CMPT 409/981: Optimization for Machine Learning

Lecture 3

Sharan Vaswani

September 12, 2024

Recap

- For an L-smooth function, $f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2} \|y - x\|^2$ for all $x, y \in \mathcal{D}$.
- For L-smooth functions lower-bounded by f^*, gradient descent with $\eta = \frac{1}{L}$ returns $\hat{w}$ such that $\|\nabla f(\hat{w})\|^2 \leq \epsilon$ and requires $T \geq \frac{2L\,[f(w_0) - f^*]}{\epsilon}$ iterations (oracle calls).
- Importantly, the GD rate does not depend on the dimension of w.
- *Lower-Bound*: When minimizing a smooth function (without additional assumptions), any *first-order* algorithm requires $\Omega\left(\frac{1}{\epsilon}\right)$ oracle calls to return a point $\hat{w}$ such that $\|\nabla f(\hat{w})\|^2 \leq \epsilon$.
- Hence, GD is optimal for minimizing smooth functions.

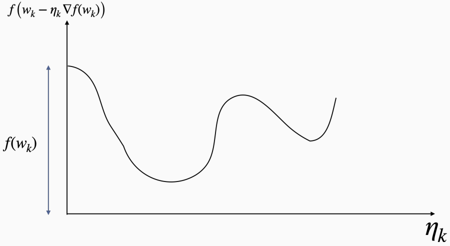
Gradient Descent

- The above results require setting the step-size to $\frac{1}{L}$. In fact, GD with any $\eta \in \left(0, \frac{2}{L}\right)$ will result in convergence to the stationary point (prove in Assignment 1).
- However, estimating L can be difficult as the functions get more complicated.
- Even for simple functions, the theoretically computed L is global (the "local" L might be much smaller) and often loose in practice. Typically we tend to overestimate L resulting in a smaller step-size.
- Instead of setting η according to L, we can "search" for a good step-size η_k in each iteration k. We will study 2 ways to do so:
    - Exact Line-search
    - Backtracking Armijo Line-search

Exact Line-search

Exact line-search: At iteration k, solve the following sub-problem:

$$\eta_k = \arg\min_{\eta} f(w_k - \eta \nabla f(w_k)).$$



After computing η_k, do the usual GD update: $w_{k+1} = w_k - \eta_k \nabla f(w_k)$.

- Can adapt to the "local" L, resulting in larger step-sizes and better performance.
- Can solve the sub-problem approximately by doing gradient descent w.r.t η (known as *hyper-gradient descent* [BCR+17]). This is computationally expensive.
- Can compute η_k analytically. This can only be done in special cases such as for quadratics.

Exact Line-search for Linear Regression

Recall linear regression: for $X \in \mathbb{R}^{n \times d}$ and $y \in \mathbb{R}^n$, we aim to solve:
$\min_{w \in \mathbb{R}^d} f(w) := \frac{1}{2} \|Xw - y\|^2 = \frac{1}{2} \left[w^\mathsf{T}(X^\mathsf{T}X)w - 2\langle X^\mathsf{T}y, w\rangle + \|y\|^2 \right]$.

For the exact line-search, we need to $\min_\eta h(\eta) := f(w_k - \eta \nabla f(w_k))$.

Since f is a quadratic, we can directly use the second-order Taylor series:

$$
\begin{aligned}
f(w_k - \eta \nabla f(w_k)) &= f(w_k) + \langle \nabla f(w_k), -\eta \nabla f(w_k)\rangle + \frac{1}{2}[-\eta \nabla f(w_k)]^\mathsf{T} \nabla^2 f(w_k)[-\eta \nabla f(w_k)] \\
\implies \nabla h(\eta_k) &= -\|\nabla f(w_k)\|^2 + \eta [\nabla f(w_k)]^\mathsf{T} \nabla^2 f(w_k)[\nabla f(w_k)] = 0 \\
\implies \eta_k &= \frac{\|\nabla f(w_k)\|^2}{\|\nabla f(w_k)\|^2_{\nabla^2 f(w_k)}}
\end{aligned}
$$

For linear regression, $\nabla^2 f(w_k) = X^\mathsf{T}X$ and $\nabla f(w_k) = X^\mathsf{T}(Xw_k - y)$.
$\implies \eta_k = \frac{\|X^\mathsf{T}(Xw_k - y)\|^2}{\|X^\mathsf{T}(Xw_k - y)\|^2_{X^\mathsf{T}X}}$. (Implement in Assignment 1)

