Lecture 11  Mirror descent

- Projected gradient descent
- Mirror descent
- MWS as a special case of mirror descent

Projected gradient descent

We want to solve the following optimization problem:

\[
\min_{x \in C} f(x)
\]

\[
\text{convex function}
\]

\[
\text{convex set}
\]

**GD**

\[
X_{t+1} \leftarrow X_t - \eta_t \nabla f(X_t)
\]

Problem: We may end outside \( C \).

**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 \( \| x \|_2 \Rightarrow \Pi_C(x) = \arg \min_{y \in C} \| x - y \|_2^2 \)
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!!!

\[
\begin{aligned}
(x_{t+1} - x_{t+1/2}) (x^* - x_{t+1}) &\leq 0 \\
(x - \Pi_C(x)) (y - \Pi_C(x)) &\leq 0 \\
&\forall y < x, \forall x
\end{aligned}
\]

(C: remember that C is convex)

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

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

Important observation: (in order to derive later the Projected GD update onillon descent)

\[
\begin{aligned}
x_{t+1/2} &\leftarrow x_t - n_t \nabla f(x_t) \\
x_{t+1} &\leftarrow \Pi_C(x_{t+1/2}) = \arg\min_{x \in C} \|x - x_{t+1/2}\|^2
\end{aligned}
\]

is equivalent to a "regularized" updating rule

\[
x_{t+1} = \arg\min_{x \in C} \left\{ f(x_t) + \nabla f(x_t)^T (x - x_t) + \frac{1}{2n_t} \|x - x_{t+1/2}\|^2 \right\}
\]
\[
\arg\min_{x \in \mathcal{C}} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{2n_t} \|x - x_t + n_t \nabla f(x_t)\|^2 \right\} = \frac{1}{2n_t} \left\langle \nabla f(x_t), x \right\rangle + \frac{1}{2} \|x - x_t\|^2
\]

and
\[
\arg\min_{x \in \mathcal{C}} \|x - x_{t+1}\|^2 = \arg\min_{x \in \mathcal{C}} \|x - x_t + n_t \nabla f(x_t)\|^2 = \frac{1}{2} \left( \frac{1}{2} \|x - x_t\|^2 + \|n_t \nabla f(x_t)\|^2 + \langle x - x_t, n_t \nabla f(x_t) \rangle \right)
\]

\[
= \arg\min_{x \in \mathcal{C}} \left\{ n_t \left\langle \nabla f(x_t), x \right\rangle + \frac{1}{2} \|x - x_t\|^2 \right\}
\]
we proceed by analysing the convergence speed of Projected Gradient Descent when applied to \( \text{Lipschitz, convex} \) functions.

**Theorem**

\[
\min_{x \in C} f(x) \quad \text{\( \text{Lipschitz} \Rightarrow \| f(x_1) - f(y) \| \leq \rho \| x_1 - y \|_2 \)}
\]

\[
\Rightarrow \| x^* \|_2 \leq 2
\]

The projected G.D. with stepsize

\[
\eta = \frac{\| x_0 - x^* \|_2}{\sqrt{T}}
\]

satisfies:

\[
f \left( \frac{1}{T} \sum_{t=1}^{T} x_t \right) - f(x^*) \leq \frac{\| x_0 - x^* \|_2}{\sqrt{T}}
\]

and

\[
\min_{t \in [T]} f(x_t) - f(x^*) \leq \frac{\| x_0 - x^* \|_2}{\sqrt{2T}}
\]

**Proof**

\[
\Rightarrow
\]
\[ \| x_{t+1} - x^* \|^2 \leq \| x_t + \frac{1}{2} - x^* \|^2 \quad \text{(projection)} \]
\[
= \| x_t - \nabla f(x_t) - x^* \|^2 \quad \text{(definition of } x_{t+1/2}) \]
\[
= \| x_t - x^* \|^2 + n^2 \| \nabla f(x_t) \|^2 - 2n \langle \nabla f(x_t), x_t - x^* \rangle \leq \]
\[
\text{convexity of } f \quad f(x^* + x_t - x^*) \]
\[
\leq \| x_t - x^* \|^2 + n^2 \| \nabla f(x_t) \|^2 + 2n \left( f(x_t) - f(x^*) \right) \]
\[
\text{adding } \sum_{t=0}^{T} \text{ for } t = 0 \rightarrow T \]
\[
2n \sum_{t=0}^{T} \left( f(x_t) - f(x^*) \right) + \| x_{t+1} - x^* \|^2 \leq \| x_0 - x^* \|^2 + n^2 \sum_{t=0}^{T} \| \nabla f(x_t) \|^2 \]
\[
\Rightarrow \sum_{t=0}^{T} \left( f(x_t) - f(x^*) \right) \leq \| x_0 - x^* \|^2 \sum_{t=0}^{T} \frac{\| \nabla f(x_t) \|^2}{2n} + n \cdot \frac{n^2 d^2}{2} \]
\[
\| \nabla f(x_t) \|^2 \leq d^2 \quad \frac{1}{T} \sum_{t=0}^{T} f(x_t) - f(x^*) \leq \| x_0 - x^* \|^2 \frac{1}{2nT} + \frac{n \cdot d^2}{2} \]

