ELEG/CISC 867: Advanced Machine Learning **Spring 2019**

Lecture 9: Convex Learning Problems

Lecturer: Xiugang Wu *04/11/2019, 04/16/2019, 04/18/2019*

Convex learning comprises an important family of learning problems, mainly because most of what can learn efficiently falls into it. Take linear predictors as example:

- Halfspaces with the 0-1 loss is a nonconvex problem, and is indeed known to be computationally hard to learn in the unrealizable case;
- Linear regression with square loss is a convex problem, and can be indeed learned efficiently;
- Logistic regression with log loss is also a convex problem and can be learned efficiently.

In general, a convex learning problem is a problem where the hypothesis class is a convex set, and the loss function is a convex function for each example. Two particular families of convex learning problems are convex-smooth-bounded problems and convex-Lipschitz-bounded problems, which will be shown to be learnable in the next two lectures.

1 Convex Learning Problems

1.1 Convexity

Convex Set. A set C in a vector space is convex if for any two vectors $\mathbf{u}, \mathbf{v}$ in C, the line segment between $\mathbf{u}$ and $\mathbf{v}$ is contained in C, i.e. the convex combination $\alpha\mathbf{u} + (1-\alpha)\mathbf{v} \in C$ for any $\alpha \in [0,1]$.

Convex Function. Let C be a convex set. A function $f : C \to \mathbb{R}$ is convex if for every $\mathbf{u}, \mathbf{v} \in C$ and $\alpha \in [0,1]$,

$$f(\alpha\mathbf{u} + (1-\alpha)\mathbf{v}) \leq \alpha f(\mathbf{u}) + (1-\alpha) f(\mathbf{v}).$$

Epigraph. The epigraph of a function f is the set

$$\text{epigraph}(f) = \{(\mathbf{x}, \beta) : f(\mathbf{x}) \leq \beta\}.$$

A function f is convex iff its epigraph(f) is a convex set.

Local Minimum and Global Minimum. An important property of convex functions is that every local minimum of the function is also a global minimum.

- Let $B(\mathbf{u}, r) = \{\mathbf{v} : \|\mathbf{u} - \mathbf{v}\| \leq r\}$ be a ball centered at $\mathbf{u}$ of radius r. We say that $f(\mathbf{u})$ is a local minimum of f at $\mathbf{u}$ if there exists some $r > 0$ s.t. $f(\mathbf{u}) \leq f(\mathbf{v})$ for any $\mathbf{v} \in B(\mathbf{u}, r)$.
- Consider an arbitrary $\mathbf{w}$ that is not necessarily in $B(\mathbf{u}, r)$. There must exist some $\alpha > 0$ s.t. $(1-\alpha)\mathbf{u} + \alpha\mathbf{w} \in B(\mathbf{u}, r)$ and therefore

  $$f(\mathbf{u}) \leq f((1-\alpha)\mathbf{u} + \alpha\mathbf{w}) \leq (1-\alpha) f(\mathbf{u}) + \alpha f(\mathbf{w}).$$

  This immediately implies that $f(\mathbf{u}) \leq f(\mathbf{w})$ for any $\mathbf{w}$.

Tangent to Convex f. Another important property of convex functions is that for every $\mathbf{w}$ we can construct a tangent to f at $\mathbf{w}$ that lies below f everywhere.

- If f is differentiable, the tangent to f at $\mathbf{w}$ is given by the affine function

$$l(\mathbf{u}) = f(\mathbf{w}) + \nabla f(\mathbf{w})^T(\mathbf{u} - \mathbf{w})$$

  where $\nabla f(\mathbf{w})$ is the gradient of f at $\mathbf{w}$, defined as

$$\nabla f(\mathbf{w}) = \left(\frac{\partial f(\mathbf{w})}{\partial w_1}, \frac{\partial f(\mathbf{w})}{\partial w_2}, \ldots, \frac{\partial f(\mathbf{w})}{\partial w_d} \right).$$

- If f is convex and differentiable, then for any $\mathbf{u}$,

$$f(\mathbf{u}) \geq l(\mathbf{u}) = f(\mathbf{w}) + \nabla f(\mathbf{w})^T(\mathbf{u} - \mathbf{w}).$$

How to Check Convexity. Let $f : \mathbf{R} \to \mathbf{R}$ be twice differentiable. Then we have

$$f \text{ is convex} \Leftrightarrow f' \text{ is monotonically nondecreasing} \Leftrightarrow f'' \text{ is nonnegative.}$$

For example, consider the following two functions which are building blocks for the square loss and logistic loss function:

- The function $f(x) = x^2$ is convex; $f'(x) = 2x$; $f''(x) = 2$.
- The function $\log(1 + e^x)$ is convex; $f'(x) = e^x/(1 + e^x) = 1/(1 + e^{-x})$ is increasing.

