

Friederike Horn & Bileam Scheuven

Justify all your claims.

Online Gradient Descent

Solution

a) The pythagorean inequality becomes apparent when we expand by \hat{y} :

$$\begin{aligned} \|y - u\|^2 &= \|y - \hat{y} + \hat{y} - u\|^2 \\ &= \|(y - \hat{y}) + (\hat{y} - u)\|^2 \end{aligned}$$

This way we can apply the parallelogram law $\|a + b\|^2 = \|a\|^2 + \|b\|^2 + 2\langle a, b \rangle$:

$$= \|y - \hat{y}\|^2 + \|\hat{y} - u\|^2 + 2\langle y - \hat{y}, \hat{y} - u \rangle$$

Since the $\langle a, b \rangle \geq 0$:

$$\geq \|y - \hat{y}\|^2 + \|\hat{y} - u\|^2$$

b) Consider the distance between the w_{t+1} and u :

$$\|w_{t+1} - u\|^2$$

We can expand w_{t+1} with our step rule before applying P_U

$$= \|w_t - \eta \nabla f_t(w_t) - u\|^2 - \epsilon$$

Where ϵ describes the nonnegative length lost in the projection. Without ϵ clearly:

$$\begin{aligned} \|w_{t+1} - u\|^2 &\leq \|w_t - \eta \nabla f_t(w_t) - u\|^2 \\ &= \|(w_t - u) + (-\eta \nabla f_t(w_t))\|^2 \\ &= \|(w_t - u)\|^2 + \|(-\eta \nabla f_t(w_t))\|^2 + 2\langle w_t - u, -\eta \nabla f_t(w_t) \rangle \\ &= \|(w_t - u)\|^2 + \eta^2 \|\nabla f_t(w_t)\|^2 - 2\eta \langle w_t - u, \nabla f_t(w_t) \rangle \end{aligned}$$

Remember the term we started with:

$$\begin{aligned} \|w_{t+1} - u\|^2 &\leq \|(w_t - u)\|^2 + \eta^2 \|\nabla f_t(w_t)\|^2 - 2\eta \langle w_t - u, \nabla f_t(w_t) \rangle \\ 2\eta \langle w_t - u, \nabla f_t(w_t) \rangle &\leq \|(w_t - u)\|^2 + \eta^2 \|\nabla f_t(w_t)\|^2 - \|w_{t+1} - u\|^2 \end{aligned}$$

Divide by 2η :

$$\langle w_t - u, \nabla f_t(w_t) \rangle \leq \frac{\|(w_t - u)\|^2 - \|w_{t+1} - u\|^2}{2\eta} + \frac{\eta}{2} \|\nabla f_t(w_t)\|^2$$

- c) From convexity we know that a (differentiable) function stays above any linear local approximation, that is to say for x, y in the domain of f :

$$\begin{aligned} f(x) &\geq f(y) + \langle \nabla f(y), x - y \rangle \\ \Leftrightarrow -\langle \nabla f(y), x - y \rangle &\geq f(y) - f(x) \\ \Leftrightarrow \langle \nabla f(y), y - x \rangle &\geq f(y) - f(x) \end{aligned}$$

Translating to our problem, we obtain:

$$f_t(w_t) - f_t(u) \leq \langle \nabla f_t(w_t), w_t - u \rangle$$

We can replace the right side with our inequality from b):

$$f_t(w_t) - f_t(u) \leq \frac{\|(w_t - u)\|^2 - \|w_{t+1} - u\|^2}{2\eta} + \frac{\eta}{2} \|\nabla f_t(w_t)\|^2$$

The sum $\sum_{t=1}^T f_t(w_t) - f_t(u)$ can then be decomposed to examine both sums on the right hand side individually:

$$\sum_{t=1}^T \frac{\|(w_t - u)\|^2 - \|w_{t+1} - u\|^2}{2\eta} = \frac{1}{2\eta} \sum_{t=1}^T (\|(w_t - u)\|^2 - \|w_{t+1} - u\|^2)$$

It can be seen that this sum is telescopic, that is, it expands to:

$$\frac{1}{2\eta} (\|w_1 - u\|^2 - \|w_2 - u\|^2 + \|w_2 - u\|^2 - \|w_3 - u\|^2 + \|w_3 - u\|^2 + \dots)$$

leaving only the first and last term.

$$= \frac{\|w_1 - u\|^2 - \|w_{T+1} - u\|^2}{2\eta}$$

Since $w_1 = 0$, $\|u\|^2 \leq D$ we obtain:

$$\begin{aligned} &\leq \frac{D^2 - \|w_{T+1} - u\|^2}{2\eta} \\ &\leq \frac{D^2}{2\eta} \end{aligned}$$

Returning to the original expression and examining the other summand:

$$\frac{\eta}{2} \sum_{t=1}^T \|\nabla f_t(w_t)\|^2$$

Since that $\|\nabla f_t\| \leq G$:

$$\begin{aligned} &\leq \frac{\eta}{2} \sum_{t=1}^T G^2 \\ &= \frac{\eta}{2} T G^2 \end{aligned}$$

This concludes the proof that:

$$\max_{u \in U} \sum_{t=1}^T f_t(w_t) - f_t(u) \leq \frac{D^2}{2\eta} + \frac{\eta}{2} T G^2$$

d) To find the optimal η we investigate the upper bound for critical points.

$$\frac{d}{d\eta} \left(\frac{D^2}{2\eta} + \frac{\eta}{2} TG^2 \right) \stackrel{!}{=} 0$$

$$\frac{-D^2}{2\eta^2} + \frac{TG^2}{2} = 0$$

$$\frac{TG^2}{2} = \frac{D^2}{2\eta^2}$$

$$\eta^2 TG^2 = D^2$$

$$\eta^2 = \frac{D^2}{TG^2}$$

$$\eta = \frac{D}{\sqrt{TG}}$$

This is a minimum, since the second derivative is $\frac{D^2}{6\eta^3}$, which is positive.

Inserting this into the original equation yields a worst case regret of:

$$\begin{aligned} & \frac{D^2}{2 \left(\frac{D}{\sqrt{TG}} \right)} + \frac{\left(\frac{D}{\sqrt{TG}} \right)}{2} TG^2 \\ &= \frac{D^2 \sqrt{TG}}{2D} + \frac{D}{2\sqrt{TG}} TG^2 \\ &= \frac{D\sqrt{TG}}{2} + \frac{D\sqrt{TG}}{2} \\ &= D\sqrt{TG} \end{aligned}$$

Formal proof of the existence of Lagrange Multipliers

Solution

a)

b)

c)

Differentiable approximation of \mathcal{L}^1 -approximation

Solution

a)

b)