Lecture Notes 16: Second-order smoothness and Cubic regularization

Instructor: Ashok Cutkosky

For most of this course we have consider *smooth* objectives, for which the second derivative $\nabla^2\mathcal{L}(\mathbf{w})$ satisfies the bound $\|\nabla^2\mathcal{L}(\mathbf{w})\|_{\text{op}} \le H$ for some H. Now, we will consider restricting the class of functions we consider a little more, to see if we can make some improved algorithms. Specifically, let us consider functions for which not just the second derivative, but also the *third* derivative is bounded. Such functions are callled *second-order smooth*:

Definition 1. *A function $\mathcal{L}$ is J-second-order smooth if $\mathcal{L}$ is twice-differentiable and the hessian satisfies for all x, y:*

$$\|\nabla^2\mathcal{L}(x) - \nabla^2\mathcal{L}(y)\|_{op} \le J\|x-y\|$$

Note that in the literature, the symbol for second-order smoothness is often ρ rather than J. This definition is related to the third derivative (or "jerk" as it is called in physics) in a directly analogous way to how Lipschitzness is related to the first derivative and ordinary smoothness is related to the second derivative. Note that in the literature, the symbol for second-order smoothness is often ρ rather than J. We name it J for "jerk" to help remember the symbol, and also because when writing it is often easy to get a ρ confused with a p.

In order to formalize the relationship between second-order smoothness and the third derivative, we need to think a little bit about what sort of object the third derivative actually is. The third derivative of a function $\mathcal{L}:\mathbb{R}^d\to\mathbb{R}$ can be specified by a three dimensional $d\times d\times d$ matrix whose ijk entry is $\frac{\partial^2\mathcal{L}(\mathbf{w})}{\partial\mathbf{w}[i]\partial\mathbf{w}[j]\partial\mathbf{w}[k]}$. Formally, in the same way that a (2-d) matrix M represents a bilinear function taking vectors v, w to the scalar $v^\top Mw$, a 3-d matrix T represents a trilinear function taking vectors x, y, z to a scalar via the operation:

$$(x,y,z)\mapsto \sum_{i=1}^d\sum_{j=1}^d\sum_{k=1}^d T[i,j,k]x[i]y[j]z[k]$$

If we denote the third derivative of a function $\mathcal{L}$ at a point $\mathbf{w}$ by $\nabla^3\mathcal{L}(\mathbf{w})$ and use $\nabla^3\mathcal{L}(\mathbf{w})(x,y,z)$ to indicate application of the above trilinear form to vectors x, y, z, then we can write the third-order Taylor approximation to $\mathcal{L}$ as:

$$\mathcal{L}(\mathbf{w}+\delta)\approx\mathcal{L}(\mathbf{w})+\langle\nabla\mathcal{L}(\mathbf{w}),\delta\rangle+\frac{\delta^\top\nabla^2\mathcal{L}(\mathbf{w})\delta}{2}+\frac{\nabla^2\mathcal{L}(\mathbf{w})(\delta,\delta,\delta)}{6}$$

Also, note that just as a 2-d matrix transforms vectors into other vectors, the 3-d matrices transform vectors into matrices. That is, since $\nabla^3\mathcal{L}(\mathbf{w})$ is the third derivative, we can use it to write a first-order Taylor expansion for $\nabla^2\mathcal{L}$:

$$\nabla^2\mathcal{L}(\mathbf{w}+\delta)[i,j]\approx\nabla^2\mathcal{L}(\mathbf{w})[i,j]+\sum_{k=1}^d\nabla^3\mathcal{L}(\mathbf{w})[i,j,k]\delta[k]$$

Mathematically, the objects $\nabla\mathcal{L}(\mathbf{w})$, $\nabla^2\mathcal{L}(\mathbf{w})$ and $\nabla^3\mathcal{L}(\mathbf{w})$ are sometimes said to be first, second, and third-order *tensors* respectively. Since it is somewhat cumbersome to continually say 2-d matrix vs 3-d matrix, we will call the first derivative a tensor, the second derivative a matrix, and the first derivative a vector.

