CONVEX OPTIMIZATION

JOHN THICKSTUN

PROJECTED SUBGRADIENT DESCENT

Let $f \in \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$ with $\mathcal{X}$ compact (ML literature uses d instead of n). If f is convex ($\implies \mathcal{X}$ convex) then there is some global minimizer x^* with $f(x^*) \leq f(x)$ for all $x \in \mathcal{X}$. If differentiable then $\nabla f(x^*) = 0$.

How do we find x^*? Minimize a linearization of f (i.e. first order; i.e. gradient descent):

$$x_{t+1} \equiv x_t - \eta \nabla f(x_t).$$

How to set η? Too large and linear approximation is bad. Too small, slow progress.

Where to start? Any $x_1 \in \mathcal{X}$. In general it might be hard to find a feasible $x_1 \in \mathcal{X}$.

What if $x_{t+1} \notin \mathcal{X}$? Replace it with

$$\Pi_{\mathcal{X}}(x) \equiv \arg\min_{y \in \mathcal{X}} \|x - y\|.$$

What if f isn't differentiable? Use a subgradient.

Recall the first order convexity lower bound: $f(x) - f(y) \leq \nabla f(x)^T (x - y)$ (picture).

Definition. *(Subgradients) Let $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$. Then $g \in \partial f(x) \subset \mathbb{R}^n$ iff for all $y \in \mathcal{X}$,*

$$f(x) - f(y) \leq g^T (x - y).$$

Proposition. *Let $\mathcal{X} \subset \mathbb{R}^n$ be convex, $f : \mathcal{X} \to \mathbb{R}$; f is convex iff $\partial f(x) \neq \emptyset, \forall x \in \boldsymbol{int}\, \mathcal{X}$.*

Proof. See Proposition 1.1 in Bubeck. □

For insight, try proving the proposition if f is differentiable.

Projected subgradient descent:

$$y_{t+1} \equiv x_t - \eta g_t, \text{ where } g_t \in \partial f(x_t),$$
$$x_{t+1} \equiv \Pi_{\mathcal{X}} \left(y_{t+1} \right).$$

Definition. *A function $f : \mathcal{X} \subset (\mathbb{R}^n, \|\cdot\|) \to \mathbb{R}$ is L-Lipschitz iff $|f(x) - f(y)| \leq L\|x - y\|$ for all $x, y \in \mathcal{X}$.*

Exercise 1: Let $f : \mathcal{X} \subset (\mathbb{R}^n, \|\cdot\|) \to \mathbb{R}$ be convex. Show that f is L-Lipschitz iff for all $x \in \mathcal{X}$ and $g \in \partial f(x)$, $\|g\|_* \leq L$.

Exercise 2: Suppose $f : \mathcal{X} \subset (\mathbb{R}^n, \|\cdot\|_2) \to \mathbb{R}$ is convex and $\mathcal{X}$ is compact. Show that f is Lipschitz or give a counterexample.

How fast do we get to x^*? It depends on the learning rate. Try a constant rate. This will cause variance issues near opt (picture) so let's average our iterates.

Theorem. *Let $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$ be convex and L-Lipschitz with diameter of $\mathcal{X}$ bounded by R. The projected subgradient method with $\eta = \frac{R}{L\sqrt{t}}$ satisfies*

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq \frac{RL}{\sqrt{t}} = O\left(\frac{1}{\sqrt{t}}\right).$$

- This is optimal (section 3.5 of Bubeck) given only first-order information of f (delayed averaging, non-constant lr, anything else can't help).
- Subgradients and projections could be expensive; this analysis ignores that.
- Let $\epsilon > 0$, we achieve ϵ accuracy in $O(1/\epsilon^2)$ iterations (bad!).
- This accuracy is independent of the dimension n (good!).
- Learning rate depends on t (weird); same rate up to factor $\log t$ with rate $\frac{R}{L\sqrt{s}}$.
- If exercise 2 is true, then the hypotheses of the theorem are satisfied if $\mathcal{X}$ is compact.