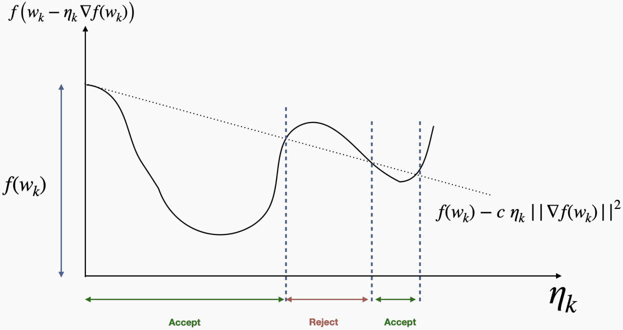
Armijo Condition

Usually, the cost of doing an exact line-search is not worth the computational effort.

Armijo condition for a prospective step-size $\tilde{\eta}_k$:

$$f(w_k - \tilde{\eta}_k \nabla f(w_k)) \leq f(w_k) - c\,\tilde{\eta}_k \,\|\nabla f(w_k)\|^2$$

where $c \in (0, 1)$ is a hyper-parameter.



Gradient Descent with Backtracking Armijo Line-search

Algorithm GD with Armijo Line-search

function GD with Armijo line-search(f, w_0, $\eta_{\max}$, $c \in (0,1)$, $\beta \in (0,1)$)
for $k = 0, \ldots, T-1$ **do**
  $\tilde{\eta}_k \leftarrow \eta_{\max}$
  while $f(w_k - \tilde{\eta}_k \nabla f(w_k)) > f(w_k) - c \cdot \tilde{\eta}_k \left\| \nabla f(w_k) \right\|^2$ **do**
    $\tilde{\eta}_k \leftarrow \tilde{\eta}_k \beta$
  end while
  $\eta_k \leftarrow \tilde{\eta}_k$
  $w_{k+1} = w_k - \eta_k \nabla f(w_k)$
end for
return w_T

Backtracking Armijo Line-search

Simplification for analysis: Assume that the backtracking line-search procedure returns the largest η that satisfies the Armijo condition. Will be referred to as *exact backtracking line-search*.

Claim: For L-smooth functions, the exact backtracking line-search procedure terminates and returns $\eta_k \geq \min\left\{\frac{2\,(1-c)}{L}, \eta_{\max}\right\}$.

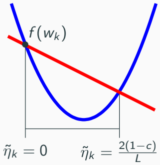
Proof: For a prospective step-size $\tilde{\eta}_k$, we will use the following two inequalities:

$$f(w_k - \tilde{\eta}_k \nabla f(w_k)) \leq \underbrace{f(w_k) - \|\nabla f(w_k)\|^2 \left(\tilde{\eta}_k - \frac{L\tilde{\eta}_k^2}{2}\right)}_{h_1(\tilde{\eta}_k)} \quad \text{(Quadratic bound using smoothness)}$$

$$f(w_k - \tilde{\eta}_k \nabla f(w_k)) \leq \underbrace{f(w_k) - \|\nabla f(w_k)\|^2 \,(c\tilde{\eta}_k)}_{h_2(\tilde{\eta}_k)} \quad \text{(Armijo condition)}$$

Backtracking Armijo Line-search

If the Armijo condition is satisfied, the backtracking line-search procedure terminates.



Smoothness:
$h_1(\tilde{\eta}_k) = f(w_k) + (\frac{L}{2}\tilde{\eta}_k^2 - \tilde{\eta}_k)\|\nabla f(w_k)\|^2$

Line search:
$h_2(\tilde{\eta}_k) = f(w_k) - c\tilde{\eta}_k\|\nabla f(w_k)\|^2$

Case (i) $\eta_{\max} \leq \frac{2(1-c)}{L}$: From smoothness, $f(w_k - \eta_{\max}\nabla f(w_k)) \leq h_1(\eta_{\max})$. For $\eta_{\max} \leq \frac{2(1-c)}{L}$, we know that $h_1(\eta_{\max}) \leq h_2(\eta_{\max})$. Hence, $f(w_k - \eta_{\max}\nabla f(w_k)) \leq h_2(\eta_{\max})$, meaning that the Armijo condition is satisfied for $\eta_{\max}$. $\implies$ if $\eta_{\max} \leq \frac{2(1-c)}{L}$, then the line-search terminates immediately and $\eta_k = \eta_{\max}$.