We can extend the definition of operator norm from matrices to tensors as follows:

Definition 2. *The* operator norm *of a tensor T is denoted by $\|T\|_{op}$ and defined by:*

$$\sup_{\|x\|,\|y\|,\|z\|\le 1}|T(x,y,z)|$$

Notice that the trilinearity of a tensor T then implies:

$$|T(x,y,z)| \le \|x\|\|y\|\|z\|\|T\|_{\text{op}}$$

With these preliminaries out of the way, we can provide the formal connection between second-order smoothness and the third derivative:

Proposition 3. *Suppose $\mathcal{L}$ is thrice differentiable. Then $\mathcal{L}$ is J-second-order smooth if $\|\nabla^3\mathcal{L}(\mathbf{w})\|_{op} \le J$ for all $\mathbf{w}$.*

Proof. Let x and y be arbitrary points. Then by the mean value theorem, there is some z such that:

$$\nabla^2\mathcal{L}(y)[i,j] = \nabla^2\mathcal{L}(x)[i,j] + \sum_{k=1}^{d} \nabla^3\mathcal{L}(z)[i,j,k](x-y)[k]$$

Therefore, for any unit vectors v and w:

$$\begin{aligned} v^\top(\nabla^2\mathcal{L}(y) - \nabla^2\mathcal{L}(x))w &= \nabla^3\mathcal{L}(z)(v,w,(x-y)) \\ &\le \|v\|\|w\|\|x-y\|\|\nabla^3\mathcal{L}(z)\|_{\text{op}} \\ &\le J\|x-y\| \end{aligned}$$

where the last line uses the boundedness of $\nabla^3\mathcal{L}(z)$ and that v and w are unit-vectors. But $\sup_{\|v\|=1,\|w\|=1} v^\top(\nabla^2\mathcal{L}(y) - \nabla^2\mathcal{L}(x))w = \|\nabla^2\mathcal{L}(y) - \nabla^2\mathcal{L}(x)\|_{\text{op}}$, so this implies $\|\nabla^2\mathcal{L}(y) - \nabla^2\mathcal{L}(x)\|_{\text{op}} \le J\|x-y\|$ and $\mathcal{L}$ is J-second-order smooth. $\square$

Next, we can provide an analog of the "key smoothness lemma":

Lemma 4. *Suppose $\mathcal{L}$ is J-second-order smooth. Then for any x and δ:*

$$\mathcal{L}(x+\delta) \le \mathcal{L}(x) + \langle\nabla\mathcal{L}(x),\delta\rangle + \frac{\delta^\top\nabla^2\mathcal{L}(x)\delta}{2} + \frac{J\|\delta\|^3}{6}$$

Proof. From fundamental theorem of calculus (used twice):

$$\begin{aligned} \mathcal{L}(x+\delta) &= \mathcal{L}(x) + \int_0^1 \langle\nabla\mathcal{L}(x+t\delta),\delta\rangle\, dt \\ &= \mathcal{L}(x) + \int_0^1\int_0^1 \langle\nabla\mathcal{L}(x),\delta\rangle + t\delta^\top\nabla^2\mathcal{L}(x+tk\delta)\delta\, dkdt \\ &= \mathcal{L}(x) + \langle\nabla\mathcal{L}(x),\delta\rangle + \int_0^1\int_0^1 t\delta^\top\nabla^2\mathcal{L}(x)\delta + t\delta^\top(\nabla^2\mathcal{L}(x+tk\delta) - \nabla^2\mathcal{L}(x))\delta\, dkdt \\ &= \mathcal{L}(x) + \langle\nabla\mathcal{L}(x),\delta\rangle + \frac{\delta^\top\nabla^2\mathcal{L}(x)\delta}{2} + \int_0^1\int_0^1 t\delta^\top(\nabla^2\mathcal{L}(x+tk\delta) - \nabla^2\mathcal{L}(x))\delta\, dkdt \\ &\le \mathcal{L}(x) + \langle\nabla\mathcal{L}(x),\delta\rangle + \frac{\delta^\top\nabla^2\mathcal{L}(x)\delta}{2} + \int_0^1\int_0^1 t\|\delta\|^2\|\nabla^2\mathcal{L}(x+tk\delta) - \nabla^2\mathcal{L}(x)\|_{\text{op}}\, dkdt \\ &\le \mathcal{L}(x) + \langle\nabla\mathcal{L}(x),\delta\rangle + \frac{\delta^\top\nabla^2\mathcal{L}(x)\delta}{2} + \int_0^1\int_0^1 t^2kJ\|\delta\|^3\, dkdt \\ &= \mathcal{L}(x) + \langle\nabla\mathcal{L}(x),\delta\rangle + \frac{\delta^\top\nabla^2\mathcal{L}(x)\delta}{2} + \frac{J\|\delta\|^3}{6} \end{aligned}$$