Proof. By convexity of f,

$$f\left(\frac{1}{t}\sum_{s=1}^{t} x_s\right) - f(x^*) \leq \frac{1}{t}\sum_{s=1}^{t} f(x_s) - f(x^*) = \frac{1}{t}\sum_{s=1}^{t}(f(x_s) - f(x^*)).$$

The distance of an iterate to opt is given by

$$\begin{aligned}
f(x_s) - f(x^*) &\leq g_s^T(x_s - x^*) && \text{(definition of a subgradient)}\\
&= \frac{1}{\eta}(x_s - y_{s+1})^T(x_s - x^*) && \text{(definition of } g_s\text{)}\\
&= \frac{1}{2\eta}\left(\|x_s - y_{s+1}\|^2 + \|x_s - x^*\|^2 - \|y_{s+t} - x^*\|^2\right) && (2a^Tb = \|a\|^2 + \|b\|^2 - \|a-b\|^2)\\
&= \frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2\right) + \frac{\eta}{2}\|g_s\|^2. && \text{(definition of } g_s\text{).}
\end{aligned}$$

By geometry (see Lemma 3.1 in Bubeck)

$$\|x_{s+1} - x^*\| \leq \|y_{s+1} - x^*\|.$$

Because $\|g_s\| \leq L$ and $\|x_1 - x^*\| \leq R$,

$$\begin{aligned}
\frac{1}{t}\sum_{s=1}^{t}(f(x_s) - f(x^*)) &\leq \frac{1}{t}\sum_{s=1}^{t}\left(\frac{1}{2\eta}\left(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2\right) + \frac{\eta}{2}\|g_s\|^2\right)\\
&\leq \frac{1}{2t\eta}\sum_{s=1}^{t}\left(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2\right) + \frac{1}{t}\sum_{s=1}^{t}\frac{\eta}{2}L^2\\
&\leq \frac{1}{2t\eta}\left(\|x_1 - x^*\|^2 - \|x_t - x^*\|^2\right) + \frac{\eta L^2}{2} \leq \frac{R^2}{2t\eta} + \frac{\eta L^2}{2} = \frac{RL}{\sqrt{t}}.
\end{aligned}$$

□

SMOOTH FIRST-ORDER OPTIMIZATION

Definition. *A differentiable function f is β-smooth iff ∇f is β-Lipschitz.*

Exercise 3: Suppose f is twice-differentiable; f is β-smooth iff $\|\nabla^2 f(x)\|_2 \leq \beta, \forall x \in \mathcal{X}$.

Accuracy in non-smooth case was $O(1/\sqrt{t})$. Can we do better if f is smooth? Intuitively yes, because linearization is a better approximation. Specifically,

Lemma. *(Quadratic Upper Bound) Let $f : \mathbb{R}^n \to \mathbb{R}$ be β-smooth. For all $x, y \in \mathbb{R}^n$,*

$$|f(x) - f(y) - \nabla f(y)^T(x - y)| \leq \frac{\beta}{2}\|x - y\|^2.$$

Proof. See Lemma 3.4 in Bubeck. Converse is also true. □

Return to simple gradient descent (forget projections; see Bubeck if you care):

$$x_{t+1} \equiv x_t - \eta \nabla f(x_t).$$

Suppose $\eta = 1/\beta$; the quadratic upper bound helps us analyze the gradient step.

Corollary. *If f is β-smooth then*

$$f(x_{t+1}) - f(x_t) \leq -\frac{1}{2\beta}\|\nabla f(x_t)\|^2.$$

Proof. Plug in to the upper bound ($x \equiv x_t$):

$$\begin{aligned} f\left(x - \frac{1}{\beta}\nabla f(x)\right) - f(x) &\leq \nabla f(x)^T \left(x - \frac{1}{\beta}\nabla f(x) - x\right) + \frac{\beta}{2}\left\|x - \frac{1}{\beta}\nabla f(x) - x\right\|^2 \\ &= -\frac{1}{\beta}\|\nabla f(x)\|^2 + \frac{1}{2\beta}\|\nabla f(x)\|^2 = -\frac{1}{2\beta}\|\nabla f(x)\|^2. \end{aligned}$$

□

This implies that gradient descent on smooth functions is a *descent* method; i.e. the function value decreases with each iteration. This is not the case for non-smooth functions.

If f is β-smooth then ∇f will approximately vanish near x^*. The corollary implies that the gradient updates near opt will also vanish. So we'll know when we're getting close.

Because the gradient steps get small, we may also be able to avoid the averaging we used in the non-smooth case.

