**Gradient Descent Method**

# 1 Unconstrained Minimization

**Our focus today:** *Unconstrained minimization* problem: given a real-valued function $f$ over $\mathbb{R}^n$, find its minimum $x^*$ (assuming it exists). That is, solve the problem

$$x^* = \arg\min_{x \in \mathbb{R}^n} f(x).$$

- *Note:* This problem is *very* general:
  - To get maximization, just minimize $-f(x)$.
  - To introduce constraints, just consider minimizing $f(x)+\psi(x)$, where $\psi(x) = 0$, if $x$ satisfies all constraints, and $+\infty$, otherwise. (So, in principle, this is stronger than LP!)
- To make our discussion simpler, we will assume though that our function $f$ is "nice". That is, $f$ is:
  - continuous;
  - (twice) differentiable. (This requirement can, and often needs to, be relaxed.)

# 2 Gradient Descent

How to solve an unconstrained minimization problem?

- **Powerful approach:** Gradient descent method.
- **Key idea:** Apply (continuous) local greedy approach.
- Start with some point $x^0$.
- In each iteration: move a bit (locally) in the direction that reduces the value of $f$ the most (greedily).

  $\Rightarrow$ Guarantees that $f(x^{t+1}) < f(x^t)$.

**Question:** What is the direction of the steepest decrease of $f$?

- Recall (multi-variate) Taylor theorem: for any $x \in \mathbb{R}^n$ and (vector) displacement $\delta \in \mathbb{R}^n$, we have that

$$f(x+\delta) = f(x) + \nabla f(x)^T \delta + \frac{1}{2}\delta^T \nabla^2 f(y) \delta,$$

  for some $y = x + \lambda\delta$ with $0 \leq \lambda \leq 1$, where

- $\nabla f(x) \in \mathbb{R}^n$ is the *gradient* of $f$ at point $x$ and

$$\nabla f(x)_i := \frac{\partial f(x)}{\partial x_i},$$

for each $i$.

- $\nabla^2 f(x) \in \mathbb{R}^{n\times n}$ is the *Hessian* of $f$ at point $x$ and

$$\nabla^2 f(x)_{ij} := \frac{\partial^2 f(x)}{\partial x_i \partial x_j},$$

for each $i$ and $j$.

- *Observe:* the gradient term in the Taylor expansion is linear in $\|\delta\|$ while the Hessian term is quadratic in $\|\delta\|$.

- Consequently, for small enough step, i.e., $\|\delta\|$, the Hessian term is negligible. That is,

$$f(x+\delta) = f(x) + \nabla f(x)^T \delta + O(\|\delta\|^2) \approx f(x) + \nabla f(x)^T \delta$$

- **Key conclusion:** Even though $f$ might be very complex, locally it is "simple", i.e., it is well approximated by, essentially, the simplest function possible: the linear function!

  $\Rightarrow$ We know how to minimize linear functions. Just take $\delta = -\eta\nabla f(x)$, for some *step size* $\eta > 0$.

**Resulting algorithm:** *Gradient descent method*:

- Start with some $x^0 \in \mathbb{R}^n$.

- In each step $t$: $x^{t+1} \leftarrow x^t - \eta\nabla f(x^t)$.

**Question:** What should $\eta$ be?

- Assume that $f$ is *$\beta$-smooth*, for some $\beta > 0$. That is,

$$\|\nabla f(y) - \nabla(x)\| \leq \beta\|y - x\|,$$

  for any $x, y \in \mathbb{R}^n$. Intuitively, $\beta$ measures how much the gradient of $f$ can change between two nearby points.

- *Equivalently (for twice differentiable functions):* $f$ is $\beta$-smooth iff $y^T\nabla^2 f(x)y \leq \beta\|y\|^2$, for any $x, y$; or, put yet another way, the maximum eigenvalue of $\nabla^2 f(x)$ is at most $\beta$.

  $\Rightarrow$ We have that

$$f(x+\delta) \leq f(x) + \nabla f(x)^T \delta + \frac{\beta}{2}\|\delta\|^2,$$

for any $x$ and $\delta$

$\Rightarrow$ *Intuitively:* For every point $x$, there is a corresponding quadratic (i.e., relatively"simple") function that upper bounds $f$ *everywhere* and agrees with $f$ at the point $x$.