$\square$

We also have the following bound on the change in the gradients, which is also an intuitive consequence of Taylor series ideas:

Lemma 5. *Suppose $\mathcal{L}$ is J-second order smooth. Then for any $\mathbf{w}$ and δ:*

$$\|\nabla\mathcal{L}(\mathbf{w}+\delta) - (\nabla\mathcal{L}(\mathbf{w}) + \nabla^2\mathcal{L}(\mathbf{w})\delta)\| \le \frac{J\|\delta\|^2}{2}$$

Proof. By the fundamental theorem of calculus:

$$\begin{aligned}
\nabla\mathcal{L}(\mathbf{w}+\delta) &= \nabla\mathcal{L}(\mathbf{w}) + \int_0^1 \nabla^2\mathcal{L}(\mathbf{w}+t\delta)\delta\, dt \\
&= \nabla\mathcal{L}(\mathbf{w}) + \nabla^2\mathcal{L}(\mathbf{w})\delta + \int_0^1 (\nabla^2\mathcal{L}(\mathbf{w}+t\delta) - \nabla^2\mathcal{L}(\mathbf{w}))\delta\, dt \\
\|\nabla\mathcal{L}(\mathbf{w}+\delta) - (\nabla\mathcal{L}(\mathbf{w}) + \nabla^2\mathcal{L}(\mathbf{w})\delta)\| &\le \left\|\int_0^1 (\nabla^2\mathcal{L}(\mathbf{w}+t\delta) - \nabla^2\mathcal{L}(\mathbf{w}))\delta\, dt\right\| \\
&\le \int_0^1 \|(\nabla^2\mathcal{L}(\mathbf{w}+t\delta) - \nabla^2\mathcal{L}(\mathbf{w}))\delta\|\, dt \\
&\le \int_0^1 \|\nabla^2\mathcal{L}(\mathbf{w}+t\delta) - \nabla^2\mathcal{L}(\mathbf{w})\|_{\text{op}}\|\delta\|\, dt \\
&\le \int_0^1 Jt\|\delta\|^2 \\
&\le \frac{J\|\delta\|^2}{2}
\end{aligned}$$

□

What can we do with second-order smooth functions? With regular smooth functions, we were trying to find critical points where the gradient is small. With second-order smooth functions, we can do much better. Not only will we be able to find critical points faster, we will be able to find *second-order stationary points*, otherwise known as (approximate) *local minima*.

Definition 6. *A point* $\mathbf{w}$ *is an* (α, β)*-second order stationary point of a function* $\mathcal{L}$ *if:*

$$\begin{aligned}
\|\nabla\mathcal{L}(\mathbf{w})\| &\le \alpha \\
\lambda_{\min}(\nabla^2\mathcal{L}(\mathbf{w})) &\ge -\beta
\end{aligned}$$

where $\lambda_{\min}(M)$ *indicates the smallest eigenvalue of a matrix* M.

Intuitively, if $\beta = 0$ in the above, that means that the hessian is positive semi-definite at this point so that if the gradient is also small, we must be at a local minimum. Thus, allowing a small positive β is a way to relax the notion of local minimum.

