COS 597C: New Directions in Theoretical Machine Learning **Fall 2017**

Lecture 2: 19 September 2017

Lecturer: Sanjeev Arora *Scribe: Mikhail Khodak*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

2.1 Gradient Descent

Our goal in this lecture is to explore what can proved about (non-stochastic) *gradient descent* (GD) when faced with optimizing a nonconvex function $f : \mathbb{R}^d \mapsto \mathbb{R}$. It is well-known that if f is strongly convex then the standard GD iteration $x_{k+1} = x_k - \eta \nabla f(x_k)$ converges quickly to a global optimum for appropriate choice of η. However, what can we say about the nonconvex case? Can we even enforce that the iteration will decrease? Will it at least arrive to a local minimum?

The answer to the first question turns out to be yes (Lemma 1). Proving the second turns out to require the introduction of random jumps into the gradient algorithm that allow it to escape from saddle points (Theorem 1). In this lecture we'll show both of these facts under common regularity assumptions on f, which together show that *perturbed gradient descent* (PGD) reaches an approximate *second-order local minimum* in polynomial time. Although these results do not guarantee global optimality for all problems, there are multiple nonconvex settings where any second-order local minimum is enough for global optimality, such as matrix completion or phase retrieval.

Results based on the work of Jin-Ge-Lee-Jordan (2017).

2.2 Reaching a Stationary Point

We consider functions f that satisfy the following two properties:

1. ρ-Hessian Lipschitzness: $\|\nabla^2 f(x) - \nabla^2 f(x')\| \leq \rho \|x - x'\|$
2. ℓ-smoothness: $\|\nabla f(x) - \nabla f(x')\| \leq \ell \|x - x'\|$

Let's first see that the function value decreases using GD with appropriate step-size:

Lemma 1. *If $\eta \leq \frac{1}{\ell}$ then taking a GD step from x_t to x_{t+1} results in $f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\eta}\|x_{t+1} - x_t\|^2$.*

Proof. We apply the ℓ-smoothness of f to get

$$\begin{aligned} f(x_{t+1}) &\leq f(x_t) + \nabla f(x_t) \cdot (x_{t+1} - x_t) + \frac{\ell}{2}\|x_{t+1} - x_t\|^2 \\ &= f(x_t) - \eta \|\nabla f(x_t)\|^2 + \frac{\eta^2 \ell}{2}\|\nabla f(x_t)\|^2 \\ &\leq f(x_t) - \frac{\eta}{2}\|\nabla f(x_t)\|^2 \\ &= f(x_t) - \frac{1}{2\eta}\|x_{t+1} - x_t\|^2 \end{aligned}$$

□

Thus despite the nonconvexity we know that for nice functions we can always decrease the function value using gradient descent. However, this doesn't tell us what kind of stationary point ($\|\nabla f\| < \varepsilon$ for small ε) we will reach after iterating. Assuming random initialization, with probability 1 we won't reach a local maximum, as we can't descend into it and so can only get stuck there by starting there. We could also reach a local minimum, which is likely the best we can do.

Finally, we could also reach a saddle point - a point that is maximal in some directions and minimal in others and has zero gradient. Visualizing this problem in two dimensions, it seems trivial to escape it - just try a couple random directions and one of them will likely lead you to descending again. However, the number of directions we'd need to try might be exponentially large in the dimension d. A standard approach here is to compute the local Hessian $\nabla^2 f$: if there is an eigenvector with a negative eigenvalue then take that descent direction. However, second-order methods are expensive, and we would like to show that a first-order algorithm can also succeed.

