Recall Results on Binary Classification

- \( Z_i = (X_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\} \)
- Admissible action set \( \mathcal{A} \subseteq \mathcal{B} := \{a : \mathbb{R}^d \to \{-1, 1\}\} \)
- True loss function \( \ell(a, (x, y)) = 1_{a(x)\neq y} = \varphi^*(a(x)y) \) with \( \varphi^*(u) := 1_{u \leq 0} \)

\[
\begin{align*}
    r(a) &= \mathbb{P}(a(X) \neq Y) & a^* &= \arg\min_{a \in \mathcal{A}} r(a) & a^{**} &= \arg\min_{a \in \mathcal{B}} r(a) \\
    R(a) &= \frac{1}{n} \sum_{i=1}^{n} 1_{a(X_i)\neq Y_i} & A^* &= \arg\min_{a \in \mathcal{A}} R(a)
\end{align*}
\]

So far we have proved:

\[
\mathbb{P}\left( r(A^*) - r(a^*) < 54 \sqrt{\frac{\text{VC}(\mathcal{A})}{n}} + \sqrt{2 \frac{\log(1/\delta)}{n}} \right) \geq 1 - \delta
\]

**Problem:** In general, computing \( A^* \) is NP hard!

**Idea:** Define convex relaxation of the original problem
Convex function (Definition 8.1)

A function \( f : \mathbb{R}^d \to \mathbb{R} \) is convex if for every \( x, \tilde{x} \in \mathbb{R}^d, \lambda \in [0, 1] \) we have

\[
f(\lambda x + (1 - \lambda)\tilde{x}) \leq \lambda f(x) + (1 - \lambda) f(\tilde{x})
\]

Convex set (Definition 8.2)

A set \( A \) is convex if for every \( a, \tilde{a} \in A, \lambda \in [0, 1] \) we have

\[
\lambda a + (1 - \lambda)\tilde{a} \in A
\]
Convex Loss Surrogates

A function $\varphi : \mathbb{R} \to \mathbb{R}_+$ is called a **convex loss surrogate** if:
- convex
- non-increasing
- $\varphi(0) = 1$

**True loss:**
$\varphi^*(u) = 1_{u \leq 0}$

**Exponential loss:**
$\varphi(u) = e^{-u}$

**Hinge loss:**
$\varphi(u) = \max\{1 - u, 0\}$

**Logistic loss:**
$\varphi(u) = \log_2(1 + e^{-u})$
Convex Soft Classifiers

1. **Soft** classifiers \( A_{\text{soft}} \subseteq B_{\text{soft}} := \{ a : \mathbb{R}^d \rightarrow \mathbb{R} \} \)

2. If \( a \in B_{\text{soft}} \), corresponding **hard** classifier is given by \( \text{sign}(a) \)

1. **Linear functions with convex parameter space:**

\[
A_{\text{soft}} = \{ a(x) = w^\top x + b : w \in C_1 \subseteq \mathbb{R}^d, b \in C_2 \subseteq \mathbb{R} \}
\]

\( C_1, C_2 \) are convex sets

2. **Majority votes (Boosting):**

\[
A_{\text{soft}} = \{ a(x) = \sum_{i=1}^{m} w_j h_j(x) : w = (w_1, \ldots, w_m) \in \Delta_m \}
\]

\( \Delta_m \) is the \( m \)-dim. simplex and \( h_1, \ldots, h_m : \mathbb{R}^d \rightarrow \mathbb{R} \) are **base classifiers**

---

**Empirical \( \varphi \)-Risk Minimization**

If \( \varphi \) and \( A_{\text{soft}} \) are convex, we are left with a convex problem

\[
R_\varphi(a) := \frac{1}{n} \sum_{i=1}^{n} \varphi(a(X_i)Y_i) \\
A_\varphi^* := \text{argmin}_{a \in A_{\text{soft}}} R_\varphi(a)
\]
Zhang’s Lemma

\[ r_\varphi(a) := \mathbb{E}\varphi(a(X)Y) \]

\[ r(a) := \mathbb{E}\varphi^*(a(X)Y) = \mathbb{P}(a(X) \neq Y) \]

\[ a^{**} := \arg\min_{a \in \mathcal{B}} r_\varphi(a) \]

\[ a^{**} := \arg\min_{a \in \mathcal{B}} r(a) \]

Zhang’s Lemma (Lemma 8.5)