One way to try to minimize second-order smooth functions is using the *cubic-regularized newton step* [1]. This algorithm makes the update:

$$\mathbf{w}_{t+1} = \operatorname*{argmin}_{\mathbf{w}} \mathcal{L}(\mathbf{w}_t) + \langle \mathcal{L}(\mathbf{w}_t), \mathbf{w} - \mathbf{w}_t\rangle + \frac{(\mathbf{w}-\mathbf{w}_t)^\top\nabla^2\mathcal{L}(\mathbf{w}_t)(\mathbf{w}-\mathbf{w}_t)}{2} + \frac{J\|\mathbf{w}-\mathbf{w}_t\|^3}{6}$$

Now, actually solving for the value of $\mathbf{w}_{t+1}$ here is itself a non-convex optimization problem. However, it turns out that since it has this very special cubic form, it is possible to reformulate it into a way that can be solved efficiently. However, let's not worry about that now and instead get some idea for how this update will perform.

Previously we've exploited the fact that a large gradient value means that it is possible to make the function value decrease a lot. Here, we will exploit a different fact: if $\nabla^2\mathcal{L}(\mathbf{w})$ has an eigenvector with large negative eigenvalue, then it is also possible to make the function value decrease a lot. Let's see how. Suppose $\nabla^2\mathcal{L}(\mathbf{w})\mathbf{v} = -\lambda\mathbf{v}$ for some $\lambda > 0$ and unit vector $\mathbf{v}$. Notice that we also have $\nabla^2\mathcal{L}(\mathbf{w})(-\mathbf{v}) = -\lambda(-\mathbf{v})$. Therefore, after possibly replacing $\mathbf{v}$ with $-\mathbf{v}$, we may as well assume that $\langle \mathbf{v}, \nabla\mathcal{L}(\mathbf{w})\rangle \le 0$. Let's consider $\mathbf{w} + \eta\mathbf{v}$ for some to-be-specified η. Then we have:

$$\mathcal{L}(\mathbf{w}+\eta\mathbf{v}) \le \mathcal{L}(\mathbf{w}) + \langle \nabla\mathcal{L}(\mathbf{w}), \eta\mathbf{v}\rangle + \frac{\eta^2\mathbf{v}^\top\nabla^2\mathcal{L}(\mathbf{w})\mathbf{v}}{2} + \frac{J\eta^3\|\mathbf{v}\|^3}{6}$$

using $\langle \nabla\mathcal{L}(\mathbf{w}), \mathbf{v}\rangle \le 0$, $\nabla^2\mathcal{L}(\mathbf{w})\mathbf{v} = -\lambda\mathbf{v}$ and $\|\mathbf{v}\| = 1$:

$$\le \mathcal{L}(\mathbf{w}) - \frac{\eta^2\lambda}{2} + \frac{J\eta^3}{6}$$

Now, set $\eta = \frac{2\lambda}{J}$:

$$\leq \mathcal{L}(\mathbf{w}) - \frac{2\lambda^3}{3J^2}$$

Now, this is similar to how we had a function decrease of $O(-\|\nabla\mathcal{L}(\mathbf{w})\|^2/H)$, but now the decrease is in terms of the eigenvalue of $\nabla^2\mathcal{L}(\mathbf{w})$ and is cubic rather than quadratic.

Thus, we have the following result:

Theorem 7. *Suppose $\mathcal{L}$ is J-second-order smooth, and for some given $\mathbf{w}$ there is a unit vector $\mathbf{v}$ such that $\mathbf{v}^\top\nabla^2\mathcal{L}(\mathbf{w})\mathbf{v} < -\lambda$ for some $\lambda \geq 0$. Let $\delta = -\frac{2\lambda}{J}\mathbf{v}sign(\langle \mathbf{v}, \nabla\mathcal{L}(\mathbf{w})\rangle)$, where $sign(x)$ is 1 if $x \geq 0$ and -1 otherwise. Then:*

$$\mathcal{L}\left(\mathbf{w}+\delta\right) \leq \mathcal{L}(\mathbf{w}) - \frac{2\lambda^3}{3J^2}$$