To gain an understanding of what to do next, let's consider what happens geometrically when we reach a stationary point. By a simple application of Cauchy-Schwarz we have the following lemma, which says that if the function value doesn't change after several iterations then gradient descent must be stuck in a ball of small radius:

Lemma 2. *If $f(x_T) - f(x_0) \geq -\mathscr{F}$ then $\forall\, t \leq T$ we have $\|x_t - x_0\| \leq \sqrt{2\eta T \mathscr{F}}$.*

Proof. We apply Cauchy-Schwarz followed by the inequality from Lemma 1:

$$\|x_t - x_0\| = \sum_{\tau=1}^{t} \|x_\tau - x_{\tau-1}\| \leq \sqrt{T \sum_{\tau=1}^{t} \|x_\tau - x_{\tau-1}\|^2} \leq \sqrt{T 2\eta (f(x_0) - f(x_T))} \leq \sqrt{2\eta T \mathscr{F}}$$

□

2.3 Escaping from Saddle Points

Lemma 2 shows that escaping from a saddle point will require escaping a ball of small radius in which the iteration is stuck. As mentioned before, the second-order solution is to take the Hessian and find the descent direction via the smallest eigenvalue. However, we can do just as well in polynomially many tries by just picking a random direction:

1. for $t = 1, \ldots, T$:
2. if no progress in a fixed number of steps: $x_t \leftarrow x_t + \xi_t$ for $x_t \sim B_0(r)$
3. $x_{t+1} \leftarrow x_t - \eta \nabla f(x_t)$

where $B_0(r)$ is the uniform distribution over the ball of radius r. For this *perturbed gradient descent* (PGD) algorithm Theorem 1 provides a guarantee for reaching an *ϵ-approximate second order stationary point*, which is defined as a point x with $\|\nabla f(x)\| \leq \varepsilon$ and $\lambda_{\min}(\nabla^2 f(x)) \geq -\sqrt{\rho\varepsilon}$ (recall that when the Hessian is positive semi-definite, i.e. has all nonnegative eigenvalues, then the stationary point is a local minimum).

Theorem 1. *PGD with $\eta = \frac{1}{\ell}$ and $r = \frac{1}{200\chi^3\sqrt{\kappa}}\sqrt{\frac{\varepsilon}{\rho}}$, where $\kappa = \frac{\ell}{\sqrt{\varepsilon\rho}}$ and $\chi = \Omega\left(\log \frac{\Delta_f d\sqrt{\kappa}}{\eta\varepsilon^2\delta}\right)$, will pass through an ε-approximate second order stationary point with $1-\varepsilon$ probability in $\mathcal{O}\left(\frac{\ell\Delta_f}{\varepsilon^2}\right)$ iterations. Here $\chi\kappa$ is the number of steps between perturbations and Δ_f is the difference between the function value at the first point of the iteration and the optimum value.*

We prove this theorem by showing that when the iteration is stuck near a saddle point then the region in the ball of sufficient radius from which regular GD won't make progress (i.e. start decreasing again) is small, so making a random jump will work after a small number of attempts. Combined with the fact the GD decreases even for non-convex functions (Lemma 1) this will complete the proof.

We first show that if GD cannot escape from a point x_0 near a saddle point $\tilde{x}$ then for large enough r_0 it can escape from all points $x_0 + re_1$, where e_1 is the eigenvector of the smallest eigenvalue of $\nabla^2 f(\tilde{x})$. The approach here is essentially to take two sequences starting $x_0, x_0 + re_1$ and show that after some time T the distance between them is large, and hence at least one must have escaped from $\tilde{x}$.

Lemma 3. *Consider $\tilde{x}$ such that $\lambda_{min}(\nabla^2 f(\tilde{x})) \leq -\sqrt{\rho\varepsilon}$ and x_0, x_0' at most distance r away from $\tilde{x}$ and $x_0' = x_0 + r_0 e_1$ for e_1 the minimum eigendirection of $\nabla^2 f(\tilde{x})$. Then for appropriate choice of T and r (depending on ε, δ, and f) and $\mathscr{F} = \tilde{\Omega}\left(\varepsilon^{\frac{3}{2}}\right)$ we have*

$$\min\{f(x_T) - f(x_0), f(x_T') - f(x_0')\} \leq -\mathscr{F}$$

where $\{x_t\}_{t=1}^T$ and $\{x_t'\}_{t=1}^T$ are sequences of GD steps starting from x_0 and x_0', respectively.