$\Rightarrow$ Our progress on minimizing this quadratic function at $x$ lowerbounds our progress on reducing the value of $f$ at $x$.

$\Rightarrow$ If we plug in our choice of $\delta = -\eta \nabla f(x)$, we get that

$$\begin{aligned} f(x+\delta) &\leq f(x) + \nabla f(x)^T \delta + \frac{\beta}{2}\|\delta\|^2 \\ &\leq f(x) - \eta\|\nabla f(x)\|^2 + \frac{\beta}{2}\eta^2\|\nabla f(x)\|^2 \\ &\leq f(x) - \frac{1}{2\beta}\|\nabla f(x)\|^2, \end{aligned}$$

for the optimal setting of $\eta = \frac{1}{\beta}$.

$\Rightarrow$ Setting $\eta = \frac{1}{\beta}$ ensures that

$$f(x^{t+1}) \leq f(x^t) - \frac{1}{2\beta}\|\nabla f(x)\|^2,$$

i.e., we make progress of at least $\frac{1}{2\beta}\|\nabla f(x)\|^2$ towards minimizing the value of $f$.

- In practice, we choose best $\eta$ adaptively in each step via binary search – this is often called *line search*.

**Remaining issue:** What if $\|\nabla f(x^k)\| = 0$ (or is just very small)?

- $x^k$ has to be a critical point – means $x^k$ is either a local minimum *or* maximum (with bad initialization) *or* a saddle point.
- If $\nabla^2 f(x^k) \succeq 0$, we know it is a local minimum.
- We can deal with the other two possibilities by perturbing our point slightly and resuming the algorithm.
- *In general:* Typically, gradient descent converges to *local* minimum.
- What if we want this local minimum to be a global one?
- We need additional (strong) assumption.
- $f$ is *convex* iff, for any $x$ and $y$,

$$f(\lambda x + (1-\lambda)y) \leq \lambda f(x) + (1-\lambda) f(y),$$

for any $0 \leq \lambda \leq 1$. That is, the epigraph of the function is a convex set.

- *Alternatively:* $f$ is convex iff $\nabla^2 f(x^k) \succeq 0$, for all $x$.

  $\Rightarrow$ The only critical points are local minimums!

- In fact, a *much* stronger property holds: all critical points are *global* minimums.

- To see that, note that by Taylor theorem convexity implies that

$$f(x+\delta) = f(x) + \nabla f(x)^T \delta + \frac{1}{2}\delta^T \nabla^2 f(x)\delta \geq f(x) + \nabla f(x)^T \delta.$$

  That is, every gradient defines a lowerbounding hyperplane for $f$ that agrees with $f$ at $x$.

  $\Rightarrow$ If $\nabla f(x) = 0$ then $f(x+\delta) \geq f(x)$ for *all* $\delta$.

- It turns out that convexity is a very widespread phenomena in optimization. But there are very important domains, e.g., deep learning, where the underlying optimization problems are inherently *non*-convex.

## 2.1 Convergence Analysis

How fast does gradient descent converge?

- Convexity allows us to bound our (sub-)optimality. Specifically, if $x^*$ is the minimum of $f$, we have that, for any $x$,

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x).$$

  $\Rightarrow$ $f(x) - f(x^*) \leq -\nabla f(x)^T (x^* - x) \leq \|\nabla f(x)\| \|x^* - x\|$, where the last inequality follows by Cauchy-Schwartz inequality.

  $\Rightarrow$ If $\|\nabla f(x)\| \leq \frac{\epsilon}{\|x^* - x\|}$, we are by at most $\epsilon$ off from optimum.

- The fact that the above near-optimality condition involves $\|x^* - x\|$ is unfortunate (but inherent!). After all, we don't know what this distance is.

- To connect this distance to the optimum to the norm of the gradient/difference in function value, and thus to get rid of this dependence, we need to make an (even stronger) assumption on $f$.

