Lecture Notes 13: Stochastic Strongly-Convex Optimization and Almost-Convexity

Instructor: Ashok Cutkosky

Previously we considered smooth and strongly-convex losses and showed that *deterministic* gradient descent is able to achieve a convergence rate of:

$$\mathcal{L}(\mathbf{w}_{T+1}) - \mathcal{L}(\mathbf{w}_\star) \leq \exp\left(-\frac{\mu}{H}T\right)(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))$$

Let's now consider the *stochastic* case in which we cannot trust our gradient evaluations to be completely accurate. Recall that previously we showed that SGD on ordinary convex functions achieves:

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \leq O\left(\frac{1}{\sqrt{T}}\right)$$

Now, we will show that if $\mathcal{L}$ is known to be μ-strongly convex, then SGD can instead obtain:

$$\frac{1}{T}\mathbb{E}\left[\sum_{t=1}^T \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)\right] \leq O\left(\frac{\log(T)}{\mu T}\right)$$

Thus, even in the stochastic case, we see a significant improvement stemming from strong-convexity. However, note that the improvement is quite a bit less dramatic than in the case of deterministic losses. The analysis is remarkably similar to the analysis for ordinary convex losses:

Theorem 1. *Suppose that $\mathcal{L}(\mathbf{w}) = \mathbb{E}[\ell(\mathbf{w}, z)]$ is a μ-strongly convex function satisfiying $\|\mathbf{w}_\star\| \leq D$ for some known D, and $\mathbb{E}[\|\nabla\ell(\mathbf{w}, z)\|^2] \leq G^2$ for all $\|\mathbf{w}\| \leq D$. Consider projected stochastic gradient descent with learning rate $\eta_t = \frac{1}{\mu t}$:*

$$\mathbf{w}_{t+1} = \prod_{\|\mathbf{w}\|\leq D}[\mathbf{w}_t - \eta_t\nabla\ell(\mathbf{w}_t, z_t)]$$

This algorithm satisfies:

$$\frac{1}{T}\sum_{t=1}^T \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \frac{G^2(\log(T)+1)}{2\mu}$$

Proof. The proof begins in exactly the same way as for previous convex analysis. Since $\|\mathbf{w}_\star\| \leq D$, by properties of the projection we have:

$$\begin{aligned}
\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2 &\leq \|\mathbf{w}_t - \eta_t\nabla\ell(\mathbf{w}_t, z_t) - \mathbf{w}_\star\|^2 \\
&= \|\mathbf{w}_t - \mathbf{w}_\star\|^2 - 2\eta_t\langle\nabla\ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_\star\rangle + \eta_t^2\|\nabla\ell(\mathbf{w}_t, z_t)\|^2 \\
\langle\nabla\ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_\star\rangle &\leq \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2\eta_t} - \frac{\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t\|\nabla\ell(\mathbf{w}_t, z_t)\|^2}{2}
\end{aligned}$$

Now, previously we simply argued that by convexity, $\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \mathbb{E}[\langle\nabla\ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_\star\rangle]$. However, now we will instead leverage strong convexity. Recall that an equivalent characterization of strong convexity is the condition:

$$\mathcal{L}(x) \geq \mathcal{L}(y) + \langle\nabla\mathcal{L}(y), x - y\rangle + \frac{\mu}{2}\|x - y\|^2$$

From this, we can conclude:

$$\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) \leq \langle \nabla \mathcal{L}(\mathbf{w}_t), \mathbf{w}_t - \mathbf{w}_\star \rangle - \frac{\mu}{2}\|\mathbf{w}_t - \mathbf{w}_\star\|^2$$
$$\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \mathbb{E}\left[\langle \nabla \ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_\star \rangle - \frac{\mu}{2}\|\mathbf{w}_t - \mathbf{w}_\star\|^2\right]$$

Plugging this into our previous equation yields:

$$\langle \nabla \ell(\mathbf{w}_t, z_t), \mathbf{w}_t - \mathbf{w}_\star \rangle \leq \frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2\eta_t} - \frac{\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}$$
$$\mathbb{E}\left[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star) + \frac{\mu}{2}\|\mathbf{w}_t - \mathbf{w}_\star\|^2\right] \leq \mathbb{E}\left[\frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2\eta_t} - \frac{\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t \|\nabla \ell(\mathbf{w}_t, z_t)\|^2}{2}\right]$$
$$\leq \mathbb{E}\left[\frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2\eta_t} - \frac{\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} + \frac{\eta_t G^2}{2}\right]$$
$$\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \mathbb{E}\left[\frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2\eta_t} - \frac{\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} - \frac{\mu\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2} + \frac{\eta_t G^2}{2}\right]$$

