Lecture 11 Mirror descent

- Projected gradient descent

- Mirror descent

- MWU as a special case of Mirror descent

Projected gradient descent

we want to solve the following optimization problem:

convex function

$$\min_{x \in C} f(x)$$

convex set

GD

$$x_{t+1} \leftarrow x_t - \eta_t \nabla f(x_t)$$

Problem: we may end outside C!!!

$\rightsquigarrow$

Projected GD

$$x_{t+1/2} \leftarrow x_t - \eta_t \nabla f(x_t)$$

$$x_{t+1} = \Pi_C(x_{t+1/2})$$

projection in C operator, using as a distance function $\|\cdot\|_2^2 \Rightarrow \Pi_C(x) = \arg\min_{y \in C} \|x-y\|_2^2$

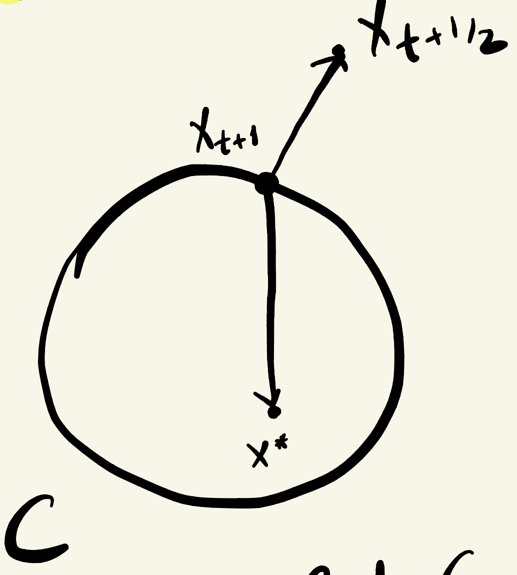
Question:

Is it always good to project?

(in other words, do we get closer to the optimal point x^* by projecting in C)

Answer:

yes!!! $\quad (x_{t+1/2}-x_{t+1})(x^*-x_{t+1}) \le 0$



$$\left((x-\Pi(x))(y-\Pi(x)) \le 0 \quad \forall y \in C, \quad \forall x \right)$$

(remember that C is convex)

$$\Rightarrow \|x_{t+1/2}-x^*\|^2 \geq \|x_{t+1/2}-x_{t+1}\|^2 + \|x^*-x_{t+1}\|^2$$

$$\Rightarrow \|x_{t+1/2}-x^*\|^2 \geq \|x_{t+1}-x^*\|^2$$

Important observation: (in order to derive later on Mirror descent)

the Projected GD updatp rule

$$x_{t+1/2} \leftarrow x_t - \eta_t \nabla f(x_t)$$

$$x_{t+1} \leftarrow \Pi_C(x_{t+1/2}) = \arg\min_{x \in C} \|x - x_{t+1/2}\|_2^2$$

is equivalent to a "regularized" updating rule

$$x_{t+1} = \arg\min_{x \in C} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta_t} \|x - x_t\|_2^2 \right\}$$

proof

$$\operatorname*{argmin}_{x\in C}\left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2n_t}\|x - x_t\|_2^2 \right\} =$$

$$= \operatorname*{argmin}_{x\in C}\left\{ \langle \nabla f(x_t), x \rangle + \frac{1}{2n_t}\|x - x_t\|_2^2 \right\} =$$

$$= \operatorname*{argmin}_{x\in C}\left\{ n_t \langle \nabla f(x_t), x \rangle + \frac{1}{2}\|x - x_t\|_2^2 \right\}$$

and

$$\operatorname*{argmin}_{x\in C} \| x - x_{t+1/2} \|^2 = \operatorname*{argmin}_{x\in C} \| x - x_t + n_t \nabla f(x_t) \|^2 =$$

$$= \operatorname*{argmin}_{x\in C}\left\{ 2 \left(\frac{1}{2}\|x - x_t\|_2^2 + \frac{1}{2}\|n_t \nabla f(x_t)\|_2^2 + \langle x - x_t, n_t \nabla f(x_t) \rangle \right) \right\}$$

$$= \operatorname*{argmin}_{x\in C}\left\{ n_t \langle \nabla f(x_t), x \rangle + \frac{1}{2}\|x - x_t\|_2^2 \right\}$$

$\blacksquare$

we proceed by analysing the convergence speed of Projected Gradient Descent when applied to Lipschitz, convex functions.