Composition of Convex Function with Affine Function. The composition of a convex function with an affine function is convex. In particular, let $f(\mathbf{w}) = g(\mathbf{w}^T\mathbf{x} + y)$ where $\mathbf{x} \in \mathbb{R}^d, y \in \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ is convex. Then f is convex in $\mathbf{w}$.

For example, consider the square loss function for linear regression and logistic loss function for logistic regression:

- Consider $f(\mathbf{w}) = (\mathbf{w}^T\mathbf{x} - y)^2$. f is convex in $\mathbf{w}$ since it is the composition of $g(a) = a^2$ onto an affine function $\mathbf{w}^T\mathbf{x} - y$.
- Consider $f(\mathbf{w}) = \log\left(1 + e^{-y\mathbf{w}^T\mathbf{x}}\right)$. f is convex in $\mathbf{w}$ since it is the composition of $g(a) = \log(1 + e^a)$ onto an affine function $-y\mathbf{w}^T\mathbf{x}$.

Maximum and Weighted Sum of Convex Functions. The maximum of convex functions is convex, and a weighted sum of convex functions with nonnegative weights is also convex. In particular, let $f_i : \mathbb{R}^d \to \mathbb{R}$ be convex for any $i \in [1 : n]$. Then both the functions

$$\max_{i \in [1:n]} f_i(x)$$

and

$$\sum_{i \in [1:n]} w_i f_i(x), \text{ with } w_i \geq 0 \; \forall i \in [1 : n]$$

are convex. For example, the function $g(x) = |x|$ is convex since $g(x) = \max\{x, -x\}$, where both x and $-x$ are convex.

1.2 Convex Learning Problems

Recall that from previous lectures, a general learning problem consists of $(\mathcal{H}, \mathcal{Z}, \ell)$, where $\mathcal{H}$ is the hypothesis class, $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ is the space of (input, label) example pairs, and $\ell : \mathcal{Y} \times \hat{\mathcal{Y}} \to \mathbb{R}_+$ is the loss function. With a slight abuse of notation, we can also reload $\mathcal{H}$ to be the set of vectors $\mathbf{w}$ that parametrizes $h_{\mathbf{w}} \in \mathcal{H}$, and reload ℓ to be a mapping $\ell : \mathcal{H} \times \mathcal{Z} \to \mathbb{R}_+$, i.e. $\ell(\mathbf{w}, z) = \ell(y, h_{\mathbf{w}}(x))$ for $\mathbf{w} \in \mathcal{H}, z \in \mathcal{Z}$.

A learning problem, $(\mathcal{H}, \mathcal{Z}, \ell)$, is convex if the hypothesis class $\mathcal{H}$ is a convex set and for all $z \in \mathcal{Z}$, the loss function $\ell(\mathbf{w}, z)$ is a convex function in $\mathbf{w}$.

- Linear regression with square loss is a convex learning problem. Here, $\mathcal{H} = \mathbb{R}^d$ is convex, $\mathcal{Z} = \mathbb{R}^d \times \mathbb{R}$, $\ell(\mathbf{w}, z) = (\mathbf{w}^T\mathbf{x} - y)^2$ is convex in $\mathbf{w}$ for any $(\mathbf{x}, y) \in \mathcal{Z}$.
- Logistic regression with log(istic) loss is a convex learning problem. Here, $\mathcal{H} = \mathbb{R}^d$ is convex, $\mathcal{Z} = \mathbb{R}^d \times \{\pm 1\}$, $\ell(\mathbf{w}, z) = \log\left(1 + e^{-y\mathbf{w}^T\mathbf{x}}\right)$ is convex in $\mathbf{w}$ for any $(\mathbf{x}, y) \in \mathcal{Z}$.