Let \( \varphi : \mathbb{R} \to \mathbb{R}_+ \) be a convex loss surrogate. For any \( \tilde{\eta} \in [0, 1], \tilde{a} \in \mathbb{R} \), let

\[ H_{\tilde{\eta}}(\tilde{a}) := \varphi(\tilde{a})\tilde{\eta} + \varphi(-\tilde{a})(1 - \tilde{\eta}), \quad \tau(\tilde{\eta}) := \inf_{\tilde{a} \in \mathbb{R}} H_{\tilde{\eta}}(\tilde{a}). \]

Assume that there exist \( c > 0 \) and \( \nu \in [0, 1] \) such that

\[
\left| \tilde{\eta} - \frac{1}{2} \right| \leq c(1 - \tau(\tilde{\eta}))^\nu \quad \text{for any } \tilde{\eta} \in [0, 1]
\]

Then, for any \( a : \mathbb{R}^d \to \mathbb{R} \) we have

\[
r(\text{sign}(a)) - r(a^{**}) \leq 2c(r_\varphi(a) - r_\varphi(a^{**}))^\nu
\]

excess risk
hard classifier

excess \( \varphi \)-risk
soft classifier
Zhang’s Lemma: Examples

- **Exponential loss:**
  \[ \tau(\tilde{\eta}) = 2\sqrt{\tilde{\eta}(1 - \tilde{\eta})} \]
  \[ c = 1/\sqrt{2} \]
  \[ \nu = 1/2 \]

- **Hinge loss:**
  \[ \tau(\tilde{\eta}) = 1 - |1 - 2\tilde{\eta}| \]
  \[ c = 1/2 \]
  \[ \nu = 1 \]

- **Logistic loss:**
  \[ \tau(\tilde{\eta}) = -\tilde{\eta} \log_2 \tilde{\eta} - (1 - \tilde{\eta}) \log_2 (1 - \tilde{\eta}) \]
  \[ c = 1/\sqrt{2} \]
  \[ \nu = 1/2 \]

Zhang’s Lemma shows that we can reliably focus on convex problems.
Subgradients (Definition 8.7)

Let \( f : C \subset \mathbb{R}^d \rightarrow \mathbb{R} \). A vector \( g \in \mathbb{R}^d \) is a subgradient of \( f \) at \( x \in C \) if
\[
    f(x) - f(y) \leq g^T(x - y) \quad \text{for any } y \in C
\]

The set of subgradients of \( f \) at \( x \) is denoted \( \partial f(x) \).

Subgradients yield global information (uniform lower bounds)

Convexity and subgradients (Theorem 8.8)

Let \( f : C \subseteq \mathbb{R}^d \rightarrow \mathbb{R} \) with \( C \) convex:
\[
    f \text{ is convex} \implies \text{for any } x \in \text{int}(C), \partial f(x) \neq \emptyset
\]
\[
    f \text{ is convex} \iff \text{for any } x \in C, \partial f(x) \neq \emptyset
\]

If \( f \) is convex and differentiable at \( x \), then \( \nabla f(x) \in \partial f(x) \)

Convex functions that are differentiable allow to infer global information (i.e., subgradients) from local information (i.e., gradients)

This is why convex problems are “typically” amenable to computations...
To prove algorithms converge we need additional local-to-global properties...
Local-to-Global Properties

- **Convex:**
  \[ f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall x, y \in \mathbb{R}^d \]

- **\( \alpha \)-Strongly Convex:**
  \[ \exists \alpha > 0 \text{ such that } f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{\alpha}{2} \| y - x \|^2 \quad \forall x, y \in \mathbb{R}^d \]

- **\( \beta \)-Smooth:**
  \[ \exists \beta > 0 \text{ such that } f(y) \leq f(x) + \nabla f(x)^T (y - x) + \frac{\beta}{2} \| y - x \|^2 \quad \forall x, y \in \mathbb{R}^d \]

- **\( \gamma \)-Lipschitz:**
  \[ \exists \gamma > 0 \text{ such that } f(x) - \gamma \| y - x \|^2 \leq f(y) \leq f(x) + \gamma \| y - x \|^2 \quad \forall x, y \in \mathbb{R}^d \]

<table>
<thead>
<tr>
<th>Loss Function</th>
<th>Strongly convex?</th>
<th>Smooth?</th>
<th>Lipschitz?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exponential loss (in ( \mathbb{R} ))</td>
<td>NO</td>
<td>NO</td>
<td>NO</td>
</tr>
<tr>
<td>Hinge loss (in ( \mathbb{R} ))</td>
<td>NO</td>
<td>NO</td>
<td>YES</td>
</tr>
<tr>
<td>Logistic loss (in ( \mathbb{R} ))</td>
<td>NO</td>
<td>YES</td>
<td>YES</td>
</tr>
</tbody>
</table>

However, we typically only need the domain to be a compact set of \( \mathbb{R} \).