Theorem

convex

$$\min_{x \in C} f(x) \quad \text{L-Lipschitz} \Rightarrow |f(x)-f(y)| \leq L \cdot \|x-y\|_2$$

$$\Rightarrow \|\nabla f(x)\|_2 \leq L$$

convex set

the projected G.D. with stepsize

$$\eta = \frac{\|x_0 - x^*\|}{L \cdot \sqrt{T}} \quad \text{satisfies :}$$

$$f\left(\frac{1}{T}\sum_{t=1}^{T} x_t\right) - f(x^*) \leq \frac{L\|x_0 - x^*\|}{\sqrt{T}}$$

and

$$\min_{t \in [T]} f(x_t) - f(x^*) \leq \frac{L \cdot \|x_0 - x^*\|}{\sqrt{T}}$$

$$\overset{\text{proof}}{\Longrightarrow}$$

proof

$$\|x_{t+1}-x^*\|^2 \leq \|x_{t+1/2}-x^*\|^2 \quad \text{(projection)}$$

$$= \|x_t-\eta\nabla f(x_t)-x^*\|^2 \quad \left(\begin{array}{c}\text{definition}\\ \text{of } x_{t+1/2}\end{array}\right)$$

$$= \|x_t-x^*\|^2+\eta^2\|\nabla f(x)\|^2-2\eta\langle\nabla f(x_t), x_t-x^*\rangle \leq$$

convexity of f

$f(x^*) \geq f(x_t)+\langle\nabla f(x_t), x^*-x_t\rangle$

$$\leq \|x_t-x^*\|^2+\eta^2\|\nabla f(x)\|^2+2\eta\left(f(x^*)-f(x_t)\right)$$

↙ adding for $t=0$ to T

$$2\eta\sum_{t=0}^{T}\left(f(x_t)-f(x^*)\right)+\|x_{T+1}-x^*\|^2 \leq \|x_0-x^*\|^2+\eta^2\sum_{t=0}^{T}\|\nabla f(x_t)\|^2$$

$$\underset{\|x_{T+1}-x^*\|^2\geq 0}{\Longrightarrow} \sum_{t=0}^{T}\left(f(x_t)-f(x^*)\right) \leq \frac{\|x_0-x^*\|^2}{2\eta}+\eta\cdot\frac{\sum_{t=0}^{T}\|\nabla f(x_t)\|^2}{2}$$

$$\overset{\|\nabla f(x_t)\|^2\leq\alpha^2}{\Longrightarrow} \frac{1}{T}\sum f(x_t)-f(x^*) \leq \frac{\|x_0-x^*\|^2}{2\eta T}+\frac{\eta\cdot\alpha^2}{2}$$

by noting that $\min_{t\in[T]} f(x_t) \leq \frac{1}{T}\sum_{t=0}^{T} f(x_t)$

$$f\left(\frac{1}{T}\sum_{t=0}^{T}x_t\right) \leq \frac{1}{T}\sum_{t=0}^{T}f(x_t) \quad \text{(Jensen)}$$

and setting $\eta = \|x_0-x^*\|_2/\alpha\cdot\sqrt{T}$ we get the theorem ∎

Observations

(1) Thus after T iterations the error is

$\frac{\|x_0-x^*\|_2 \cdot \lambda}{\sqrt{T}}$. To get an error

of ε we need $T = O\left(\frac{\|x_0-x^*\|_2^2 \cdot \lambda^2}{\varepsilon^2}\right)$

iterations.