The reason that we define a convex learning problem in the above way is precisely because of the following fact: If the loss function ℓ is a convex function and $\mathcal{H}$ is a convex set, then the $\mathrm{ERM}_{\mathcal{H}}$ problem is a convex optimization problem (i.e. minimizing a convex function over a convex set). To see this, recall that the $\mathrm{ERM}_{\mathcal{H}}$ problem is given by

$$\mathrm{ERM}_{\mathcal{H}}(z^n) = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{H}} L(\mathbf{w}, z^n) = \operatorname*{argmin}_{\mathbf{w} \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(\mathbf{w}, z_i).$$

If $\ell(\mathbf{w}, z_i)$ is convex in $\mathbf{w}$ for any $i \in [1 : n]$, then $L(\mathbf{w}, z^n)$ is also convex in $\mathbf{w}$. Therefore, $\mathrm{ERM}_{\mathcal{H}}$ is a problem of minimizing a convex function over a convex set, i.e. a convex optimization problem. Under mild conditions, such problems can be solved efficiently using generic optimization algorithms.

2 Convex-Lipschitz/Smooth-Bounded Learning Problems

2.1 Learnability of Convex Learning Problems

We have seen that for many cases, implementing ERM for convex learning problems can be done efficiently. But is convexity a sufficient condition for the learnability of a problem? Unfortunately the answer is negative as demonstrated in the following example; note that here in general one may not be able to resort to VC theory for the learnability since VC theory only deals with binary classification problems.

Example 2.1 *Consider a linear regression problem with square loss, where $\mathcal{H} = \mathbb{R}$ and $\ell(w, (x, y)) = (wx - y)^2$; clearly, this problem is convex, but we will now show that it is not PAC learnable. Suppose that it is indeed learnable, with algorithm A being a successful PAC learner for the problem. Then by definition, for any $\epsilon, \delta > 0$ and any P on $\mathcal{Z}$,*

$$P^n(|L(A(Z^n), P) - \min_{w \in \mathbb{R}} L(w, P)| \leq \epsilon) \geq 1 - \delta \tag{1}$$

whenever $n \geq n_{\mathcal{H}}(\epsilon, \delta)$.

Now choose $\epsilon = 1/100, \delta = 1/2$, let $n \geq n_{\mathcal{H}}(\epsilon, \delta)$, and set $\mu = \frac{\log(100/99)}{2n}$. Consider two distributions P_1 and P_2 supported on two particular examples $z_a = (1, 0)$ and $z_b = (\mu, -1)$, where

$$P_1(z_a) = \mu, \;\; P_1(z_b) = 1 - \mu;$$
$$\text{and } P_2(z_a) = 0, \;\; P_2(z_b) = 1.$$

Under both distributions, the probability that all examples in Z^n appear to be z_b is at least 99%. Note that this is trivially true under P_2, whereas under P_1 this probability is $(1-\mu)^n \geq e^{-2\mu n} = 0.99$.

Let $\hat{w} = A(z^n)$ for z^n consisting of all z_b examples. We will argue that no matter what value $\hat{w}$ takes, the condition (1) will be violated under P_1 or P_2, and therefore the problem is not PAC learnable.

- *Suppose $\hat{w} < -\frac{1}{2\mu}$. We can show that condition (1) is violated under P_1. In particular, $L(\hat{w}, P_1) \geq P_1(z_a)\ell(\hat{w}, z_a) = \mu(\hat{w})^2 \geq \frac{1}{4\mu}$ whereas $\min_w L(w, P_1) \leq L(0, P_1) = P_1(z_b)\ell(0, z_b) = 1-\mu$, and therefore $L(\hat{w}, P_1) - \min_w L(w, P_1) \geq \frac{1}{4\mu} - (1-\mu) \geq \epsilon$.*
- *Suppose $\hat{w} \geq -\frac{1}{2\mu}$. We can show that condition (1) is violated under P_2. In particular, $L(\hat{w}, P_2) \geq \ell(\hat{w}, z_b) = (\hat{w}\mu + 1)^2 \geq \frac{1}{4}$ whereas $\min_w L(w, P_2) = 0$, and therefore $L(\hat{w}, P_2) - \min_w L(w, P_2) \geq \epsilon$.*

A possible solution to the above issue of convex learning problems being non-learnable, is to add another constraint on the hypothesis class. In addition to the convexity requirement, we require that $\mathcal{H}$ will be bounded, i.e. $\|w\| \leq B, \forall w \in \mathcal{H}$ for some $B > 0$. However, boundedness and convexity alone are still not sufficient for ensuring that the problem is learnable, as demonstrated in the following example.