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 \( n = \| x_0 - x^* \|^2 / d^2 \sqrt{T} \) we get the Regev.
Observations

1) Thus after $T$ iterations the error is

$$\frac{\|x_0-x^*\|_2}{\sqrt{T}}.$$ To get an error of $\varepsilon$ we need $T = O\left(\frac{\|x_0-x^*\|_2^2\cdot \varepsilon}{\varepsilon^2}\right)$ iterations.

2) We can also use a variable step size $n_t = \frac{\|x_0-x^*\|_2}{\sqrt{t+1}}$ and get that

$$\min_{t \geq 0} f(x_t) - f(x^*) \leq \frac{\|x_0-x^*\|_2^2 + \frac{\varepsilon^2}{\varepsilon^2}}{\sum n_t} = \sum n_t = O\left(\frac{\|x_0-x^*\|_2^2 \cdot \log T}{\sqrt{T}}\right).$$

3) We still get that

$$\frac{1}{T} \sum_{t=0}^{T} (f_t(x_t) - f_t(x^*)) \leq \frac{\|x_0-x^*\|_2^2}{\sqrt{T}},$$

where the function $f$ changes every time (online setting).
What happens if we do not have "very" useful information w.r.t. the $\ell_2$-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 $\ell_2$-norm.
Bregman Divergence

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

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

Observations and examples

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

2. $D_\psi(x, y) > 0$ if $x, y \in C$ ($\psi$ is strictly convex)

3. $D_\psi(x, y) = 0$ if $x = y$ ($\psi$ is strictly convex)

4. $\nabla_x D_\psi(x, y) = \nabla \psi(x) - \nabla \psi(y)$
5. Generalized Law of cosines
(usually this property is called Generalized Pythagorean theorem)

\[ D_\psi(x, y) + D_\psi(y, z) = \]

\[ = \psi(x) - \psi(y) - \langle \nabla \psi(y), x-y \rangle \]

\[ + \psi(y) - \psi(z) - \langle \nabla \psi(z), y-z \rangle = \]

\[ = \langle \nabla \psi(y), x-z \rangle + D_\psi(x, z) \]

\[ + \langle \nabla \psi(z), x-z \rangle - \langle \nabla \psi(y), x-y \rangle \]

\[ = D_\psi(x, z) + \langle \nabla \psi(z) - \nabla \psi(y), x-y \rangle \]

"Angle" between

\[ z-y \] and \[ x-y \] (tiny \( \psi(x) = \frac{1}{2} \| x \|^2 \)

to re-derive the law of cosines)

6. Let \( \psi(x) = \frac{1}{2} \| x \|^2 \) \( \Rightarrow \)

\[ D_\psi(x, y) = \]

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

\[ = \frac{1}{2} \| x \|^2 + \frac{1}{2} \| y \|^2 - \langle x, y \rangle = \frac{1}{2} \| x-y \|^2 \]
\[ \text{let } \varphi(x) = \frac{1}{n} \sum_{i=1}^{n} x_i \log x_i , \quad x \in \Delta^n \]

\[ \text{simplex} \]

\[ D_\varphi(x,y) = \sum x_i \log x_i - \sum y_i \log y_i ; \]

\[ - \langle 1 + \log y, x-y \rangle = \]

\[ = \sum x_i \log x_i - \sum y_i \log y_i ; \]

\[ \leq \sum (x_i - y_i) \log y_i ; \]

\[ \leq \sum (x_i - y_i) \log y_i ; \]

\[ = \sum x_i \log x_i - \sum x_i \log y_i ; \]

\[ = \sum x_i \log x_i / y_i ; \]

\[ = \Delta \log (x \mid y) \]

\[ \text{Projections with Bregman divergence} \]

\[ \text{let } \Pi(x) = \arg \min_{y \in \mathcal{C}} D_\varphi(y;x) \text{ be the projection operation using Bregman divergence of } \varphi. \]
• $\Pi(x)$ is uniquely determined because
  
  \[ g(y) = D\phi(y,x) = \psi(y) - \psi(x) - \langle \nabla \psi(x), y - x \rangle \]

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

  \[ \langle \nabla \psi(x) - \nabla \psi(\Pi(x)), y - \Pi(x) \rangle \geq 0 \quad \forall x, \forall y \in C \]

  \[ \text{proof} \]