Case (ii): If $\eta_{\max} > \frac{2\,(1-c)}{L}$: While backtracking, if $\tilde{\eta}_k = \frac{2\,(1-c)}{L}$, then $f(w_k - \eta_{\max}\nabla f(w_k)) \leq h_1(\tilde{\eta}_k) = h_2(\tilde{\eta}_k)$, the line-search terminates immediately and $\eta_k = \frac{2\,(1-c)}{L}$. If the Armijo condition is satisfied for a step-size η_k s.t. $h_2(\eta_k) < h_1(\eta_k)$, then $f(w_k - \eta_k\nabla f(w_k)) \leq h_2(\eta_k) < h_1(\eta_k) \implies c\eta_k \geq \eta_k - \frac{L\eta_k^2}{2} \implies \eta_k \geq \frac{2(1-c)}{L}$.

Putting everything together, the step-size η_k returned by the Armijo line-search satisfies $\eta_k \geq \min\left\{\frac{2\,(1-c)}{L}, \eta_{\max}\right\}$.

Gradient Descent with Backtracking Armijo Line-search

Claim: For L-smooth functions lower-bounded by f^*, gradient descent with exact backtracking Armijo line-search (with $c = 1/2$) returns point $\hat{w}$ such that $\|\nabla f(\hat{w})\|^2 \leq \epsilon$ and requires $T \geq \frac{\max\{2L, 2/\eta_{\max}\} \, [f(w_0) - \min_w f(w)]}{\epsilon}$ iterations.

Proof: Since η_k satisfies the Armijo condition and $w_{k+1} = w_k - \eta_k \nabla f(w_k)$,

$$
\begin{aligned}
f(w_{k+1}) &\leq f(w_k) - c\,\eta_k \, \|\nabla f(w_k)\|^2 \\
&\leq f(w_k) - \left(\min\left\{\frac{1}{2L}, \frac{\eta_{\max}}{2}\right\}\right) \|\nabla f(w_k)\|^2
\end{aligned}
$$

(Result from previous slide with $c = 1/2$)

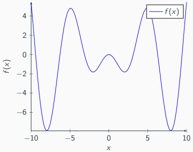
Continuing the proof as before,

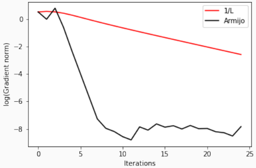
$$
\implies \|\nabla f(\hat{w})\|^2 \leq \frac{\max\{2L, 2/\eta_{\max}\} \, [f(w_0) - \min_w f(w)]}{T}
$$

The claim can be proved by the same reasoning as in Lecture 2.

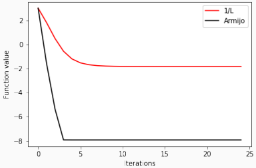
Gradient Descent with Backtracking Armijo Line-search – Example

$\min_{x\in[-10,10]} f(x) := -x\sin(x)$. Compare GD (with $x_0 = 4$) with (i) $\eta = 1/L \approx 0.1$ and (ii) Armijo line-search ($\eta_{\max} = 10, c = 1/2, \beta = 0.9$).

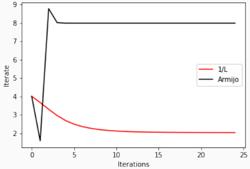




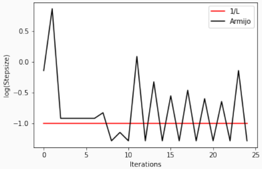
(a) Gradient norm



(b) Function value



(c) Iterate



(d) Stepsize

Questions?

Convex Optimization

For smooth functions, GD requires $\Theta(1/\epsilon)$ iterations to converge to an ϵ-approximate stationary point. Alternatively, if we care about global optimization (reach the vicinity of the true minimizer), any algorithm requires $\Theta(1/\epsilon^d)$ iterations.

Convex functions: Class of functions where local optimization can result in convergence to the global minimizer of the function.

In general, convex optimization involves minimizing a convex function over a convex set $\mathcal{C}$.

Examples of convex optimization in ML

Ridge regression: $\min_{w \in \mathbb{R}^d} \frac{1}{2} \|Xw - y\|^2 + \frac{\lambda}{2} \|w\|^2$.