Example 2.2 *Consider the same setup in the previous example but now let $\mathcal{H} = \{w \in \mathbb{R} : \|w\| \leq 1\}$, $z_a = (1/\mu, 0)$ and $z_b = (1, -1)$. Let $\hat{w} = A(z^n)$ for z^n consisting of all z_b examples. Then again, no matter what value $\hat{w}$ takes, the condition (1) will be violated under P_1 or P_2.*

- *If $\hat{w} < -\frac{1}{2}$, then condition (1) is violated under P_1. In particular, $L(\hat{w}, P_1) \geq \frac{1}{4\mu}$ whereas $\min_w L(w, P_1) \leq L(0, P_1) = 1-\mu$, and therefore $L(\hat{w}, P_1) - \min_w L(w, P_1) \geq \frac{1}{4\mu} - (1-\mu) \geq \epsilon$.*
- *If $\hat{w} \geq -\frac{1}{2}$, then condition (1) is violated under P_2. In particular, $L(\hat{w}, P_2) \geq \frac{1}{4}$ whereas $\min_w L(w, P_2) = 0$, and therefore $L(\hat{w}, P_2) - \min_w L(w, P_2) \geq \epsilon$.*

This example shows that we need further assumptions on the learning problem, and this time the remedy is in Lipschitzness or smoothness of the loss function. This motivates us to define two particular families of convex learning problems, i.e. convex-Lipschitz-bounded problems and convex-smooth-bounded problems.

2.2 Convex-Lipschitz-Bounded Learning Problems

Lipschitzness. Let $C \subseteq \mathbb{R}^d$. A function $f : \mathbb{R}^d \to \mathbb{R}^k$ is ρ-Lipschitz over C if for any $\mathbf{w}_1, \mathbf{w}_2 \in C$ we have $\|f(\mathbf{w}_1) - f(\mathbf{w}_2)\| \leq \rho\|\mathbf{w}_1 - \mathbf{w}_2\|$.

Intuitively, a Lipschitz function cannot change too fast. In particular, if $f : \mathbb{R} \to \mathbb{R}$ is differentiable, then by the mean value theorem we have

$$f(w_1) - f(w_2) = f'(u)(w_1 - w_2)$$

for some $u \in [w_1, w_2]$; therefore, if the derivative of f is everywhere bounded (in absolute value) by ρ, then the function is ρ-Lipschitz.

Some examples of Lipschitz functions are as follows:

- The function $f(x) = |x|$ is 1-Lipschitz over $\mathbb{R}$. To see this, note that by the triangle inequality, for any x_1, x_2,
$$|x_1| - |x_2| = |x_1 - x_2 + x_2| - |x_2| \leq |x_1 - x_2| + |x_2| - |x_2| = |x_1 - x_2|$$
and also, by switching the role of x_1 and x_2,
$$|x_2| - |x_1| \leq |x_1 - x_2|.$$

Therefore, we have

$$|f(x_1) - f(x_2)| = \big||x_1| - |x_2|\big| = \max\{|x_1| - |x_2|, |x_2| - |x_1|\} \leq |x_1 - x_2|$$

and hence $f(x) = |x|$ is 1-Lipschitz over $\mathbb{R}$.

- The function $f(x) = \log(1 + e^x)$ is 1-Lipschitz over $\mathbb{R}$. This follows from the fact that

$$|f'(x)| = \left|\frac{1}{1 + e^{-x}}\right| \leq 1.$$

- The function $f(x) = x^2$ is not ρ-Lipschitz over $\mathbb{R}$ for any ρ. This is because for $x_1 = 0$ and $x_2 = 1 + \rho$,

$$|f(x_2) - f(x_1)| = (1 + \rho)^2 > \rho(1 + \rho) = \rho|x_2 - x_1|.$$

  However, $f(x) = x^2$ is ρ-Lipschitz over the set $C = \{x : |x| \leq \rho/2\}$ because for any $x_1, x_2 \in C$,

$$|f(x_2) - f(x_1)| = |x_1 + x_2||x_1 - x_2| \leq 2(\rho/2)|x_2 - x_1| = \rho|x_2 - x_1|.$$

  One can also make sense of this by looking at the absolute value $|f'(x)|$ of the derivative of $f(x)$, which is unbounded over $\mathbb{R}$ but bounded by ρ over C.

