CMO: Projected Gradient Descent

Eklavya Sharma

Projected gradient descent is an algorithm for convex constrained optimization that is similar to gradient descent. In this algorithm, we use gradient descent and if we ever move out of the feasible set, we project the point back to the feasible set.

Algorithm 1 `proj-grad-desc`(f, C, x_0): Minimize $f : \mathbb{R}^d \mapsto \mathbb{R}$ (in C^1, not necessarily convex) over the convex feasible set C. $x_0 \in C$ is the initial point.

$x^{(\min)} = x^{(0)}$
for t from 0 to ∞ **do**
  Choose step size α_t.
  $x^{(t+1)} = \text{proj}_C(x^{(t)} - \alpha_t \nabla_f(x^{(t)}))$
  $x^{(\min)} = \text{argmin}_{x \in \{x^{(\min)}, x^{(t+1)}\}} f(x)$
  if (stopping criterion) **then**
    return $x^{(\min)}$
  end if
end for

1 Finding projection: Example

Projected gradient descent requires a subroutine for finding projection. There is no easy general method for this. As an example we'll see how it's done for the constraint $Ax = b$, where A is an m by d matrix.

$$\text{proj}_{Ax=b}(z) = \underset{\substack{x \\ Ax=b}}{\text{argmin}} \frac{1}{2} \|x - z\|^2$$

Since $\|x - z\|^2$ is a convex function, a KKT point will give us the global minimum.

$$L(x, \mu) = \frac{1}{2} \|x - z\|^2 + \mu^T (Ax - b)$$

By stationarity, we get

$$(\nabla_x L)(x, \mu) = (x - z) + A^T \mu = 0 \implies x = z - A^T \mu$$

By primal feasibility, we get

$$b = Ax = A(z - A^T \mu) \implies \mu = (AA^T)^{-1}(Az - b)$$
$$\implies x = (I - A^T (AA^T)^{-1} A)z - A^T (AA^T)^{-1} b$$

The above x is $\text{proj}_{Ax=b}(z)$.

We can plug the above equation into the algorithm to get a simpler expression:

$$x^{(t+1)} = x^{(t)} - \alpha_t (I - A^T (AA^T)^{-1} A) \nabla_f(x^{(t)})$$

2 Convergence Analysis

Theorem 1 (Proved eariler)**.** *Let $f \in C^1_L$. Then $\forall x, y \in \mathbb{R}^d$*

$$f(y) \leq f(x) + \nabla_f(x)^T(y - x) + \frac{L}{2}\|y - x\|^2$$

Theorem 2 (Proved earlier)**.** *Let C be a convex set. Let $z \notin C$. Then $\forall x \in C$*

$$(\text{proj}_C(z) - z)^T(x - \text{proj}_C(z)) \geq 0$$

Let the objective function f in the projected gradient algorithm be C^1_L.

Theorem 3. *The projected gradient algorithm converges if we choose step size less than $\frac{2}{L}$ (but it may not converge to a local minimum).*

Proof.

$$\begin{aligned}
&x_{t+1} = \text{proj}_C(x_t - \alpha_t \nabla_f(x_t)) \\
&\implies (x_{t+1} - x_t + \alpha_t \nabla_f(x_t))^T(x_t - x_{t+1}) \geq 0 && \text{(by theorem 2)} \\
&\implies \nabla_f(x_t)^T(x_{t+1} - x_t) \leq -\frac{\|x_{t+1} - x_t\|^2}{\alpha_t} && (1)
\end{aligned}$$

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \nabla_f(x_t)^T(x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2 && \text{(by theorem 1)} \\
&\leq f(x_t) + \|x_{t+1} - x_t\|^2 \left(\frac{L}{2} - \frac{1}{\alpha_t}\right) && \text{(by equation (1))}
\end{aligned}$$

If we choose $\alpha_t < \frac{2}{L}$, then $f(x_{t+1}) < f(x_t)$. Assuming that f is lower-bounded, this means that the algorithm will converge. □

Theorem 4 (Proved earlier)**.** *Let f be C^1 and convex. Then*

$$\forall u, v \in \mathbb{R}^d, f(v) \geq f(u) + \nabla_f(u)^T(v - u)$$

Theorem 5. *When f is convex, x_{min} converges to a minimum if we choose step size less than $\frac{1}{L}$. Also, let x^* be a minimum of f, $E_T = \min_{0 \leq t \leq T}(f(x_t) - f(x^*))$ and $\alpha = \min_{t=0}^{T} \alpha_t$. Then*

$$E_T \leq \frac{\|x_0 - x^*\|^2}{2\alpha T}$$

Proof.