Logistic regression: $\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \langle X_i, w \rangle\right)\right)$

Support vector machines: $\min_{w \in \mathbb{R}^d} \sum_{i=1}^{n} \max\left\{0, 1 - y_i \langle X_i, w \rangle\right\} + \frac{\lambda}{2} \|w\|^2$

Planning in MDPs in RL: $\max_{\mu \in \mathcal{F}_\rho} \langle \mu, r \rangle$ where $\mathcal{F}_\rho$ is the flow-polytope.

Convex Sets

A set $\mathcal{C}$ is convex if every point along the line joining two points in $\mathcal{C}$ also lies in the set.

For points x, y, the *convex combination* of x, y is $z_\theta := \theta x + (1-\theta)y$ for $\theta \in [0, 1]$.

A set $\mathcal{C}$ is convex iff $\forall x, y \in \mathcal{C}$, the convex combination $z_\theta \in \mathcal{C}$ for all $\theta \in [0, 1]$.

Examples of convex sets:

- Positive orthant $\mathbb{R}^d_+ : \{x | x \geq 0\}$.
- Hyper-plane: $\{x | Ax = b\}$.
- Half-space: $\{x | Ax \leq b\}$.
- Norm-ball: $\{x | \, \|x\|_p \leq r\}$ for $p \geq 1$.
- Norm-cone: $\{(x, r) | \, \|x\|_p \leq r\}$ for $p \geq 1$.

Convex Sets

Q: Prove that the hyper-plane (set of linear equations): $\mathcal{H} := \{x | Ax = b\}$ is a convex set.

If $x, y \in \mathcal{H}$, then, $Ax = b$ and $Ay = b$. Consider a point $z_\theta := \theta x + (1-\theta)y$ for $\theta \in [0,1]$.

$$Az_\theta = A[\theta x + (1-\theta)y] = \theta Ax + (1-\theta)Ay = b.$$

Hence, $z_\theta \in \mathcal{H}$ for all $\theta \in [0,1]$ and $\mathcal{H}$ is a convex set.

Q: Prove that the ball of radius r centered at point x_c: $\mathcal{B}(x_c, r) := \{x | \, \|x - x_c\|_p \leq r\}$ for $p \geq 1$ is convex.

If $x, y \in \mathcal{B}(x_c, r)$, then, $\|x - x_c\|_p \leq r$ and $\|y - x_c\|_p \leq r$. Consider a point $z_\theta := \theta x + (1-\theta)y$ for $\theta \in [0,1]$.

$$\begin{aligned}
\|z_\theta - x_c\|_p &= \|\theta(x - x_c) + (1-\theta)(y - x_c)\|_p \\
&\leq \|\theta(x - x_c)\|_p + \|(1-\theta)(y - x_c)\|_p && \text{(Triangle inequality for norms)} \\
&\leq \theta \|(x - x_c)\|_p + (1-\theta)\|(y - x_c)\|_p && \text{(Homogeneity of norms)}
\end{aligned}$$

$$\implies \|z - x_c\|_p \leq r$$

Hence, $z_\theta \in \mathcal{B}(x_c, r)$ for all $\theta \in [0,1]$ and $\mathcal{B}(x_c, r)$ is a convex set.

Convex Sets

Q: Prove that the set of symmetric PSD matrices: $S_+^n = \{X \in \mathbb{R}^{n \times n} | X \succeq 0, X = X^\intercal\}$ is convex.

• Intersection of convex sets is convex $\implies$ can prove the convexity of a set by showing that it is an intersection of convex sets.

Example: We know that a half-space: $\langle a_i, x \rangle \leq b_i$ is a convex set. The set of inequalities $Ax \leq b$ is an intersection of half-spaces and is hence convex.

Questions?

Convex Functions

Zero-order definition: A function f is convex iff its domain $\mathcal{D}$ is a convex set, and for all $x, y \in \mathcal{D}$ and $\theta \in [0, 1]$,

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

i.e. the function is below the chord between two points.

- Alternatively, f is convex iff the set formed by the area above the function is a convex set.

Examples of convex functions:

- All p-norms $\|x\|_p$ with $p \geq 1$.
- $f(x) = 1/\sqrt{x}$, $f(x) = -\log(x)$, $f(x) = \exp(-x)$
- Negative entropy: $f(x) = x\log(x)$
- Logistic loss: $f(x) = \log(1 + \exp(-x))$
- Linear functions $f(x) = \langle a, x \rangle$

Convex Functions

First-order condition: If f is differentiable, it is convex iff its domain $\mathcal{D}$ is a convex set and for all $x, y \in \mathcal{D}$,

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

i.e. the function is above the tangent to the function at any point x.