Proof. Notice by definition of δ we have $\langle\nabla\mathcal{L}(\mathbf{w}), \delta\rangle \leq 0$. Further, $\delta^\top\nabla^2\mathcal{L}(\mathbf{w})\delta < -\frac{4\lambda^3}{J^2}$ and $\|\delta\|^3 = \frac{8\lambda^3}{J^3}$. Therefore

$$\begin{aligned}\mathcal{L}(\mathbf{w}+\delta) &\leq \mathcal{L}(\mathbf{w}) + \langle\nabla\mathcal{L}(\mathbf{w}), \delta\rangle + \frac{\delta^\top\nabla^2\mathcal{L}(\mathbf{w})\delta}{2} + \frac{\|\delta\|^3 J}{6}\\ &< \mathcal{L}(\mathbf{w}) - \frac{2\lambda^3}{3J^2}\end{aligned}$$

□

Now, let's define

$$\delta_t^\star = \operatorname*{argmin}_{\delta}\langle\nabla\mathcal{L}(\mathbf{w}_t), \delta\rangle + \frac{\delta^\top\nabla^2\mathcal{L}(\mathbf{w})\delta}{2} + \frac{\|\delta\|^3 J}{6}$$

and so the cubic regularized newton step algorithm is:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \delta_t^\star$$

Let's also define the "progress" function:

$$P_t(\delta) = \langle\nabla\mathcal{L}(\mathbf{w}_t), \delta\rangle + \frac{\delta^\top\nabla^2\mathcal{L}(\mathbf{w})\delta}{2} + \frac{\|\delta\|^3 J}{6}$$

so that

$$\mathcal{L}(\mathbf{w}_{t+1}) \leq \mathcal{L}(\mathbf{w}) + P_t(\delta_t^\star)$$

and $\delta_t^\star = \text{argmin } P_t(\delta)$. Now, by Theorem 7, we have just seen that if the smallest eigenvalue of $\nabla^2\mathcal{L}(\mathbf{w})$ is denoted by $-\lambda(\mathbf{w})$, then if $\lambda(\mathbf{w}) \geq 0$, we have the bound:

$$P_t(\delta_t^\star) \leq -\frac{2\lambda(\mathbf{w})^3}{3J^2}$$

Our overall goal to analyze the cubic-regularized newton step algorithm is roughly to show that, at each iteration for any ϵ, either we have found a point with both $\|\nabla\mathcal{L}(\mathbf{w})\| \leq \epsilon$ and $\lambda(\mathbf{w}) \leq \sqrt{\epsilon}$, or $\mathcal{L}(\mathbf{w})$ will decrease by at least $\epsilon^{3/2}$. Therefore, since we can only decrease by $\epsilon^{3/2}$ at most $\frac{\mathcal{L}(\mathbf{w}_1)-\mathcal{L}(\mathbf{w}_\star)}{\epsilon^{3/2}}$ times, we see that by setting $\epsilon = O(1/T^{2/3})$, there must be some iteration at which we do not decrease the function value be $\epsilon^{3/2}$, and so we find an $(\epsilon, \sqrt{\epsilon})$-second-order stationary point.

Theorem 8. *After T steps of cubic-regularized newton step, there must exist some $t \in \{1, \ldots, T\}$ such that:*

$$\begin{aligned}\|\nabla\mathcal{L}(\mathbf{w}_{t+1})\| &\leq \frac{6^{2/3}\Delta^{2/3}J^{4/3}}{T^{2/3}}\\ \lambda(\mathbf{w}_{t+1}) &\leq 2\frac{6^{1/3}\Delta^{1/3}J^{2/3}}{T^{1/3}}\end{aligned}$$

That is, $\mathbf{w}_{t+1}$ is a $\left(\frac{6^{2/3}\Delta^{2/3}J^{4/3}}{T^{2/3}}, 2\frac{6^{1/3}\Delta^{1/3}J^{2/3}}{T^{1/3}}\right)$ second-order stationary point.

