Gradient Descent with Weighted Inner Product

Problem

Our goal is to solve the optimization problem

$$\min_x f(x),$$

where $f\colon \mathbb{R}^n \to \mathbb{R}$ is a convex and differentiable function.

Let $P \in \mathbb{R}^{n\times n}$ be a symmetric positive definite matrix. This matrix induces a new (weighted) inner product defined as $\langle x, y\rangle_P = \langle Px, y\rangle$ which, in turn, induces a new gradient operator $\nabla_P f(x)$ with respect to this inner product. (Why does this make sense?)

This motivates us to consider a generalized version of gradient descent using the new gradient:

$$x_{k+1} = x_k - \alpha \nabla_P f(x_k), \tag{1}$$

where $\alpha > 0$ is the step size.

TODO:

- Find an explicit form of $\nabla_P f(x)$.
- Think of other ingredients/assumptions you need (such as the Lipschitzness of $\nabla_P f$) and prove the convergence of (1).
- Hint: you should understand well main ingredients of the standard proof of GD in the convex case and adjust them to your setting.

Base Statements

Lemma 2.28. If f is L–smooth and $\gamma > 0$, then for all $x, y \in \mathbb{R}^d$,

$$f(x - \gamma\nabla f(x)) - f(x) \le -\gamma\left(1 - \frac{\gamma L}{2}\right)\|\nabla f(x)\|^2. \tag{10}$$

If moreover $\inf f > -\infty$, then for all $x \in \mathbb{R}^d$,

$$\frac{1}{2L}\|\nabla f(x)\|^2 \le f(x) - \inf f.$$

General Proof of Convergence of Gradient Descent

Theorem Consider the Problem (Differentiable Function) and assume that f is convex and L-smooth, for some $L > 0$. Let $(x_t)_{t\in\mathbb{N}}$ be the sequence of iterates generated by the (GD) algorithm, with a stepsize satisfying $0 < \gamma \leq \frac{1}{L}$. Then, for all $x^* \in \arg\min f$, for all $t \in \mathbb{N}$, we have:

$$f(x_t) - \inf f \leq \frac{\|x_0 - x^*\|^2}{2\gamma t}.$$

Proof Let f be convex and L–smooth. It follows that

$$\begin{aligned}\|x_{t+1} - x^*\|^2 &= \left\|x_t - x^* - \frac{1}{L}\nabla f(x_t)\right\|^2 \\ &= \|x_t - x^*\|^2 - 2\cdot\frac{1}{L}\langle x_t - x^*, \nabla f(x_t)\rangle + \frac{1}{L^2}\|\nabla f(x_t)\|^2 \\ &\overset{(1)}{\leq} \|x_t - x^*\|^2 - \frac{1}{L^2}\|\nabla f(x_t)\|^2. \end{aligned} \tag{18}$$

Thus, $\|x_t - x^*\|^2$ is a decreasing sequence in t, and consequently

$$\|x_t - x^*\| \leq \|x_0 - x^*\|. \tag{19}$$

Calling upon (10) and subtracting $f(x^*)$ from both sides gives

$$f(x_{t+1}) - f(x^*) \leq f(x_t) - f(x^*) - \frac{1}{2L}\|\nabla f(x_t)\|^2. \tag{20}$$

Applying convexity we have that

$$\begin{aligned} f(x_t) - f(x^*) &\leq \langle \nabla f(x_t), x_t - x^*\rangle \\ &\leq \|\nabla f(x_t)\| \cdot \|x_t - x^*\| \\ &\overset{(19)}{\leq} \|\nabla f(x_t)\| \cdot \|x_0 - x^*\|. \end{aligned} \tag{21}$$

Suppose now that $x_0 \neq x^*$, otherwise the proof is finished. Isolating $\|\nabla f(x_t)\|$ in the above and inserting in (20) gives

$$f(x_{t+1}) - f(x^*) \overset{(20)+(21)}{\leq} f(x_t) - f(x^*) - \frac{1}{2L}\frac{1}{\|x_0 - x^*\|^2}(f(x_t) - f(x^*))^2 \tag{22}$$

Let $\beta = \frac{1}{2L}\frac{1}{\|x_0-x^*\|^2}$ and $\delta_t = f(x_t) - f(x^*)$. Since $\delta_{t+1} \leq \delta_t$, and by manipulating (22) we have that

$$\delta_{t+1} \leq \delta_t - \beta\delta_t^2 \xleftrightarrow{\times \frac{1}{\delta_t\delta_{t+1}}} \beta\frac{\delta_t}{\delta_{t+1}} \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t} \xleftrightarrow{\delta_{t+1}\leq\delta_t} \beta \leq \frac{1}{\delta_{t+1}} - \frac{1}{\delta_t}.$$

Summing up both sides over $t = 0, \ldots, T-1$ and using telescopic cancellation we have that

$$T\beta \leq \frac{1}{\delta_T} - \frac{1}{\delta_0} \leq \frac{1}{\delta_T}.$$

Re-arranging the above we have that

$$f(x^T) - f(x^*) = \delta_T \leq \frac{1}{\beta T} = \frac{2L\|x^0 - x^*\|^2}{T}.$$

Proof of Convergence of Gradient Descent with weighted inner product

Main Results

We consider gradient descent in a space equipped with a weighted inner product:

$$\langle x, y \rangle_P := \langle Px, y \rangle = x^\top P y,$$

where P is a symmetric positive definite matrix.

In this geometry, the gradient descent update takes the form:

$$x_{t+1} = x_t - \eta P^{-1} \nabla f(x_t),$$

where $\eta > 0$ is the learning rate and $\nabla f(x_t)$ is the usual Euclidean gradient.

This is equivalent to performing preconditioned gradient descent with preconditioner P^{-1}, which adapts the step direction to the local geometry defined by P.