Theorem. *Let f be convex and β-smooth on $\mathbb{R}^n$. Gradient descent with $\eta = \frac{1}{\beta}$ satisfies*

$$f(x_t) - f(x^*) \leq \frac{\beta \|x_1 - x^*\|^2}{2t} = O\left(\frac{1}{t}\right).$$

- Use last point in the smooth case, versus averaging in non-smooth.
- Error drops like $1/t$ for smooth instead of $1/\sqrt{t}$ for non-smooth.
- We now achieve ϵ accuracy in $O(1/\epsilon)$ operations (better!).
- Not optimal: lower bound is $\Omega(1/\sqrt{\epsilon})$; room for acceleration.

Proof. Because this is a descent method (corollary) the last-point function value is bounded by the average function value of the iterates; i.e.

$$f(x_t) - f(x^*) \leq \frac{1}{t-1} \sum_{s=1}^{t-1} (f(x_{s+1}) - f(x^*)).$$

By the corollary and convexity (first-order condition)

$$f(x_{s+1}) \leq f(x_s) - \frac{1}{2\beta} \|\nabla f(x_s)\|^2 \leq f(x^*) + \nabla f(x_s)^T (x_s - x^*) - \frac{1}{2\beta} \|\nabla f(x_s)\|^2.$$

Completing the square, we have

$$f(x_{s+1}) - f(x^*) \leq \frac{\beta}{2} \left(\frac{2}{\beta} \nabla f(x_s)^T (x_s - x^*) - \| \frac{1}{\beta} \nabla f(x_s) \|^2 \right)$$

$$= \frac{\beta}{2} \left(\|x_s - x^*\|^2 - \|x_s - x^*\|^2 + \frac{2}{\beta} \nabla f(x_s)^T (x_s - x^*) - \| \frac{1}{\beta} \nabla f(x_s) \|^2 \right)$$

$$= \frac{\beta}{2} \left(\|x_s - x^*\|^2 - \|x_s - x^* - \frac{1}{\beta} \nabla f(x_s)\|^2 \right) = \frac{\beta}{2} \left(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2 \right).$$

And the sum telescopes:

$$f(x_t) - f(x^*) \leq \frac{\beta}{2(t-1)} \sum_{s=1}^{t-1} \left(\|x_s - x^*\|^2 - \|x_{s+1} - x^*\|^2 \right)$$

$$= \frac{\beta}{2(t-1)} \left(\|x_1 - x^*\|^2 - \|x_{t+1} - x^*\|^2 \right) \leq \frac{\beta \|x_0 - x^*\|^2}{2(t-1)}.$$

□

STRONG CONVEXITY

Definition. *A function f is α-strongly convex iff*

$$f(x) - f(y) \leq g^T(x-y) - \frac{\alpha}{2}\|x-y\|^2 \text{ for all } x, y, g \in \partial f(x).$$

Exercise 4: Suppose f is twice-differentiable; f is α-strongly convex iff

$$\nabla^2 f(x) \succeq \alpha I, \text{ for all } x \in \mathcal{X}.$$

I.e. f is bounded below by a quadratic with curvature α. This should help us; intuitively, if x is far from x^* then $\|\nabla f(x)\|$ will be large, so we will take big steps towards x^*.

Theorem. *Let $f : \mathcal{X} \subset \mathbb{R}^n \to \mathbb{R}$ be α-strongly convex and L-Lipschitz. The projected subgradient method with $\eta_s = \frac{2}{\alpha(s+1)}$ satisfies*

$$f\left(\sum_{s=1}^{t} \frac{2s}{t(t+1)} x_s\right) - f(x^*) \leq \frac{2L^2}{\alpha(t+1)} = O\left(\frac{1}{t}\right).$$

- Strong convexity lets us drop the R bound; we can quickly forget our initialization.
- Like β-smoothness, α-strong convexity upgrades our rate to $O(1/\epsilon)$.