  $\Pi(x)$ is a minimizer of $D\phi(y,x)$ in $C \Rightarrow$

  \[ \Rightarrow \langle \nabla D\phi(\Pi(x), x), y - \Pi(x) \rangle \geq 0 \quad \forall y \in C \]

  \[ \Rightarrow \langle \nabla \psi(\Pi(x)) - \nabla \psi(x), y - \Pi(x) \rangle \geq 0 \quad \forall y \in C \]

  • $D\phi(y, \Pi(x)) + D\phi(\Pi(x), x) \leq D\phi(y, x) \quad \forall x, \forall y \in C$

  \[ \Rightarrow D\phi(x^*, x) > D\phi(x^*, \Pi(x)), x^* \in C \]
In order to construct mirror descent, we will slightly change the update rule of projected G.D.

\[ x_{t+1} = \text{argmin}_{x \in C} \left[ f(x_t) + \frac{1}{2} \nabla f(x_t)^T (x - x_t) + \frac{1}{2n_t} \|x - x_t\|^2 \right] \]

\[ x_{t+1} = \text{argmin}_{x \in C} \left( f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{n_t} D_\psi(x, x_t) \right) \]

which is equivalent to

\[ x_{t+1} = \text{argmin}_{x \in C} \left( f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{n_t} D_\psi(x, x_t) \right) \]

\( x \) \( \xrightarrow{\text{unconstrained}} \)

\[ x_{t+1} = \text{argmin}_{x \in C} D_\psi(x, x_t + 1/2) \]

\( \xrightarrow{\text{projection step}} \)

proof
\[ x_{t+1} = \arg\min_{x \in \mathcal{C}} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{n_t} \nabla \psi(x_t, x, \nabla f(x_t)) \right\} = \]
\[ = \arg\min_{x \in \mathcal{C}} \left\{ n_t \langle \nabla f(x_t), x \rangle + \psi(x) - \langle \nabla \psi(x_t), x \rangle \right\} = \]
\[ = \arg\min_{x \in \mathcal{C}} \left\{ \psi(x) - \langle \nabla \psi(x_t) - n_t \nabla f(x_t), x \rangle \right\} \]
\[ x_{t+1} = \arg\min_{x \in \mathcal{C}} \left\{ f(x_t) + \langle \nabla f(x_t), x - x_t \rangle + \frac{1}{n_t} \nabla \psi(x_t, x, \nabla f(x_t)) \right\} = \]
\[ x_{t+1} = \arg\min_{x \in \mathcal{C}} \nabla \psi(x_t, x_t + \frac{1}{2} \nabla \psi(x_t, x, \nabla f(x_t)) \]

The first step is unconstrained, therefore the gradient should be equal to 0 \[ \Rightarrow \nabla f(x_t) + \frac{1}{n_t} (\nabla \psi(x) - \nabla \psi(x_t)) = 0 \Rightarrow \]
\[ \Rightarrow \nabla \psi(x_t) = \nabla \psi(x_t) - n_t \nabla f(x_t) \]
\[ \Rightarrow \nabla \psi(x_t + \frac{1}{2}) = \nabla \psi(x_t) - n_t \nabla f(x_t) \]
\[ \Rightarrow x_{t+1} = \nabla \psi(x_t, x_t + \frac{1}{2}) = \nabla \psi(x_t) - n_t \nabla f(x_t) \]
\[ x_{t+1} = x_t + \frac{1}{2} \nabla \psi(x_t, x_t + \frac{1}{2}) = \nabla \psi(x_t) - n_t \nabla f(x_t) \]
$$\begin{align*}
\text{argmin}_{x \in C} \{ \psi(x) - \psi(x_{t+1/2}) - \langle \nabla \psi(x_{t+1/2}), x - x_{t+1/2} \rangle \} \\
= \text{argmin}_{x \in C} \{ \psi(x) - \langle \nabla \psi(x_{t+1/2}), x \rangle \} \\
= \text{argmin}_{x \in C} \{ \psi(x) = \langle \nabla \psi(x_t) - n_t + \nabla f(x_t), x \rangle \}
\end{align*}$$

**Mirror descent**

$$x_{t+1/2} = (\nabla \psi)^{-1}( \nabla \psi(x_t) - n_t + \nabla f(x_t))$$

$$x_{t+1} = \text{argmin}_{x \in C} D_{\psi}(x, x_{t+1/2})$$

Theorem (convergence speed)

$$D_{\psi}(x^*, y) \geq \| x^* - y \|^2$$

Arbitrary norm

If is $\alpha$-Hölder w.r.t. $\| \cdot \|$ then MD with stepsize $n = \frac{\sqrt{\alpha \cdot D_{\psi}(x^*|x_0)}}{\sqrt{t}}$ satisfies