For a convex f, consider w^* such that $\nabla f(w^*) = 0$, then using convexity, for all $y \in \mathcal{D}$, $f(y) \geq f(w^*)$. If w^* is a stationary point i.e. $\|\nabla f(w^*)\|^2 = 0$, then it is a global minimum. Hence, local optimization to make the gradient zero results in convergence to a global minimum!

Q: For a convex f, if $\nabla f(w^*) = 0$, then is w^* a unique minimizer of f?

Second-order condition: If f is twice differentiable, it is convex iff its domain $\mathcal{D}$ is a convex set and for all $x \in \mathcal{D}$,

$$\nabla^2 f(x) \succeq 0$$

i.e. the Hessian is positive semi-definite ("curved upwards") for all x.

Convex Functions

Q: Prove that $f(x) = \max_i x_i$ is a convex function.

$$f\left(\theta x + (1-\theta)y\right) = \max_i[\theta x_i + (1-\theta)y_i] \leq \theta \max_i x_i + (1-\theta)\max_i y_i = \theta f(x) + (1-\theta)f(y)$$

Hence, by using the zero-order definition of convexity, $f(x)$ is convex.

Q: Prove that $f(x) = \frac{1}{2}x^2$ is a convex function.

$$f(y) - f(x) - \langle \nabla f(x), y - x\rangle = \frac{y^2}{2} - \frac{x^2}{2} - x(y-x) = \frac{1}{2}\left[y^2 + x^2 - 2xy\right] = \frac{(x-y)^2}{2} \geq 0$$

Hence, by using the first-order definition of convexity, $f(x)$ is convex.

Convex Functions

Q: Prove that $f(x) = \log(1 + \exp(-x))$ is a convex function.

$$f'(x) = \frac{-\exp(-x)}{1 + \exp(-x)} = \frac{-1}{1 + \exp(x)}$$
$$f''(x) = \frac{\exp(x)}{(1 + \exp(x))^2} > 0$$

Hence, by using the second-order definition of convexity, $f(x)$ is convex.

Q: Prove that the ridge regression loss function: $f(w) = \frac{1}{2}\left\|Xw - y\right\|^2 + \frac{\lambda}{2}\left\|w\right\|^2$ is convex
Recall that $\nabla^2 f(w) = X^\intercal X + \lambda I_d$. For vector v, let us consider $v^\intercal \nabla^2 f(w) v$,

$$v^\intercal \nabla^2 f(w) v = v^\intercal [X^\intercal X + \lambda I_d] v = v^\intercal [X^\intercal X] v + \lambda v^\intercal v = [Xv]^\intercal [Xv] + \lambda \left\|v\right\|^2 = \left\|Xv\right\|^2 + \lambda \left\|v\right\|^2$$
$$\implies v^\intercal \nabla^2 f(w) v \geq 0 \implies \nabla^2 f(w) \succeq 0.$$

Hence, by using the second-order definition of convexity, $f(w)$ is convex.

Convex Functions

Operations that preserve convexity: if $f(x)$ and $g(x)$ are convex functions, then $h(x)$ is convex if,

- $h(x) = \alpha f(x)$ for $\alpha \geq 0$ (Non-negative scaling)
  E.g: For $w \in R^d$, $f(w) = \|w\|^2$ is convex, and hence $h(w) = \frac{\lambda}{2}\|w\|^2$ for $\lambda \geq 0$ is convex.
- $h(x) = \max\{f(x), g(x)\}$ (Point-wise maximum)
  E.g: $f(w) = 0$ and $g(w) = 1 - w$ are convex functions, and hence $h(w) = \max\{0, 1 - w\}$ is convex.
- $h(x) = f(Ax + b)$ (Composition with affine map)
  E.g.: $f(w) = \max\{0, 1 - w\}$ is convex, and hence $h(w) = \max\{0, 1 - y_i\langle w, x_i\rangle\}$ for $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$ is convex
- $h(x) = f(x) + g(x)$ (Sum)
  E.g.: $f(w) = \max\{0, 1 - y_i\langle w, x_i\rangle\}$ is convex, and hence
  $h(w) = \sum_{i=1}^{n} \max\{0, 1 - y_i\langle w, x_i\rangle\} + \frac{\lambda}{2}\|w\|^2$ is convex.