Now, it is time to invoke our particular magical choice for $\eta_t = \frac{1}{\mu t}$:

$$\mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \mathbb{E}\left[\frac{\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2\eta_t} - \frac{\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2\eta_t} - \frac{\mu\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2} + \frac{\eta_t G^2}{2}\right]$$
$$= \mathbb{E}\left[\frac{\mu\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2t} - \frac{\mu\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2t} - \frac{\mu\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2} + \frac{G^2}{2\mu t}\right]$$
$$= \mathbb{E}\left[\frac{(t-1)\mu\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2} - \frac{t\mu\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2} + \frac{G^2}{2\mu t}\right]$$

now, sum over t and telescope:

$$\sum_{t=1}^{T} \mathbb{E}[\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_\star)] \leq \sum_{t=1}^{T} \mathbb{E}\left[\frac{(t-1)\mu\|\mathbf{w}_t - \mathbf{w}_\star\|^2}{2} - \frac{t\mu\|\mathbf{w}_{t+1} - \mathbf{w}_\star\|^2}{2} + \frac{G^2}{2\mu t}\right]$$
$$= \mathbb{E}\left[-\frac{T\mu\|\mathbf{w}_{T+1} - \mathbf{w}_\star\|^2}{2} + \sum_{t=1}^{T} \frac{G^2}{2\mu t}\right]$$
$$\leq \mathbb{E}\left[\sum_{t=1}^{T} \frac{G^2}{2\mu t}\right]$$
$$\leq \frac{G^2(\log(T) + 1)}{2\mu}$$

□

There is one quite subtle point in the above argument: although we assumed $\|\mathbf{w}_\star\| \leq D$, it doesn't seem that it was used anywhere in the analysis! This is because the assumption $\mathbb{E}[\|\nabla \ell(\mathbf{w}, z)\|^2] \leq G^2$ actually *requires* some kind of restriction on $\mathbf{w}$. To see why, observe the following fact about strongly convex functions:

Lemma 2. *Suppose $\mathcal{L}(\mathbf{w})$ is μ-strongly-convex. Then $\|\nabla \mathcal{L}(\mathbf{w})\| \geq \mu\|\mathbf{w} - \mathbf{w}_\star\|$.*

Proof. By mean value theorem, for some x we have:

$$\nabla \mathcal{L}(\mathbf{w}) - \nabla \mathcal{L}(\mathbf{w}_\star) = \nabla^2 \mathcal{L}(x)(\mathbf{w} - \mathbf{w}_\star)$$

Now, since $\nabla \mathcal{L}(\mathbf{w}_\star) = 0$, we have:

$$\nabla \mathcal{L}(\mathbf{w}) = \nabla^2 \mathcal{L}(x)(\mathbf{w} - \mathbf{w}_\star)$$

Therefore:

$$\begin{aligned}\|\nabla \mathcal{L}(\mathbf{w})\| &= \sup_{\|v\|=1} \langle v, \nabla \mathcal{L}(\mathbf{w})\rangle \\ &= \sup_{\|v\|=1} v^\top \nabla^2 \mathcal{L}(x)(\mathbf{w}-\mathbf{w}_\star) \\ &\geq \frac{(\mathbf{w}-\mathbf{w}_\star)^\top \nabla^2 \mathcal{L}(x)(\mathbf{w}-\mathbf{w}_\star)}{\|\mathbf{w}-\mathbf{w}_\star\|}\end{aligned}$$

now use strong convexity

$$\geq \frac{\mu\|\mathbf{w}-\mathbf{w}_\star\|^2}{\|\mathbf{w}-\mathbf{w}_\star\|}$$

□

This result implies that it is actually impossible to have $\mathbb{E}[\|\nabla \ell(\mathbf{w}, z)\|^2] \leq G^2$ for any fixed constant G unless we also guarantee that $\|\mathbf{w}-\mathbf{w}_\star\|$ is bounded. Since it's always possible for the noise in the stochastic gradients to push $\mathbf{w}_t$ very far away from $\mathbf{w}_\star$, we needed to assume $\|\mathbf{w}_\star\| \leq D$ and use projections to keep the distance bounded, which in turn allows for the bounded gradient assumption to hold.