$$\begin{aligned}
&f(x^*) - f(x_t) \geq \nabla_f(x_t)^T(x^* - x_t) && \text{(by theorem 4)} \\
&\implies f(x_t) \leq f(x^*) + \nabla_f(x_t)^T(x_t - x^*)
\end{aligned}$$

$$\begin{aligned}
f(x_{t+1}) &\leq f(x_t) + \nabla_f(x_t)^T(x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2 && \text{(by theorem 1)} \\
&\leq \big(f(x^*) + \nabla_f(x_t)^T(x_t - x^*)\big) + \nabla_f(x_t)^T(x_{t+1} - x_t) + \frac{L}{2}\|x_{t+1} - x_t\|^2 \\
&\leq f(x^*) + \nabla_f(x_t)^T(x_{t+1} - x^*) + \frac{L}{2}\|x_{t+1} - x_t\|^2
\end{aligned}$$

$$
\begin{aligned}
& x_{t+1} = \operatorname{proj}_C(x_t - \alpha_t \nabla_f(x_t)) \\
& \implies (x_{t+1} - x_t + \alpha_t \nabla_f(x_t))^T (x^* - x_{t+1}) \geq 0 && \text{(by theorem 2)} \\
& \implies (x_{t+1} - x_t)^T (x_{t+1} - x^*) + \alpha_t \nabla_f(x_t)^T (x_{t+1} - x^*) \leq 0 \\
& \implies \nabla_f(x_t)^T (x_{t+1} - x^*) \leq -\frac{1}{\alpha_t} (x_{t+1} - x_t)^T (x_{t+1} - x^*)
\end{aligned}
$$

$$
\begin{aligned}
& f(x_{t+1}) - f(x^*) \\
& \leq \nabla_f(x_t)^T (x_{t+1} - x^*) + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
& \leq -\frac{1}{\alpha_t} (x_{t+1} - x_t)^T (x_{t+1} - x^*) + \frac{L}{2} \|x_{t+1} - x_t\|^2 \\
& = -\frac{1}{\alpha_t} (\delta_{t+1} - \delta_t)^T \delta_{t+1} + \frac{L}{2} \|\delta_{t+1} - \delta_t\|^2 && \text{(where } \delta_t = x_t - x^*\text{)} \\
& = -\frac{\|\delta_{t+1}\|^2}{\alpha_t} + \frac{L}{2} \|\delta_{t+1} - \delta_t\|^2 + \frac{\delta_t^T \delta_{t+1}}{\alpha_t} \\
& = -\frac{\|\delta_{t+1}\|^2}{\alpha_t} + \frac{L}{2} \|\delta_{t+1} - \delta_t\|^2 + \frac{\|\delta_{t+1}\|^2 + \|\delta_t\|^2 - \|\delta_{t+1} - \delta_t\|^2}{2\alpha_t} \\
& = \frac{\|\delta_t\|^2 - \|\delta_{t+1}\|^2}{2\alpha_t} + \frac{1}{2} \left(L - \frac{1}{\alpha_t} \right) \|\delta_{t+1} - \delta_t\|^2
\end{aligned}
$$

Let $E(x) = f(x) - f(x^*)$. If we always choose $\alpha_t < \frac{1}{L}$, then

$$
0 \leq E(x_{t+1}) < \frac{\|\delta_t\|^2 - \|\delta_{t+1}\|^2}{2\alpha_t} \implies \|\delta_t\| > \|\delta_{t+1}\|
$$

Let $E_T = \min_{0 \leq t \leq T} E(x_t)$ and $\alpha = \min_{t=0}^{T} \alpha_t$. Then

$$
\begin{aligned}
E_T &\leq \frac{1}{T} \sum_{t=0}^{T-1} E(x_{t+1}) \\
&< \frac{1}{T} \sum_{t=0}^{T-1} \frac{\|\delta_t\|^2 - \|\delta_{t+1}\|^2}{2\alpha_t} \\
&\leq \frac{1}{2\alpha T} \sum_{t=0}^{T-1} (\|\delta_t\|^2 - \|\delta_{t+1}\|^2) \\
&= \frac{1}{2\alpha T} (\|\delta_0\|^2 - \|\delta_T\|^2) \\
&\leq \frac{\|\delta_0\|^2}{2\alpha T}
\end{aligned}
$$

When $T \to \infty$, $E_T \to 0$. Since $E_T = E(x_{\min})$, $x_{\min}$ converges to a minimum. □