- The affine function $f(\mathbf{w}) = \mathbf{w}^T\mathbf{v} + b$ is $\|\mathbf{v}\|$-Lipschitz because by Cauchy-Schwartz inequality,

$$|f(\mathbf{w}_1) - f(\mathbf{w}_2)| = |\mathbf{w}_1^T\mathbf{v} - \mathbf{w}_2^T\mathbf{v}| = |(\mathbf{w}_1 - \mathbf{w}_2)^T\mathbf{v}| \leq \|\mathbf{v}\|\|\mathbf{w}_1 - \mathbf{w}_2\|.$$

Composition of Lipschitz Functions. Let $f(\mathbf{x}) = g_1(g_2(\mathbf{x}))$ where g_1 is ρ_1-Lipschitz and g_2 is ρ_2-Lipschitz. Then f is $\rho_1\rho_2$-Lipschitz because

$$\begin{aligned}|f(\mathbf{x}_1) - f(\mathbf{x}_2)| &= |g_1(g_2(\mathbf{x}_1)) - g_1(g_2(\mathbf{x}_2))| \\ &\leq \rho_1\|g_2(\mathbf{x}_1) - g_2(\mathbf{x}_2)\| \\ &\leq \rho_1\rho_2\|\mathbf{x}_1 - \mathbf{x}_2\|.\end{aligned}$$

In particular, if g_2 is the affine function $g_2(\mathbf{x}) = \mathbf{v}^T\mathbf{x} + b$, then f is $\rho_1\|\mathbf{v}\|$-Lipschitz.

Convex-Lipschitz-Bounded Learning Problem. A learning problem, $(\mathcal{H}, \mathcal{Z}, \ell)$, is called Convex-Lipschitz-Bounded, with parameter ρ, B if:

- The hypothesis class $\mathcal{H}$ is a convex set and $\|\mathbf{w}\| \leq B$ for all $\mathbf{w} \in \mathcal{H}$;
- For all $z \in \mathcal{Z}$, the loss function $\ell(\mathbf{w}, z)$ is convex in $\mathbf{w}$ and ρ-Lipschitz.

Example 2.3 *Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq \rho\}$ and $\mathcal{Y} = \mathbb{R}$. Let $\mathcal{H} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leq B\}$ and let the loss function be $\ell(\mathbf{w}, (\mathbf{x}, y)) = |\mathbf{w}^T\mathbf{x} - y|$. This corresponds to a regression problem with the absolute-value loss, where we assume that the instances are in a ball of radius ρ and we restrict the hypothesis to be homogenous linear functions parametrized by $\mathbf{w}$ whose norm is bounded by B. This problem is Convex-Lipschitz-Bounded, with parameter ρ, B.*

2.3 Convex-Smooth-Bounded Learning Problems

Smoothness. A differentiable function $f : \mathbb{R}^d \to \mathbb{R}$ is β-smooth if its gradient is β-Lipschitz, i.e. for any $\mathbf{v}, \mathbf{w}$ we have $\|\nabla f(\mathbf{v}) - \nabla f(\mathbf{w})\| \leq \beta\|\mathbf{v} - \mathbf{w}\|$.

It is possible to show that smoothness implies that for all $\mathbf{v}, \mathbf{w}$ we have

$$f(\mathbf{v}) \leq f(\mathbf{w}) + \nabla f(\mathbf{w})^T(\mathbf{v} - \mathbf{w}) + \frac{\beta}{2}\|\mathbf{v} - \mathbf{w}\|^2. \tag{2}$$

On the other hand, convexity implies that

$$f(\mathbf{v}) \geq f(\mathbf{w}) + \nabla f(\mathbf{w})^T(\mathbf{v} - \mathbf{w}).$$

Therefore, when a function is convex and smooth, we have both upper and lower bounds on the difference between the function $f(\mathbf{v})$ and its first order approximation $f(\mathbf{w}) + \nabla f(\mathbf{w})^T(\mathbf{v} - \mathbf{w})$.