Note that this result does not rely on smoothness. In fact, as you saw on the homework, adding a smoothness assumption enables us to remove the bounded gradient assumption, at the cost of an extra $\frac{H}{\mu}$ factor in the error.

Almost Convexity

At this point, we have considered losses that are H-smooth and non-convex, losses that are convex and H-smooth, and losses that are μ=strongly convex and H-smooth. When written in terms of bounds on the Hessian, these conditions correspond to:

- H-smoothness: $-HI \preceq \nabla^2 \mathcal{L}(\mathbf{w}) \preceq HI$.
- Convexity: $0 \preceq \nabla^2 \mathcal{L}(\mathbf{w})$.
- μ-strong convexity: $\mu I \preceq \nabla^2 \mathcal{L}(\mathbf{w})$.

It seems that something is missing from this categorization of the Hessian: what if $\nabla^2 \mathcal{L}(\mathbf{w})$ is bounded from below by some negative number between $-H$ and 0? This condition has a few different names in the literature, but we will call it γ-almost convexity:

Definition 3. *A twice-differentiable function $\mathcal{L}$ is γ-almost convex if for all $\mathbf{w}$:*

$$-\gamma I \preceq \nabla^2 \mathcal{L}(\mathbf{w})$$

Thus, all H-smooth functions are automatically H-almost convex, and all convex functions are 0-almost convex. The natural question now is whether we can take advantage of γ-almost convexity to obtain improved bounds.

On the homework, you will design an *accelerated* algorithm for H-smooth and μ-strongly convex objectives such that after T iterations, such that the algorithm outputs a point $\hat{\mathbf{w}}$ with:

$$\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) \leq \exp\left(-C\sqrt{\frac{\mu}{H}}T\right)\frac{H\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{2}$$

where C is an absolute constant.

From this, we have the following result:

Theorem 4. *Run the accelerated gradient descent for strongly-convex functions for $N = \frac{\log(\frac{H^2\|\mathbf{w}_\star\|^2}{\epsilon^2})}{C\sqrt{\frac{\mu}{H}}}$ iterations starting at the origin. Then the output $\hat{\mathbf{w}}$ satisfies:*

$$\mathcal{L}(\hat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_\star) \leq \frac{\epsilon^2}{2H}\|\nabla \mathcal{L}(\hat{\mathbf{w}})\| \qquad \leq \epsilon$$

Proof. Recall that smooth functions satisfy: $\|\nabla\mathcal{L}(\mathbf{w})\|^2 \leq 2H(\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star))$. Thus we have

$$\begin{aligned}\|\nabla\mathcal{L}(\hat{\mathbf{w}})\|^2 &\leq 2H(\mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_\star)) \\ &\leq 2H\exp\left(-C\sqrt{\frac{\mu}{H}}\frac{\log(\frac{H^2\|\mathbf{w}_\star - \mathbf{w}_1\|^2}{\epsilon^2})}{C\sqrt{\frac{\mu}{H}}}\right)\frac{H\|\mathbf{w}_\star\|^2}{2} \\ &= \epsilon^2\end{aligned}$$

□

How can we use this fact to optimize a γ-almost convex function? The key idea is regularization. For any given point $\mathbf{w}_t$, consider the function $F_t(\delta) = \mathcal{L}(\mathbf{w}_t + \delta) + \gamma\|\delta\|^2$. Note that so long as $\mathcal{L}$ is γ-strongly-convex, the hessian of this function satisfies $\nabla^2 F_t(\delta) = \nabla^2\mathcal{L}(\mathbf{w}_t + \delta) + 2\gamma I \succeq \gamma I$, so that F_t is γ-strongly convex and $H + 2\gamma \leq 3H$ smooth. Let us further assume that $\mathcal{L}(\mathbf{w})$ is G-Lipschitz. Then if $\delta_\star = \operatorname{argmin} F_t(\delta)$, we have $\|\delta_\star\| \leq \frac{G}{2\gamma}$ because $0 = \nabla F_t(\delta_\star) = \nabla\mathcal{L}(\delta_\star) + 2\gamma\delta_\star$, so $\delta_\star = \frac{-\nabla\mathcal{L}(\delta_\star)}{2\gamma}$.

