intercal(\boldsymbol{w} - \boldsymbol{w}^\star) \geq f(\boldsymbol{w}) - f(\boldsymbol{w}^\star).$$

(We'll discuss this inequality more later.) So, we can bound the drop in potential as follows:

$$\begin{aligned}
\Phi(\boldsymbol{w}) - \Phi(\boldsymbol{w} + \boldsymbol{\delta}^\star(\boldsymbol{w})) &= \nabla f(\boldsymbol{w})^\intercal(\boldsymbol{w} - \boldsymbol{w}^\star) - \frac{1}{2\beta}\|\nabla f(\boldsymbol{w})\|_2^2 \\
&\geq \big(f(\boldsymbol{w}) - f(\boldsymbol{w}^\star)\big) + \big(f(\boldsymbol{w} + \boldsymbol{\delta}^\star(\boldsymbol{w})) - f(\boldsymbol{w})\big) \\
&= f(\boldsymbol{w} + \boldsymbol{\delta}^\star(\boldsymbol{w})) - f(\boldsymbol{w}^\star).
\end{aligned}$$

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

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

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

$$\sum_{t=1}^{T}\left(\Phi(\boldsymbol{w}^{(t-1)}) - \Phi(\boldsymbol{w}^{(t)})\right) \geq \sum_{t=1}^{T}\left(f(\boldsymbol{w}^{(t)}) - f(\boldsymbol{w}^\star)\right).$$

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

$$T\left(f(\boldsymbol{w}^{(T)}) - f(\boldsymbol{w}^\star)\right).$$

So we are left with the inequality

$$f(\boldsymbol{w}^{(T)}) - f(\boldsymbol{w}^\star) \leq \frac{1}{T}\left(\Phi(\boldsymbol{w}^{(0)}) - \Phi(\boldsymbol{w}^{(T)})\right) = \frac{\beta}{2T}\left(\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^\star\|_2^2 - \|\boldsymbol{w}^{(T)} - \boldsymbol{w}^\star\|_2^2\right).$$

Gradient descent on Lipschitz convex functions

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

For any $L > 0$, we say that a differentiable function $f\colon \mathbb{R}^d \to \mathbb{R}$ 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 $\boldsymbol{w}$ could be very different from the gradient at a nearby $\boldsymbol{w}'$, so the function value at $\boldsymbol{w} - \eta\nabla f(\boldsymbol{w})$ could be worse than the function value at $\boldsymbol{w}$. 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 $\boldsymbol{w}^{(1)}, \boldsymbol{w}^{(2)}, \ldots$ such that the function value at these points is approximately minimal *on average*.

Motivation for gradient descent on Lipschitz convex functions

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

$$f(\boldsymbol{w}^\star) \geq f(\boldsymbol{w}) + \nabla f(\boldsymbol{w})^\intercal(\boldsymbol{w}^\star - \boldsymbol{w}),$$

which rearranges to

$$(-\nabla f(\boldsymbol{w}))^\intercal(\boldsymbol{w}^\star - \boldsymbol{w}) \geq f(\boldsymbol{w}) - f(\boldsymbol{w}^\star).$$

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

Analysis of gradient descent on Lipschitz convex functions

We again monitor the change in the potential function

$$\Phi(\boldsymbol{w}) := \frac{1}{2\eta}\|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2,$$

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

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

$$\begin{aligned}
\Phi(\boldsymbol{w}) - \Phi(\boldsymbol{w} - \eta\nabla f(\boldsymbol{w})) &= \frac{1}{2\eta}\|\boldsymbol{w} - \boldsymbol{w}^\star\|_2^2 - \frac{1}{2\eta}\|\boldsymbol{w} - \eta\nabla f(\boldsymbol{w}) - \boldsymbol{w}^\star\|_2^2 \\
&= (-\nabla f(\boldsymbol{w}))^\intercal(\boldsymbol{w} - \boldsymbol{w}^\star) - \frac{\eta}{2}\|\nabla f(\boldsymbol{w})\|_2^2 \\
&\geq f(\boldsymbol{w}) - f(\boldsymbol{w}^\star) - \frac{L^2\eta}{2},
\end{aligned}$$

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

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

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

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

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

$$\frac{1}{T}\sum_{t=1}^{T}\Big(f(\boldsymbol{w}^{(t-1)}) - f(\boldsymbol{w}^{\star})\Big) \leq \frac{\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^{\star}\|_2^2}{2\eta T} + \frac{L^2\eta}{2}.$$

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

$$f(\boldsymbol{w}^{(t^*)}) - f(\boldsymbol{w}^{\star}) \leq \frac{1}{T}\sum_{t=1}^{T}\Big(f(\boldsymbol{w}^{(t-1)}) - f(\boldsymbol{w}^{\star})\Big) \leq \frac{\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^{\star}\|_2^2}{2\eta T} + \frac{L^2\eta}{2}.$$

The right-hand side is $O(1/\sqrt{T})$ when we choose $\eta = 1/\sqrt{T}$.[1] Alternatively, if we compute the average point

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

then by Jensen's inequality we have

$$f(\bar{\boldsymbol{w}}_T) = f\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{w}^{(t-1)}\right) \leq \frac{1}{T}\sum_{t=1}^{T} f(\boldsymbol{w}^{(t-1)}).$$

So the bound for $\boldsymbol{w}^{(t^*)}$ also applies to $\bar{\boldsymbol{w}}_T$:

$$f(\bar{\boldsymbol{w}}_T) - f(\boldsymbol{w}^{\star}) \leq \frac{1}{T}\sum_{t=1}^{T}\Big(f(\boldsymbol{w}^{(t-1)}) - f(\boldsymbol{w}^{\star})\Big) \leq \frac{\|\boldsymbol{w}^{(0)} - \boldsymbol{w}^{\star}\|_2^2}{2\eta T} + \frac{L^2\eta}{2}.$$

[1] A similar guarantee holds when the step size used for the t-th update is $\eta_t = 1/\sqrt{t}$.