- Assume that $f$ is $\alpha$-*strong convexity.* That is, assume that, for any $x$ and $y$,

$$y^T \nabla^2 f(x) y \geq \alpha \|y\|^2.$$

  $\Rightarrow$ The smallest eigenvalue of $\nabla^2 f(x)$ is always at least $\alpha$.

  $\Rightarrow$ "Normal" convexity would correspond to $\alpha = 0$ (but we require $\alpha > 0$ here).

$\Rightarrow$ We can now strengthen our lowerbounding inequality we got from convexity. Specifically, for any $x$ and $\delta$ we have that

$$f(x+\delta) \geq f(x) + \nabla f(x)^T \delta + \frac{1}{2}\delta^T \nabla^2 f(x) \delta \geq f(x) + \nabla f(x)^T \delta + \frac{\alpha}{2}\|\delta\|^2.$$

That is, for each point $x$, there is a quadratic function that *lowerbounds* $f$ everywhere and agrees with $f$ at $x$.

- Now, the key consequence of $\alpha$-strong convexity we will need is that, for any $x$,

$$f(x^*) \geq f(x) + \nabla f(x)^T (x^* - x) + \frac{\alpha}{2}\|x - x^*\|^2.$$

And, as a result, by re-arranging, we get that

$$\nabla f(x)^T (x - x^*) \geq f(x) - f(x^*) + \frac{\alpha}{2}\|x - x^*\|^2. \tag{1}$$

- Now, to get the convergence bound, let us just put together everything we derived so far:
  - Let us use $\|x^t - x^*\|^2$ as a measure of our progress/potential.
  - Let's analyze its change in one step:

$$\begin{aligned}
\|x^{t+1} - x^*\|^2 &= \|x^t - \eta \nabla f(x^t) - x^*\|^2 \\
&= \|x^t - x^*\|^2 - 2\eta \nabla f(x^t)^T (x^t - x^*) + \eta^2 \|\nabla f(x^t)\|^2 \\
&\leq \|x^t - x^*\|^2 - \eta \left( 2(f(x^t) - f(x^*) + \frac{\alpha}{2}\|x^t - x^*\|^2) - \eta \|\nabla f(x^t)\|^2 \right),
\end{aligned}$$

    where the last line follows by (1).
  - Further, observe that as each gradient step guarantees making progress of at least $\frac{1}{2\beta}\|\nabla f(x^t)\|^2$ (whenever we set $\eta = \frac{1}{\beta}$, which we do here), it has to be that

$$f(x^t) - f(x^*) \geq f(x^t) - f(x^{t+1}) \geq \frac{1}{2\beta}\|\nabla f(x^t)\|^2$$

  - Plugging this back into our derivation and re-arranging, we obtain:

$$\begin{aligned}
\|x^{t+1} - x^*\|^2 &\leq \|x^t - x^*\|^2 - \eta \left( 2(f(x^t) - f(x^*) + \frac{\alpha}{2}\|x^t - x^*\|^2) - \eta \|\nabla f(x^t)\|^2 \right) \\
&\leq \|x^t - x^*\|^2 - \frac{1}{\beta}\left( \frac{1}{\beta}\|\nabla f(x^t)\|^2 + \alpha \|x^t - x^*\|^2 - \frac{1}{\beta}\|\nabla f(x^t)\|^2 \right) \\
&\leq \|x^t - x^*\|^2 - \frac{\alpha}{\beta}\|x^t - x^*\|^2 = \left(1 - \frac{1}{\kappa}\right)\|x^t - x^*\|^2,
\end{aligned}$$

    where $\kappa := \frac{\beta}{\alpha}$ is the *condition number* of $f$. (Intuitively, condition number tells us how "nicely" it behaves, i.e., how well can we "sandwich" the function $f$ locally by two quadratic functions. The smaller condition number the faster convergence.)

$\Rightarrow$ After $O(\kappa \log \frac{(f(x^0)-f(x^*)}{\epsilon})$ steps we obtain a solution that is within $\epsilon$ of the optimal value (in norm)!
*Note:* The dependence on $\epsilon$ is only logarithmic, which essentially allows us to solve the problem exactly by taking sufficiently large $\epsilon$.