(2) we can also use a variable step

size $n_t = \frac{\|x_0-x^*\|_2}{\lambda\sqrt{t+1}}$ and get that

$$\min_{t\in[T]} f(x_t) - f(x^*) \leq \frac{\|x_0-x^*\|_2^2 + \lambda^2 \cdot \sum n_t^2}{\sum n_t} =$$

$$= O\left(\frac{\|x_0-x^*\|^2 \cdot \lambda \cdot \log T}{\sqrt{T}}\right)$$

(3) We still get that

$$\frac{1}{T}\sum_{t=0}^{T}\left(f_t(x_t) - f_t(x^*)\right) \leq \frac{\|x_0-x^*\|_2 \cdot \lambda}{\sqrt{T}}$$

where the function f changes every time (online setting)

What happens if we do not have "very" useful information w.r.t the ℓ_2-norm of the gradient?

For example we know that $\|\nabla f(x)\|_\infty \leq 1$ but only that $\|\nabla f(x)\|_2 \leq \sqrt{n}$.

Mirror descent

Overview of what we will do:

(1) We will define Bregman divergence which "should be viewed" as a distance function

(2) We will redefine the projected G.D update rule to use the Bregman divergence as the projection function (instead of $\|\cdot\|_2$)

(3) choosing appropriate Bregman divergences we get bounds w.r.t other norms than the ℓ_2-norm

Bregman Divergence

Let $\varphi : C \to \mathbb{R}$ be a strictly convex function, continuously differentiable, over a closed convex set C. Then we define the Bregman divergence as

$$D_\varphi(x,y) := \varphi(x) - \varphi(y) - \langle \nabla\varphi(y), x-y \rangle$$

observations and examples

① $D_\varphi(x,y)$ measures how good is the first order approximation of $\varphi(x)$ at y.

② $D_\varphi(x,y) \geq 0 \quad \forall\, x,y \in C$ (φ is strictly convex)

③ $D_\varphi(x,y) = 0$ iff $x = y$ (φ is strictly convex)

④ $\nabla_x D_\varphi(x,y) = \nabla\varphi(x) - \nabla\varphi(y)$

⑤ Generalized law of cosines
(usually this property is called Generalized pythagorean theorem)

$$D_\varphi(x,y) + D_\varphi(y,z) =$$

$$= \varphi(x) - \varphi(y) - \langle \nabla\varphi(y), x-y \rangle$$

$$+ \varphi(y) - \varphi(z) - \langle \nabla\varphi(z), y-z \rangle =$$

$$= \underbrace{\varphi(x) - \varphi(z) - \langle \nabla\varphi(z), x-z \rangle} + D_\varphi(x,z)$$

$$+ \langle \nabla\varphi(z), x-z \rangle - \langle \nabla\varphi(y), x-y \rangle - \langle \nabla\varphi(z), y-z \rangle =$$

$$= D_\varphi(x,z) + \underbrace{\langle \nabla\varphi(z) - \nabla\varphi(y), x-y \rangle}_{\text{"angle" between } z-y \text{ and } x-y}$$

(try $\varphi(x) = \frac{1}{2}\cdot\|x\|^2$ to re-derive the law of cosines)

⑥ Let $\varphi(x) = \frac{1}{2}\|x\|^2 \Rightarrow D_\varphi(x,y) =$

$$= \frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2 - \langle y, x-y \rangle = \frac{1}{2}\|x\|^2 - \frac{1}{2}\|y\|^2 + \|y\|^2 - \langle x, y \rangle =$$

$$= \frac{1}{2}\|x\|^2 + \frac{1}{2}\|y\|^2 - \langle x, y \rangle = \frac{1}{2}\|x-y\|^2$$