Proof. We prove by contradiction, using which we have by Lemma 2 that $\forall\, t \leq T$

$$\max\{\|x_t - \tilde{x}\|, \|x_t' - \tilde{x}\|\} \leq \{\|x_t - x_0\| + \|x_0 - \tilde{x}\|, \|x_t' - x_0'\| + \|x_0' - \tilde{x}\|\} \leq \sqrt{2\eta T\mathscr{F}} + r$$

Denoting $\mathcal{H} = \nabla^2 f(\tilde{x})$ and $\mathscr{S} = \sqrt{2\eta T\mathscr{F}} + r$ we can track the difference $w_t = x_t - x_t'$ between the two GD sequences using the gradient update:

$$w_{t+1} = w_t - \eta\left[\nabla f(x_t) - \nabla f(x_t')\right] = (I - \eta\mathcal{H})w_t - \eta\Delta_t w_t = (I - \eta\mathcal{H})^{t+1} w_0 - \eta\sum_{\tau=0}^{t}(I - \eta\mathcal{H})^{t-\tau}\Delta_\tau w_\tau$$

where $\Delta_t = \int_0^t \left[\nabla^2 f(x_t' + \theta(x_t - x_t')) - \mathcal{H}\right] d\theta$, which has norm $\|\Delta_t\| \leq \rho \max\{\|x_t - \tilde{x}\|, \|x_t' - \tilde{x}\|\} \leq \rho\mathscr{S}$ using the Hessian Lipschitzness of f.

We now consider the following statement, which bounds the second quantity above in terms of the first:

$$\left\|\eta\sum_{\tau=0}^{t-1}(I - \eta\mathcal{H})^{t-1-\tau}\Delta_\tau w_\tau\right\| \leq \frac{1}{2}\|(I - \eta\mathcal{H})^t w_0\|$$

For $t = 0$ this is obvious. We therefore assume it holds for $t' \leq t$ and prove inductively, which gives

$$\|w_t'\| = \|(I - \eta\mathcal{H})^{t'} w_0\| + \left\|\eta\sum_{\tau=0}^{t'-1}(I - \eta\mathcal{H})^{t'-1-\tau}\Delta_\tau w_\tau\right\| \leq 2\|(I - \eta\mathcal{H})^{t'} w_0\|$$

Letting $\gamma = \lambda_{\min}(\nabla^2 f(\tilde{x}))$ we have for $t < T$

$$\left\|\eta\sum_{\tau=0}^{t}(I - \eta\mathcal{H})^{t-\tau}\Delta_\tau w_\tau\right\| \leq \eta\rho\mathscr{S}\sum_{t=0}^{\tau}\|(I-\eta\mathcal{H})^{t-\tau}\|\|w_\tau\| \leq \eta\rho\sum_{\tau=0}^{t}(1+\eta\gamma)^t\|w_0\| \leq \eta\rho\mathscr{S}(t+1)\|(I-\eta\mathcal{H})^{t+1}w_0\|$$

which completes the induction so long as we pick r and η such that $2\eta\rho\mathscr{S}T \leq 1$. Thus we finally have

$$\|w_T\| \geq \|(I - \eta\mathcal{H})^T w_0\| - \left\|\eta\sum_{\tau=0}^{T-1}(I - \eta\mathcal{H})^{T-1-\tau}\Delta_\tau w_\tau\right\| \geq \frac{1}{2}\|(I - \eta\mathcal{H})^T w_0\| \geq \frac{(1 + \eta\sqrt{\rho\varepsilon})^T r_0}{2}$$

which for appropriate parameters gives $\|w_T\| \geq 2^{2\log\frac{4\mathscr{S}}{r_0} - 1} r_0 \geq 2\mathscr{S}$, a contradiction since it follows from the assumption that $\max\{\|x_T - \tilde{x}\|, \|x_T' - \tilde{x}\|\} \leq \mathscr{S}$. □

In Lemma 3 we showed at least one of two points sufficiently separated in the minimum eigendirection at $\tilde{x}$ will escape a saddle point. To complete the proof of Theorem 1 we use this to show that the *stuck region*, i.e. the region of the ball around a stationary point that is not approximately second order from which GD cannot escape, is small.