## 3 Dealing with lack of $\alpha$-strongly convexity

- What to do if $f$ is *not* $\alpha$-strongly convex for any $\alpha > 0$? (This is often the case in applications.)
- A different analysis gives a (much weaker) convergence bound of $O(\frac{\beta\|x^*-x^0\|^2}{\varepsilon})$. (Here, the dependence on $\epsilon$ is polynomial, so in this regime we can only get approximate answers.)
- Alternatively, we could (almost, i.e., up to $O(\log \frac{1}{\varepsilon}$ factor) recover this weaker bound by *making* $f$ $\alpha$-strongly convex, with $\alpha = \frac{\varepsilon}{2\|x^*-x^0\|^2}$, by adding $\alpha\|x - x^0\|^2$ to it. (Note, we do not need to know $\|x^* - x^0\|^2$ exactly. Doing iterative doubling will suffice here.)
- This is an example of a more general technique called *regularization.*

  $\Rightarrow$ Adding this new term corresponding to adding $\alpha \cdot I$ to the Hessian $\nabla^2 f(x)$ of $f$. So, $f$ is indeed $\alpha$-strongly convex now and we can use the convergence analysis from above.

  $\Rightarrow$ *Problem:* The minimizer of $f$ changed! Still, one can show that the value attained at the new minimizer is withing $\frac{\varepsilon}{2}$ of the optimum. (Left as an exercise,)

## 4 Projections

- What to do if we want to solve *constrained* minimization? (E.g., max flow.)
- Just project (in $\ell_2$-norm) on the feasible space!
- The way we measured progress was by keeping track of $\|x^t - x^*\|^2$. But: an $\ell_2$-norm projection will never increase this quantity! Specifically, we have that if $\Pi(x)$ denotes the projected point $x$, we have that

  $$\|\Pi(x^t) - x^*\|^2 = \|\Pi(x^t) - \Pi(x^*)\|^2 \leq \|x^t - x^*\|^2,$$

  since the projection $\Pi$ is contractive.

  $\Rightarrow$ The analysis follows unchanged.

## 5 Dealing with Lack of $\beta$-Smoothness

- We can either use Subgradient descent, i.e., a variant of gradient descent that uses subgradients instead of gradients, or *smoothing*, a way to introduce a proxy objective function that is $\beta$-smooth while approximating the objective function well. (The latter is always preferable, as long as we can find a sufficiently good smoothening proxy.)

- For maximum flow, it is the best to smoothen the objective function $\|\cdot\|_\infty$ via *soft max* function:

$$\mathrm{smax}_\delta(x) := \delta \ln \left( \frac{\sum_{i=1}^n e^{\frac{x_i}{\delta}} + e^{\frac{-x_i}{\delta}}}{2n} \right), \tag{2}$$

  where $\delta >$ is a parameter.

- For every $\delta > 0$, the function $\mathrm{smax}_\delta$ is convex and $\frac{1}{\delta}$-smooth. (Exercise.)

- For any $x$ we have that, $\|x\|_\infty - \delta \ln(2n) \leq \mathrm{smax}_\delta(x) \leq \|x\|_\infty$. (Exercise)

- So, there is a trade-off between how well we approximate $\|\cdot\|_\infty$ and how smooth the resulting function is.

- Plugging the smoothened maximum flow formulation, with $\delta = \frac{\varepsilon}{2}$ into our bounds for gradient descent (with no $\alpha$-convexity), we get an $\epsilon$-approximate solution after

$$O\left(\frac{\beta\|x^0 - x^*\|^2}{\varepsilon}\right) = O\left(\frac{m}{\varepsilon^2}\right)$$

  iterations, where we use the fact that by choosing $x_0$ to be an all-zero vector (and then projecting it on the space of unit s-t flows) and noticing that optimal solution never flows more than 1 on any coordinate, $\|x^0 - x^*\|^2 \leq m$.

- As we can compute projections in nearly-linear time, the resulting algorithm runs in $\widetilde{O}(m^2\varepsilon^{-2})$ time.