(f) Let $\varphi(x) = \sum_{i=1}^{n} x_i \log x_i$, $x \in \Delta^n$

simplex

$$D_\varphi(x,y) = \sum x_i \log x_i - \sum y_i \log y_i$$

$$- \langle 1 + \log y, x - y \rangle =$$

$$= \sum x_i \log x_i - \sum y_i \log y_i - \sum \cancel{(x_i - y_i)}$$

→ 0 (prob. distri)

$$- \sum (x_i - y_i) \log y_i =$$

$$= \sum x_i \log x_i - \sum x_i \log y_i =$$

$$= \sum x_i \log x_i / y_i = KL(x \| y)$$ ☺

Projections with Bregman divergence

Let $\Pi(x) = \underset{y \in C}{\operatorname{argmin}} D_\varphi(y,x)$ be the projection operation using Bregman divergence of φ.

- $\Pi(x)$ is uniquely determined because

$$g(y) = D_\varphi(y,x) = \varphi(y) - \varphi(x) - \langle \nabla\varphi(x), y-x \rangle$$

is strictly convex, and C is a closed convex set $\Rightarrow$ unique minimizer

- $\langle \nabla\varphi(x) - \nabla\varphi(\Pi(x)), y - \Pi(x) \rangle \leq 0 \quad \forall x, \forall y \in C$

<u>proof</u>

$\Pi(x)$ is a minimizer of $D_\varphi(\cdot, x)$ in $C \Rightarrow$

$$\Rightarrow \langle \nabla D_\varphi(\Pi(x), x), y - \Pi(x) \rangle \geq 0 \quad \forall y \in C$$

$$\Rightarrow \langle \nabla\varphi(\Pi(x)) - \nabla\varphi(x), y - \Pi(x) \rangle \geq 0 \quad \forall y \in C$$

- $D_\varphi(y, \Pi(x)) + D_\varphi(\Pi(x), x) \leq D_\varphi(y, x)$

$\forall x \quad , \forall y \in C$

$$\Rightarrow D_\varphi(x^*, x) \geq D_\varphi(x^*, \Pi(x)), \quad x^* \in C$$

☺

In order to construct Mirror descent we will slightly change the update rule of projected G.D.

<u>P.G.D.</u>

$$x_{t+1} = \underset{x \in C}{\arg\min} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2\eta_t} \|x - x_t\|_2^2 \right\}$$

↙

MD

$$x_{t+1} = \underset{x \in C}{\arg\min} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta_t} D_\varphi(x, x_t) \right\}$$

which is equivalent to

$$x_{t+\frac{1}{2}} = \underset{\boxed{x}}{\arg\min} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta_t} D_\varphi(x, x_t) \right\} \rightsquigarrow \text{unconstrained}$$

$$x_{t+1} = \underset{x \in C}{\arg\min}\, D_\varphi(x, x_{t+1/2}) \rightsquigarrow \text{projection step}$$

proof

⟹

- $$x_{t+1} = \underset{x \in C}{\operatorname{argmin}} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta_t} D_\varphi(x, x_t) \right\} =$$

$$= \underset{x \in C}{\operatorname{argmin}} \left\{ \eta_t \langle \nabla f(x_t), x \rangle + \varphi(x) - \langle \nabla \varphi(x_t), x \rangle \right\}$$

$$= \underset{x \in C}{\operatorname{argmin}} \left\{ \varphi(x) - \langle \nabla \varphi(x_t) - \eta_t \nabla f(x_t), x \rangle \right\}$$