Hence, the SVM loss function: $f(w) := \sum_{i=1}^{n} \max\{0, 1 - y_i\langle X_i, w\rangle\} + \frac{\lambda}{2}\|w\|^2$ is convex.

Convex Functions

Q: Prove that ℓ_1-regularized logistic regression:
$f(w) := \sum_{i=1}^{n} \log\left(1 + \exp\left(-y_i \langle X_i, w \rangle\right)\right) + \lambda \left\|w\right\|_1$ is convex.

We have proved that the logistic loss $f(x) = \log(1 + \exp(-x))$ is convex. Since composition with an affine map is convex, and the sum of convex functions is convex, the first term is convex. Since all norms are convex, and a non-negative scaling of a convex function is convex, the second term is convex. Hence, $f(w)$ is convex.

Another way to prove convexity for logistic regression is to compute the Hessian and show that it is positive semi-definite (In Assignment 1)

Jensen's Inequality

- Recall the zero-order definition of convexity: $\forall x, y \in \mathcal{D}$ and $\theta \in [0,1]$, $f(\theta x + (1-\theta)x) \leq \theta f(x) + (1-\theta) f(y)$.
- This can be generalized to n points $\{x_1, x_2, \ldots, x_n\}$, i.e. for $p_i \geq 0$ and $\sum_i p_i = 1$,

$$f(p_1\, x_1 + p_2\, x_2 + \ldots + p_n\, x_n) \leq p_1\, f(x_1) + p_2\, f(x_2) + \ldots + p_n\, f(x_n) \implies f\left(\sum_{i=1}^{n} p_i x_i\right) \leq \sum_{i=1}^{n} p_i f(x_i)$$

- If X is a discrete r.v. that can take value x_i with probability p_i, and f is convex, then,

$$f\left(\mathbb{E}[X]\right) \leq \mathbb{E}\left[f(X)\right]. \qquad \text{(Jensen's inequality)}$$

- Jensen's inequality can be used to prove inequalities like the AM-GM inequality: $\sqrt{ab} \leq \frac{a+b}{2}$.
- *Proof*: Choose $f(x) = -\log(x)$ as the convex function, and consider two points a and b with $\theta = 1/2$. By Jensen's inequality,

$$-\log\left(\frac{a+b}{2}\right) \leq \frac{-\log(a) - \log(b)}{2} \implies \log\left(\frac{a+b}{2}\right) \geq \log(\sqrt{ab}) \implies \frac{a+b}{2} \geq \sqrt{ab}.$$

Holder's Inequality

Q: Prove Holder's inequality, for $p, q \geq 1$ s.t. $\frac{1}{p} + \frac{1}{q} = 1$ and $x, y \in R^d$, $\langle x, y \rangle \leq \|x\|_p \, \|y\|_q$.

By repeating the AM-GM proof, but for a general $\theta \in [0, 1]$, for $a, b \geq 0$, we can prove that,

$$a^\theta b^{1-\theta} \leq \theta a + (1-\theta) b$$

Use $a = \frac{|x_i|^p}{\sum_{j=1}^n |x_j|^p}$, $b = \frac{|y_i|^q}{\sum_{j=1}^n |y_j|^q}$, $\theta = 1/p$, and using the fact that $1 - \theta = 1 - 1/p = 1/q$

$$\left(\frac{|x_i|^p}{\sum_{j=1}^n |x_j|^p}\right)^{1/p} \left(\frac{|y_i|^q}{\sum_{j=1}^n |y_j|^q}\right)^{1/q} \leq \frac{1}{p} \frac{|x_i|^p}{\sum_{j=1}^n |x_j|^p} + \frac{1}{q} \frac{|y_i|^p}{\sum_{j=1}^n |y_j|^p}$$

Summing both sides from $i = 1$ to n,

$$\sum_{i=1}^n \frac{|x_i|}{\left(\sum_{j=1}^n |x_j|^p\right)^{1/p}} \frac{|y_i|}{\left(\sum_{j=1}^n |y_j|^q\right)^{1/q}} \leq 1 \implies \sum_i x_i y_i \leq \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} \left(\sum_{i=1}^n |y_i|^q\right)^{1/q}$$

$$\implies \langle x, y \rangle \leq \|x\|_p \, \|y\|_q$$

Questions?

References i

Atilim Gunes Baydin, Robert Cornish, David Martinez Rubio, Mark Schmidt, and Frank Wood, *Online learning rate adaptation with hypergradient descent*, arXiv preprint arXiv:1703.04782 (2017).