Therefore, by Theorem 4, after $N = \frac{\log(\frac{9H^2G^2}{4\gamma^2\epsilon^2})}{C\sqrt{\frac{\gamma}{6H}}}$ we obtain a point $\hat{\delta}$ such that:

$$\begin{aligned}F_t(\hat{\delta}) - F_t(\delta_\star) &\leq \frac{\epsilon^2}{6H} \\ \|\nabla F_t(\hat{\delta})\| &\leq \epsilon\end{aligned}$$

Our overall algorithm will be take the following idea: given any iterate $\mathbf{w}_t$, we will form the function F_t and optimize it using accelerated gradient descent to get some $\hat{\delta}_t$. Then we set $\mathbf{w}_{t+1} = \mathbf{w}_t + \hat{\delta}_t$ and repeat the process. Why should this be a good idea? From the definition, we have that that $F_t(\hat{\delta}_t) \geq \mathcal{L}(\mathbf{w}_{t+1})$, and we might expect to be able to optimize F_t quickly because F_t is γ-strongly convex. Formally, if $\delta_{t,\star} = \operatorname{argmin} F_t(\delta)$, we have the following:

$$\begin{aligned}\mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_{t+1}) &= \mathcal{L}(\mathbf{w}_t) - \mathcal{L}(\mathbf{w}_t + \hat{\delta}_t) - \gamma\|\hat{\delta}_t\|^2 + \gamma\|\hat{\delta}_t\|^2 \\ &= F_t(0) - F_t(\hat{\delta}_t) + \gamma\|\hat{\delta}_t\|^2 \\ &\geq F_t(\delta_{t,\star}) - F_t(\hat{\delta}_t) + \gamma\|\hat{\delta}_t\|^2 \\ \mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star) + \sum_{t=1}^T F_t(\hat{\delta}_t) - F_t(\delta_{t,\star}) &\geq \gamma\sum_{t=1}^T\|\hat{\delta}_t\|^2 \\ \mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star) + \frac{T\epsilon^2}{6H} &\geq \gamma\sum_{t=1}^T\|\hat{\delta}_t\|^2\end{aligned}$$

Now, we also have:

$$\begin{aligned}\epsilon &\geq \|\nabla F_t(\hat{\delta}_t)\| \\ &= \|\nabla\mathcal{L}(\mathbf{w}_{t+1}) + 2\gamma\hat{\delta}_t\| \\ \epsilon + 2\gamma\|\hat{\delta}_t\| &\geq \|\nabla\mathcal{L}(\mathbf{w}_{t+1})\| \\ 2\epsilon^2 + 8\gamma^2\|\hat{\delta}_t\|^2 &\geq \|\nabla\mathcal{L}(\mathbf{w}_{t+1})\|^2 \\ \frac{\epsilon^2}{4\gamma} + \gamma\|\hat{\delta}_t\|^2 &\geq \frac{\|\nabla\mathcal{L}(\mathbf{w}_{t+1})\|^2}{8\gamma}\end{aligned}$$

Therefore, we have:

$$\begin{aligned}
\sum_{t=1}^{T} \frac{\|\nabla \mathcal{L}(\mathbf{w}_{t+1})\|^2}{8\gamma} &\leq \mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star) + \frac{T\epsilon^2}{6H} + \frac{T\epsilon^2}{4\gamma} \\
\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_{t+1})\|^2 &\leq 8\gamma(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star)) + \frac{4T\gamma\epsilon^2}{3H} + 2T\epsilon^2 \\
\frac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_{t+1})\|^2 &\leq \frac{\gamma(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))}{T} + \frac{4\gamma\epsilon^2}{3H} + 2\epsilon^2 \\
&\leq \frac{\gamma(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))}{T} + 4\epsilon^2
\end{aligned}$$

Therefore, if we set $\epsilon = \sqrt{\frac{\gamma(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))}{T}}$, we have

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_{t+1})\|^2 \leq \frac{5\gamma(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))}{T}$$

However, the total number of iterations here is $M = NT$, so:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} \|\nabla \mathcal{L}(\mathbf{w}_{t+1})\|^2 &\leq \frac{5\gamma(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))}{M/N} \\
&\leq \frac{5\gamma(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))}{M} \frac{\log(\frac{H^2G^2}{\gamma^2\epsilon^2})}{C\sqrt{\frac{\gamma}{2H}}} \\
&\leq \frac{4\gamma(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))}{M} \frac{\log(\frac{H^2G^2T}{\gamma^3(\mathcal{L}(\mathbf{w}_1) - \mathcal{L}(\mathbf{w}_\star))})}{C\sqrt{\frac{\gamma}{2H}}} \\
&= O\left(\frac{\sqrt{\gamma H}\log(HGM/\gamma)}{M}\right)
\end{aligned}$$