- $$x_{t+\frac{1}{2}} = \underset{x}{\operatorname{argmin}} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{\eta_t} D_\varphi(x, x_t) \right\}$$

$$x_{t+1} = \underset{x \in C}{\operatorname{argmin}} D_\varphi(x, x_{t+1/2})$$

the first step is unconstrained, therefore the gradient should be equal to 0 ⇒

$$\Rightarrow \nabla f(x_t) + \frac{1}{\eta_t} \left(\nabla \varphi(x) - \nabla \varphi(x_t) \right) = 0 \Rightarrow$$

$$\Rightarrow \nabla \varphi(x) = \nabla \varphi(x_t) - \eta_t \nabla f(x_t)$$

$$\Rightarrow \nabla \varphi(x_{t+1/2}) = \nabla \varphi(x_t) - \eta_t \nabla f(x_t)$$

$$\Rightarrow x_{t+1/2} = (\nabla \varphi)^{-1} \left(\nabla \varphi(x_t) - \eta_t \nabla f(x_t) \right)$$

$$x_{t+1} = \underset{x \in C}{\operatorname{argmin}} D_\varphi(x, x_{t+1/2}) = \quad \Longrightarrow$$

$$= \arg\min_{x \in C} \left\{ \varphi(x) - \varphi(x_{t+1/2}) - \langle \nabla \varphi(x_{t+\frac{1}{2}}), x - x_{t+\frac{1}{2}} \rangle \right\}$$

$$= \arg\min_{x \in C} \left\{ \varphi(x) - \langle \nabla \varphi(x_{t+1/2}), x \rangle \right\} =$$

$$= \arg\min_{x \in C} \left\{ \varphi(x) - \langle \nabla \varphi(x_t) - \eta_t \nabla f(x_t), x \rangle \right\}$$

▨

Mirror descent

$$x_{t+1/2} = (\nabla \varphi)^{-1} \left(\nabla \varphi(x_t) - \eta_t \nabla f(x_t) \right)$$

$$x_{t+1} = \arg\min_{x \in C} D_\varphi(x, x_{t+1/2})$$

strictly convex, etc...

Theorem (convergence speed)

$D_\varphi(x,y) \geq a\|x-y\|^2$ ← arbitrary norm

f is α-Lipschitz w.r.t $\|\cdot\|_*$ then MD with stepsize $\eta = \frac{\sqrt{a \cdot D_\varphi(x^*, x_0)}}{\alpha \sqrt{T}}$ satisfies

$$f\left(\frac{1}{T}\sum x_t\right) - f(x^*) \leq 2 \frac{\sqrt{D_\varphi(x^*, x_0)} \cdot \alpha}{\sqrt{aT}}$$

$$\min_{t \in [T]} f(x_t) - f(x^*) \leq \quad ''$$

proof

As in the P. GD algorithm where we proved that

$$\|x_{t+1}-x^*\|^2 \leq \|x_t-x^*\|^2 + 2n\left(f(x^*)-f(x_t)\right) + n^2\|\nabla f(x_t)\|_2^2$$

here it suffice to prove that

$$D_\varphi(x^*, x_{t+1}) \leq D_\varphi(x^*, x_t) + n\left(f(x^*)-f(x_t)\right) + \frac{n^2}{a}\|\nabla f(x_t)\|_*^2$$

(then by summing for $t=0,\dots,T$ we get a telescopic sum etc etc)

we start by taking the generalized law of cosines:

$$D_\varphi(x^*, x_{t+1}) + D_\varphi(x_{t+1}, x_t) = D_\varphi(x^*, x_t) + \langle \nabla\varphi(x_t) - \nabla\varphi(x_{t+1}), x^* - x_{t+1} \rangle$$

moreover since $x_{t+1} = \Pi_c(x_{t+1/2})$ we have

$\langle \nabla_{x_{t+1}} D_\varphi(x_{t+1}, x_{t+1/2}), y - x_{t+1} \rangle \geq 0$ (because x_{t+1} is a minimizer)

$$\langle \nabla\varphi(x_{t+1}) - \nabla\varphi(x_{t+1/2}), x_{t+1} - y \rangle \leq 0 \quad \forall y \in C$$

$$\Rightarrow \langle \nabla\varphi(x_{t+1}), x_{t+1} - y \rangle \leq \langle \nabla\varphi(x_{t+1/2}), x_{t+1} - y \rangle \quad \forall y \in C$$