The proof follows the intuition resulting from the fact that taking random vector of length r in a random direction has projection that is Gaussian with variance $\frac{r^2}{d}$ in any fixed direction, so in expectation a step of size r in a random direction will have length $\frac{r}{\sqrt{d}}$ in any given (absolute) direction; here we are specifically interested in having a positive projection along the descent direction given by the negative eigenvalue.

Lemma 4. *Consider $\tilde{x}$ such that $\|\nabla f(\tilde{x})\| \leq \varepsilon$ and $\lambda_{min}(\nabla^2 f(\tilde{x})) \leq -\sqrt{\rho\varepsilon}$. For T let $\{x_t\}_{t=1}^T$ be a sequence of PGD steps starting from a perturbation of radius r. Then for T, r, and $\mathscr{F}$ as in Lemma 3 we have with probability at least $1-\delta$ that*

$$f(x_T) - f(\tilde{x}) \leq -\frac{\mathscr{F}}{2}$$

Proof. We first decompose the function decrease into two parts:

$$f(x_T) - f(\tilde{x}) = [f(x_T) - f(x_0)] + [f(x_0) - f(\tilde{x})] \leq f(x_T) - f(x_0) + \varepsilon r + \frac{\ell r^2}{2}$$

where the second inequality follows by applying the ℓ-Lipschitzness of f and the perturbation radius r. Choosing r such that $\varepsilon r + \frac{\ell r^2}{2} \leq \frac{\mathscr{F}}{2}$ we define the stuck region

$$\mathcal{X}_{\text{stuck}} = \left\{ x_0 : x_0 \in B_{\tilde{x}}(r) \text{ and } f(x_T) - f(x_0) \geq -\frac{\mathscr{F}}{2} \right\}$$

This is the region of the ball of radius r around $\tilde{x}$ from which gradient descent will not escape, so we want to show that it is small.

We set $r_0 = \delta r \sqrt{\frac{2\pi}{d}}$ and see from Lemma 3 that $\mathcal{X}_{\text{stuck}}$ has width at most r_0 in the minimum eigenvector direction of $\nabla^2 f(\tilde{x})$, which implies that $\text{Vol}(\mathcal{X}_{\text{stuck}}) \leq \text{Vol}(B_0^{(d-1)}(r))r_0$. Note here the correspondence of the value of r_0 with the projection of a random vector on any given direction discussed above - we can expect to travel this much in the right direction, but still need to calculate the volume of the stuck region because we don't know where it is along this line. Applying the formula for the volume of spheres in $\mathbb{R}^d$ and for appropriate choices of parameters δ and r we have that

$$\frac{\text{Vol}(\mathcal{X}_{\text{stuck}})}{\text{Vol}(B_{\tilde{x}}^{(d)}(r))} \leq \frac{r_0 \text{Vol}(B_{\tilde{x}}^{(d-1)}(r))}{\text{Vol}(B_{\tilde{x}}^{(d)}(r))} = \frac{r_0 \Gamma\left(\frac{d}{2}+1\right)}{r\sqrt{\pi}\Gamma\left(\frac{d}{2}+\frac{1}{2}\right)} \leq \frac{r_0}{r\sqrt{\pi}}\sqrt{\frac{d+1}{2}} \leq \delta$$

Thus picking a point at random from the ball of radius r will succeed in escaping the saddle point with probability at least $1-\delta$. □

The main result, Theorem 1, then follows by combining this lemma with Lemma 1.