Self-Bounded Functions. Setting $\mathbf{v} = \mathbf{w} - \frac{1}{\beta}\nabla f(\mathbf{w})$ in (2), we obtain

$$\frac{1}{2\beta}\|\nabla f(\mathbf{w})\|^2 \leq f(\mathbf{w}) - f(\mathbf{v}).$$

If we further assume that $f(\mathbf{v}) \geq 0$ for all $\mathbf{v}$, then we conclude that smoothness implies that

$$\|\nabla f(\mathbf{w})\|^2 \leq 2\beta f(\mathbf{w}). \tag{3}$$

A function satisfying the above property is called a self-bounded function; a nonnegative smooth function is hence self-bounded.

Some examples of smooth functions are as follows:

- The function $f(x) = x^2$ is 2-smooth since $f'(x) = 2x$. Note that for this particular convex smooth function, conditions (2) and (3) hold with equality.
- The function $f(x) = \log(1 + e^x)$ is 1/4-smooth since $f'(x) = \frac{1}{1+e^{-x}}$ and

$$f''(x) = \frac{e^{-x}}{(1 + e^{-x})^2} = \frac{1}{(1 + e^{-x})(1 + e^x)} \leq 1/4.$$

  Since this function is nonnegative, condition (3) also holds, i.e. the function is self-bounded.

Composition of Smooth Function on Affine Function. Let $f(\mathbf{w}) = g(\mathbf{w}^T\mathbf{x} + b)$ where g is β-smooth. Then f is $\beta\|\mathbf{x}\|^2$-smooth. To see this note that by chain rule, $\nabla f(\mathbf{w}) = g'(\mathbf{w}^T\mathbf{x} + b)\mathbf{x}$ and therefore

$$\begin{aligned}
\|\nabla f(\mathbf{v}) - \nabla f(\mathbf{w})\| &= \|g'(\mathbf{v}^T\mathbf{x} + b)\mathbf{x} - g'(\mathbf{w}^T\mathbf{x} + b)\mathbf{x}\| \\
&= \|\mathbf{x}\| \cdot |g'(\mathbf{v}^T\mathbf{x} + b) - g'(\mathbf{w}^T\mathbf{x} + b)| \\
&\leq \|\mathbf{x}\| \cdot \beta\|(\mathbf{v} - \mathbf{w})^T\mathbf{x}\| \\
&\leq \|\mathbf{x}\|^2 \cdot \beta\|\mathbf{v} - \mathbf{w}\|.
\end{aligned}$$

In particular, if $f(\mathbf{w}) = (\mathbf{w}^T\mathbf{x} - y)^2$ where $\mathbf{x} \in \mathbb{R}^d$ and $y \in \mathbb{R}$, then f is $2\|x\|^2$-smooth; if $f(\mathbf{w}) = \log\left(1 + e^{-y\mathbf{w}^T\mathbf{x}}\right)$ where $\mathbf{x} \in \mathbb{R}^d$ and $y \in \{\pm 1\}$, then f is $\|x\|^2/4$-smooth.

Convex-Smooth-Bounded Learning Problem. A learning problem, $(\mathcal{H}, \mathcal{Z}, \ell)$, is called Convex-Smooth-Bounded, with parameter β, B if:

- The hypothesis class $\mathcal{H}$ is a convex set and $\|\mathbf{w}\| \leq B$ for all $\mathbf{w} \in \mathcal{H}$;
- For all $z \in \mathcal{Z}$, the loss function $\ell(\mathbf{w}, z)$ is convex, nonnegative and β-smooth.

Example 2.4 *Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\|^2 \leq \beta/2\}$ and $\mathcal{Y} = \mathbb{R}$. Let $\mathcal{H} = \{\mathbf{w} \in \mathbb{R}^d : \|\mathbf{w}\| \leq B\}$ and let the loss function be $\ell(\mathbf{w}, (\mathbf{x}, y)) = (\mathbf{w}^T\mathbf{x} - y)^2$. This corresponds to a regression problem with the sqaure loss, where we assume that the instances are in a ball of radius $\sqrt{\beta/2}$ and we restrict the hypothesis to be homogenous linear functions parametrized by $\mathbf{w}$ whose norm is bounded by B. This problem is Convex-Smooth-Bounded, with parameter β, B.*