Proof. Observe that $\sum_{s=1}^{t} \frac{2s}{t(t+1)} = 1$. By convexity of f,

$$f\left(\sum_{s=1}^{t} \frac{2s}{t(t+1)} x_s\right) - f(x^*) \leq \sum_{s=1}^{t} \frac{2s}{t(t+1)} \left(f(x_s) - f(x^*)\right).$$

And by strong convexity,

$$f(x_s) - f(x^*) \leq g_s^T(x_s - x^*) - \frac{\alpha}{2}\|x_s - x^*\|^2.$$

Using the same algebra as in the L-Lipschitz case,

$$\begin{aligned} f(x_s) - f(x^*) &\leq \frac{1}{2\eta_s}\left(\|x_s - x^*\|^2 - \|y_{s+1} - x^*\|^2\right) + \frac{\eta_s}{2}\|g_s\|^2 - \frac{\alpha}{2}\|x_s - x^*\|^2 \\ &\leq \frac{\eta_s}{2}L^2 + \left(\frac{1}{2\eta_s} - \frac{\alpha}{2}\right)\|x_s - x^*\|^2 - \frac{1}{2\eta_s}\|x_{s+1} - x^*\|^2 \\ &= \frac{1}{\alpha(s+1)}L^2 + \left(\frac{\alpha(s+1)}{4} - \frac{2\alpha}{4}\right)\|x_s - x^*\|^2 - \frac{\alpha(s+1)}{4}\|x_{s+1} - x^*\|^2 \\ &= \frac{1}{\alpha(s+1)}L^2 + \frac{\alpha}{4}\Big((s-1)\|x_s - x^*\|^2 - (s+1)\|x_{s+1} - x^*\|^2\Big). \end{aligned}$$

Note that $\frac{s}{s+1} \leq 1$ and summing the series we have

$$\frac{2}{t(t+1)}\sum_{s=1}^{t} s\left(f(x_s) - f(x^*)\right) \leq \frac{2L^2}{\alpha(t+1)} - \frac{2}{t(t+1)}\frac{\alpha s(s+1)\|x_{t+1} - x^*\|^2}{4}. \qquad \square$$

Finally, what happens if f is both smooth and strongly convex? Now we have both a quadratic lower bound and a quadratic upper bound; we know that the first-order approximation will be not so bad from the upper bound, and we know that we'll make good progress from the lower bound. The rate will be governed by the condition number of f.

Definition. *Let f be β-smooth and α-strongly convex. The condition number of f is $\kappa \equiv \frac{\beta}{\alpha}$.*

Theorem. *Let $f : \mathbb{R}^n \to \mathbb{R}$ be α-strongly convex and β-smooth. Then projected gradient descent with $\eta = \frac{1}{\beta}$ satisfies*

$$f(x_{t+1}) - f(x^*) \leq e^{-t/\kappa}\|x_1 - x^*\|^2 = O(e^{-t/\kappa}).$$

- Notice smoothness lets us to bound function value distance using iterate distance.
- Can achieve ϵ accuracy with $O(\kappa \log(1/\epsilon))$ iterations!

Proof. Recall that $\nabla f(x^*) = 0$ and therefore by β-smoothness

$$f(x_{t+1}) - f(x^*) \leq \frac{\beta}{2}\|x_{t+1} - x^*\|^2.$$

By definition of the gradient descent algorithm,

$$\|x_{t+1} - x^*\|^2 = \|x_t - \eta\nabla f(x_t) - x^*\|^2$$
$$= \|x_t - x^*\|^2 - \frac{2}{\beta}\nabla f(x_t)^T(x_t - x^*) + \frac{1}{\beta^2}\|\nabla f(x_t)\|^2.$$

Combining the smoothness corollary and strong convexity,

$$0 \leq f(x_{t+1}) - f(x^*) = f(x_t) - f(x^*) + f(x_{t+1}) - f(x_t)$$
$$\leq \nabla f(x_t)^T(x_t - x^*) - \frac{\alpha}{2}\|x_t - x^*\|^2 - \frac{1}{2\beta}\|\nabla f(x_t)\|^2.$$

And therefore

$$\|x_{t+1} - x^*\|^2 \leq \left(1 - \frac{\alpha}{\beta}\right)\|x_t - x^*\|^2$$
$$\leq \left(1 - \frac{\alpha}{\beta}\right)\|x_t - x^*\|^2 \leq \left(1 - \frac{\alpha}{\beta}\right)^t\|x_1 - x^*\|^2 \leq e^{-t/\kappa}\|x_1 - x^*\|^2.$$

□