by setting $y = x^*$ and using the law of cosines we get $\Longrightarrow$

$$D_\varphi(x^*,x_{t+1}) + D_\varphi(x_{t+1},x_t) = D_\varphi(x^*,x_t) + \langle \nabla\varphi(x_t) - \nabla\varphi(x_{t+1}), x^* - x_{t+1}$$

$$\Downarrow$$

$$D_\varphi(x^*,x_{t+1}) + D_\varphi(x_{t+1},x_t) \leq D_\varphi(x^*,x_t) + \langle \nabla\varphi(x_t) - \nabla\varphi(x_{t+1/2}), x^* - x_{t+1} \rangle$$

$$\Downarrow \quad \nabla\varphi(x_{t+1/2}) = \nabla\varphi(x_t) - n_t \nabla f(x_t)$$

$$D_\varphi(x^*,x_{t+1}) + D_\varphi(x_{t+1},x_t) \leq D_\varphi(x^*,x_t) + n_t \langle \nabla f(x_t), x^* - x_{t+1} \rangle$$

$$\Downarrow \text{ rearranging}$$

$$D_\varphi(x^*,x_{t+1}) \leq D_\varphi(x^*,x_t) + \boxed{n_t \langle \nabla f(x_t), x^* - x_t \rangle} - n_t \langle \nabla f(x_t), x_{t+1} - x_t \rangle - D_\varphi(x_{t+1},x_t)$$

$$\Downarrow \text{ first order condition}$$

$$D_\varphi(x^*,x_{t+1}) \leq D_\varphi(x^*,x_t) + n_t\big(f(x^*) - f(x_t)\big) - n_t \langle \nabla f(x_t), x_{t+1} - x_t \rangle - D_\varphi(x_{t+1},x_t)$$

$D_\varphi(x,y) \geq a \cdot \|x-y\|^2$

$$\Downarrow \quad n_t \langle \nabla f(x_t), x_{t+1} - x_t \rangle \leq \; \leq n_t \cdot \|\nabla f(x_t)\|_* \, \|x_{t+1} - x_t\|$$

$$D_\varphi(x^*,x_{t+1}) \leq D_\varphi(x^*,x_t) + n_t\big(f(x^*) - f(x_t)\big) + n_t \|\nabla f(x_t)\|_* \|x_{t+1} - x_t\| - a\|x_{t+1} - x_t\|^2$$

$$D_\varphi(x^*, x_{t+1}) \le D_\varphi(x^*, x_t) + \eta_t(f(x^*) - f(x_t)) + \eta_t\|\nabla f(x_t)\|_*\|x_{t+1} - x_t\| - a\|x_{t+1} - x_t\|^2$$

$$\eta_t\|\nabla f(x_t)\|_*\|x_{t+1} - x_t\| - a\|x_{t+1} - x_t\|^2 \le \frac{\eta_t^2\|\nabla f(x_t)\|_*^2}{a} \qquad \Downarrow$$

$$D_\varphi(x^*, x_{t+1}) \le D_\varphi(x^*, x_t) + \eta_t(f(x^*) - f(x_t)) + \frac{\eta_t^2\|\nabla f(x_t)\|_*^2}{a}$$

to make everything work remember we need to select φ s.t. $D_\varphi(x, y) \ge a\|x - y\|^2$

∎

Thus, to achieve ε-optimality we need $2\frac{\sqrt{D_\varphi(x^*, x_0)\cdot\lambda}}{\sqrt{aT}} \le \varepsilon \Rightarrow$

$$\Rightarrow T = O\left(\frac{D_\varphi(x^*, x_0)\,\lambda^2}{a\varepsilon^2}\right)$$

f is λ- Lipschitz w.r.t norm $\|\cdot\|$ which may not be the ℓ_2-norm !!!