Proof. Let us define $\Delta = \mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star)$ and $\epsilon = \frac{6^{2/3}\Delta^{2/3}J^{4/3}}{T^{2/3}}$. Now, let's consider some index $t \in \{1, \dots, T\}$ and do some casework.

First, suppose $\lambda(\mathbf{w}_t) > \sqrt{\epsilon}$. Then by Theorem 7, we have:

$$\mathcal{L}(\mathbf{w}_{t+1}) < \mathcal{L}(\mathbf{w}_t) - \frac{2\epsilon^{3/2}}{3J^2}$$

Next, let's suppose $\lambda(\mathbf{w}_t) \le \sqrt{\epsilon}$. Now, notice that since $\delta_t^\star = \text{argmin } P_t(\delta)$, we have $\nabla P_t(\delta_t^\star) = 0$. Therefore:

$$\nabla \mathcal{L}(\mathbf{w}_t) + \nabla^2 \mathcal{L}(\mathbf{w}_t)\delta_t^\star + \frac{\delta_t^\star \|\delta_t^\star\| J}{2} = 0 \tag{1}$$

Further, since $\mathbf{w}_{t+1} = \mathbf{w}_t + \delta_t^\star$, by Lemma 5, there is some vector $\mathbf{x}$ with $\|\mathbf{x}\| \le \frac{J\|\delta_t^\star\|^2}{2}$ such that:

$$\nabla \mathcal{L}(\mathbf{w}_t) + \nabla^2 \mathcal{L}(\mathbf{w}_t)\delta_t^\star = \nabla \mathcal{L}(\mathbf{w}_{t+1}) + \mathbf{x}$$

Therefore:

$$\nabla \mathcal{L}(\mathbf{w}_{t+1}) + \mathbf{x} + \frac{\delta_t^\star \|\delta_t^\star\| J}{2} = 0$$

which in turn implies:

$$\|\nabla \mathcal{L}(\mathbf{w}_{t+1})\| \le \|\mathbf{x}\| + \frac{\|\delta_t^\star\|^2 J}{2} \le J\|\delta_t^\star\|^2$$

Now, let's suppose that $\|\delta_t^\star\| \le \frac{\sqrt{\epsilon}}{J}$. Then we have

$$\|\nabla \mathcal{L}(\mathbf{w}_{t+1})\| \le \epsilon$$

And further, since $\lambda(\mathbf{w}) \le \sqrt{\epsilon}$, we have for all unit vectors $\mathbf{v}$:

$$\begin{aligned}
\mathbf{v}^\top \mathcal{L}(\mathbf{w}_{t+1})\mathbf{v} &= \mathbf{v}^\top(\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t))\mathbf{v} + \mathbf{v}^\top \mathcal{L}(\mathbf{w}_t)\mathbf{v} \\
&\ge -J\|\mathbf{w}_{t+1} - \mathbf{w}_t\| + \mathbf{v}^\top \mathcal{L}(\mathbf{w}_t)\mathbf{v} \\
&\ge -J\|\delta_t^\star\| - \sqrt{\epsilon} \\
&\ge -2\sqrt{\epsilon}
\end{aligned}$$

so that $\lambda(\mathbf{w}_{t+1}) \le 2\sqrt{\epsilon}$ and so $\mathbf{w}_{t+1}$ is an $(\epsilon, 2\sqrt{\epsilon})$ second order stationary point.

Now, suppose instead that $\|\delta_t^\star\| > \frac{\sqrt{\epsilon}}{J}$. Then, taking the inner product of equation (1) with $\delta_t^\star$ on both sides, we have:

$$\begin{aligned}
\langle \nabla \mathcal{L}(\mathbf{w}_t), \delta_t^\star \rangle + {\delta_t^\star}^\top \nabla^2 \mathcal{L}(\mathbf{w}_t)\delta_t^\star + \frac{\|\delta_t^\star\|^3 J}{2} &= 0 \\
P_t(\delta_t^\star) + \frac{{\delta_t^\star}^\top \nabla^2 \mathcal{L}(\mathbf{w})\delta_t^\star}{2} + \frac{\|\delta_t^\star\|^3 J}{3} &= 0 \\
P_t(\delta_t^\star) &= -\frac{{\delta_t^\star}^\top \nabla^2 \mathcal{L}(\mathbf{w})\delta_t^\star}{2} - \frac{\|\delta_t^\star\|^3 J}{3} \\
&\le \frac{\lambda(\mathbf{w})\|\delta_t^\star\|^2}{2} - \frac{\|\delta_t^\star\|^3 J}{3} \\
&\le \frac{\sqrt{\epsilon}\|\delta_t^\star\|^2}{2} - \frac{\|\delta_t^\star\|^3 J}{3} \\
&< -\frac{\epsilon^{3/2}}{6J^2}
\end{aligned}$$

In summary we have the following possibilities:

1. $\mathbf{w}_{t+1}$ is an $(\epsilon, 2\sqrt{\epsilon})$ second-order stationary point.

2. $\lambda(\mathbf{w}_t) > \sqrt{\epsilon}$ and $\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) < -\frac{2\epsilon^{3/2}}{3J}$.

3. $\lambda(\mathbf{w}_t) \le \sqrt{\epsilon}$ and $\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) < -\frac{\epsilon^{3/2}}{6J^2}$.

We can compactify these cases a little bit by noticing that the conclusion of the third case is strictly weaker than the conclusion of the second case:

1. $\mathbf{w}_{t+1}$ is an $(\epsilon, 2\sqrt{\epsilon})$ second-order stationary point.

2. If case 1 does not occur, then $\mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) < -\frac{\epsilon^{3/2}}{6J^2}$.

Suppose case 1 never occurs. Then by definition of ϵ, this would imply:

$$-\Delta \le \mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_1) = \sum_{t=1}^{T} \mathcal{L}(\mathbf{w}_{t+1}) - \mathcal{L}(\mathbf{w}_t) < -\frac{T\epsilon^{3/2}}{6J^2} = -\Delta$$

so that $-\Delta < -\Delta$, which is a contradiction. Therefore there must be at least one iteration in which case 1 happens, which implies the desired result. □

There has been a fair amount of recent work on second-order smooth optimization. In the deterministic setting, see [2] for an algorithm that does not require the ability to compute the entire hessian in order to achieve faster convergence rates (it's also a nice application of the almost-convex optimization algorithm we saw earlier). In the stochastic setting, see [3] for a proof that SGD actually converges faster on second-order smooth functions than the analysis we provided previously in lecture, or [4] for a slightly more complicated algorithm but much simpler proof of a similar result. The lower bounds in this setting are not as well understood. In the case of stochastic methods, the results of [3, 4] are known to be tight, as recently shown by [5]. However, in the deterministic setting there is a bit of a gap between the lower bounds and the best known algorithms see [6] and [7] for a description of the lower bounds currently known in this case.

References

[1] Yurii Nesterov and Boris T Polyak. "Cubic regularization of Newton method and its global performance". In: *Mathematical Programming* 108.1 (2006), pp. 177–205.

[2] Yair Carmon et al. ""Convex Until Proven Guilty": Dimension-Free Acceleration of Gradient Descent on Non-Convex Functions". In: *International Conference on Machine Learning*. PMLR. 2017, pp. 654–663.

[3] Cong Fang, Zhouchen Lin, and Tong Zhang. "Sharp analysis for nonconvex sgd escaping from saddle points". In: *Conference on Learning Theory*. PMLR. 2019, pp. 1192–1234.

[4] Ashok Cutkosky and Harsh Mehta. "Momentum improves normalized sgd". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 2260–2268.

[5] Yossi Arjevani et al. "Second-order information in non-convex stochastic optimization: Power and limitations". In: *Conference on Learning Theory*. PMLR. 2020, pp. 242–299.

[6] Yair Carmon et al. "Lower bounds for finding stationary points i". In: *Mathematical Programming* (2019), pp. 1–50.

[7] Yair Carmon et al. "Lower bounds for finding stationary points II: first-order methods". In: *Mathematical Programming* 185.1-2 (2021).