$$f\left(\frac{1}{t} \sum x_t\right) - f(x^*) \leq 2 \sqrt{\frac{D_{\psi}(x^*|x_0) \cdot \alpha}{\sqrt{t}}}$$

$$\min_{t \geq 1} f(x_t) - f(x^*) \leq \frac{\alpha}{2}$$
As in the P. 60 algorithm where we proved that
\[ \|x_{t+1} - x^*\|^2 \leq \|x_t - x^*\|^2 + 2n (f(x^*) - f(x_t)) + n^2 \| \nabla f(x_t) \|^2 \]
leave it suffice to prove that
\[ D_\phi(x^*, x_{t+1}) \leq D_\phi(x^*, x_t) + n (f(x^*) - f(x_t)) + \frac{n^2 \| \nabla f(x_t) \|^2}{a} \]
(then by summing for \( t=0, \ldots, T \) we get a telescopic sum etc etc)

we start by taking the generalized law of cosines:
\[ D_\phi(x^*, x_{t+1}) + D_\phi(x_{t+1}, x_t) = D_\phi(x^*, x_t) + \langle \nabla \phi(x_{t+1}) - \nabla \phi(x_{t+1/2}), x_{t+1} - x^* \rangle \]
moreover since \( x_{t+1} = \Pi_c(x_{t+1/2}) \) we have

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

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

by setting \( y = x^* \) and using the law of cosines we get \( \Rightarrow \)
\[
D_\varphi(x^{*}_i, x^{*}_{i+1}) + D_\varphi(x^{*}_{i+1}, x^*_i) = D_\varphi(x^*_i, x^*_t) + \nabla \varphi(x^*_i) - \nabla \varphi(x^*_t) \leq x^*_i - x^*_t >
\]

\[
\Rightarrow \nabla \varphi(x^*_{i+1}) = \nabla \varphi(x^*_i) - n_+ \nabla f(x^*_i)
\]

\[
D_\varphi(x^*_i, x^{*}_{i+1}) + D_\varphi(x^{*}_{i+1}, x^*_i) \leq D_\varphi(x^*_i, x^*_t) + n_+ < \nabla f(x^*_i), x^*_t - x^*_i >
\]

\[
\Rightarrow \text{ rearranging }
\]

\[
D_\varphi(x^*_i, x^{*}_{i+1}) \leq D_\varphi(x^*_i, x^*_t) + n_+ < \nabla f(x^*_i), x^*_t - x^*_i > - n_+ < \nabla f(x^*_i), x^*_t - x^*_i >
\]

\[
\Rightarrow \text{ first order condition } - D_\varphi(x^{*}_{i+1}, x^*_i)
\]

\[
D_\varphi(x^*_i, x^{*}_{i+1}) \leq D_\varphi(x^*_i, x^*_t) + n_+ (f(x^*_i) - f(x^*_t)) - n_+ < \nabla f(x^*_i), x^*_t - x^*_i >
\]

\[
\Rightarrow D_\varphi(x, y) \geq \alpha \|x - y\|^2
\]

\[
D_\varphi(x^*_i, x^{*}_{i+1}) \leq D_\varphi(x^*_i, x^*_t) + n_+ (f(x^*_i) - f(x^*_t)) + n_+ \| \nabla f(x^*_i) \| \| x^*_t - x^*_i \|
\]

\[
\Rightarrow \| x^*_t - x^*_i \|^2 >
\]
\[ D_\rho(x^k, x_{t+1}) \leq D_\rho(x^k, x_t) + \nu_f(f(x^k) - f(x_t)) + \nu_f \| \nabla f(x_t) \|_x \| x_{t+1} - x_t \|_x \\
- \alpha \| x_{t+1} - x_t \|^2 \\
- \alpha \| x_{t+1} - x_t \|^2 \\
\leq \frac{\nu^2_f \| \nabla f(x_t) \|_x^2}{\alpha} \]

\[ D_\rho(x^k, x_{t+1}) \leq D_\rho(x^k, x_t) + \nu_f(f(x^k) - f(x_t)) + \frac{\nu^2_f \| \nabla f(x_t) \|_x^2}{\alpha} \]

to make everything work remember we need to select \( \frac{\nu^2_f \| \nabla f(x_t) \|_x^2}{\alpha} \gtrsim 0 \).

Thus, to achieve \( \epsilon \)-optimality we need \( \frac{2 \sqrt{D_\rho(x^k, x_0)} \cdot \delta}{\sqrt{\alpha T}} \leq \epsilon = \frac{\nu^2_f \| \nabla f(x_t) \|_x^2}{\alpha} \)

\( = \frac{T}{\delta} = O\left( \frac{D_\rho(x^k, x_0)^{\frac{1}{2}}}{\delta^2} \right) \)

\( f \) is \( \delta \)-Hölder w.r.t. norm \( \| \cdot \| \) which may not be the \( l_2 \)-norm !!!