Friederike Horn & Bileam Scheuvens

Justify all your claims.

Online Gradient Descent

Solution

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

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

$$= ||(y-\hat{y})+(\hat{y}-u)||^2$$

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:

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

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

$$= ||(w_t-u)||^2 + ||(-\eta\nabla f_t(w_t))||^2 + 2\langle w_t-u, -\eta\nabla f_t(w_t)\rangle$$

$$= ||(w_t-u)||^2 + \eta^2||\nabla f_t(w_t)||^2 - 2\eta\langle w_t-u, \nabla f_t(w_t)\rangle$$

Remember the term we started with:

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

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

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:

$$f(x) \geq f(y) + \langle \nabla f(y), x - y \rangle$$

$$\Leftrightarrow -\langle \nabla f(y), x - y \rangle \geq f(y) - f(x)$$

$$\Leftrightarrow \langle \nabla f(y), y - x \rangle \geq f(y) - f(x)$$

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:

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

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:

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

$$\leq \frac{D^2}{2\eta}$$

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 $\nabla |f_t| \leq G$:

$$\leq \frac{\eta}{2} \sum_{t+1}^{T} G^2$$

$$= \frac{\eta}{2} T G^2$$

This concludes the proof that:

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

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

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

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:

$$\frac{D^2}{2\left(\frac{D}{\sqrt{T}G}\right)}+\frac{\left(\frac{D}{\sqrt{T}G}\right)}{2}TG^2$$
$$=\frac{D^2\sqrt{T}G}{2D}+\frac{D}{2\sqrt{T}G}TG^2$$
$$=\frac{D\sqrt{T}G}{2}+\frac{D\sqrt{T}G}{2}$$
$$=D\sqrt{T}G$$

Formal proof of the existence of Lagrange Multipliers

Let $f, g\mathbb{R}^d \to \mathbb{R}$.

Solution

We want to show that the problem $\min_x f(x)$ under the constraint $g(x)=0$ is solved by $\nabla f(x)+\nu g(x)=0$. For this we first introduce the parametrisation function $h(t)$, such that the image of h is $\{x|g(x)=0\}$. For this function we know that $g(h(t))=0\forall t$. In particular, this means that it is a constant function and thererfore by definition $\frac{dg(h(t))}{dt}=0$. Writing this out with the chain rule we obtain:

$$\nabla g(h(t))Dh(t)=0.$$

On the other hand $f(h(t))$ also fulfills the condition that it only is defined over all x for which $g(x)=0$. Taking the derivative and setting to zero to find the minimum we find:

$$\nabla f(h(t*))Dh(t*)=0,$$

with $h(t*)=x*$, the minimal point of the optimization problem. This means that $\nabla f(h(t))$ and as $\nabla g(h(t))$ are perpendicular to $range(Dh(t))$. But we can assume that $Dh(t)$ has full rank the space perpendicular to $Dh(p)$ has only dimension $d-(d-1)=1$ and therefore $\nabla g(h(t))=\nu\nabla f(h(t*))$ and thus the proposition:

$$\nabla f(x*)+\nu g(x*)=0$$

holds true.

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

Solution

a) Given:

$$\underset{x,y}{min}\ ||y||_1$$

$$\text{Subject to } y - Ax + b = 0$$

We can rearrange into the Lagrangian:

$$\mathcal{L}(y,x,y) = \underset{x,y}{min}\ ||y||_1 + \lambda(y - Ax + b)$$

The dual is then:

$$g(\lambda) = \underset{y,x}{inf}\ \mathcal{L}(y,x,\lambda)$$

To obtain our constrained version we can differentiate this wrt x and y. Starting with y:

$$\underset{y}{inf}\ ||y||_1 + \lambda^t y$$

It can be seen that this is lower bounded by 0 if the all entries of λ are in $[-1,1]$, as any negative terms are canceled out by the norm. Without this restriction, however, it is unbounded. Therefore:

$$= \begin{cases} 0 & \text{if} |\lambda_i| \leq 1\ \forall i \\ -\infty & otherwise \end{cases}$$

Differentiating wrt x yields:

$$\underset{x}{inf}\ -\lambda^t Ax$$

For dual feasibilty we require $\lambda A = 0$, yielding our second constraint. Substituting this for the infimum into the dual problem leaves:

$$g(\lambda) = \underset{\lambda}{max}\ \lambda^t b$$

$$\text{subject to } |\lambda_i| \leq 1\ \forall i,\ \lambda^t A = 0$$

b)

c)