Mixability in Statistical Learning

Tim van Erven

Joint work with: Peter Grünwald, Mark Reid, Bob Williamson
Summary

• **Stochastic mixability** → fast rates of convergence in different settings:
  
  • statistical learning (margin condition)

  • sequential prediction (mixability)
Outline

• Part 1: Statistical learning
  • Stochastic mixability (definition)
  • Equivalence to margin condition

• Part 2: Sequential prediction

• Part 3: Convexity interpretation for stochastic mixability

• Part 4: Grünwald’s idea for adaptation to the margin
Notation
Notation

- Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
- Predict \(Y\) from \(X\): \(\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathcal{A}\}\)
- Loss: \(\mathcal{L} : \mathcal{Y} \times \mathcal{A} \rightarrow [0, \infty]\)
Notation

• Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)

• Predict \(Y\) from \(X\): \(\mathcal{F} = \{ f : \mathcal{X} \to \mathcal{A} \}\)

• Loss: \(\ell : \mathcal{Y} \times \mathcal{A} \to [0, \infty]\)

Classification

\(\mathcal{Y} = \{0, 1\}, \mathcal{A} = \{0, 1\}\)

\(\ell(y, a) = \begin{cases} 0 & \text{if } y = a \\ 1 & \text{if } y \neq a \end{cases}\)
Notation

• Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)

• Predict \(Y\) from \(X\): \(\mathcal{F} = \{f : \mathcal{X} \rightarrow \mathcal{A}\}\)

• Loss: \(\ell : \mathcal{Y} \times \mathcal{A} \rightarrow [0, \infty]\)

<table>
<thead>
<tr>
<th>Classification</th>
<th>Density estimation</th>
</tr>
</thead>
<tbody>
<tr>
<td>(\mathcal{Y} = {0, 1}, \mathcal{A} = {0, 1})</td>
<td>(\mathcal{A} = \text{density functions on } \mathcal{Y})</td>
</tr>
</tbody>
</table>
| \(\ell(y, a) = \begin{cases} 
0 & \text{if } y = a \\
1 & \text{if } y \neq a
\end{cases}\) | \(\ell(y, p) = -\log p(y)\) |
Notation

- Data: \((X_1, Y_1), \ldots, (X_n, Y_n)\)
- Predict \(Y\) from \(X\): \(\mathcal{F} = \{ f : \mathcal{X} \rightarrow \mathcal{A} \}\)
- Loss: \(\ell : \mathcal{Y} \times \mathcal{A} \rightarrow [0, \infty]\)

### Classification

\(\mathcal{Y} = \{0, 1\}, \mathcal{A} = \{0, 1\}\)

\(\ell(y, a) = \begin{cases} 0 & \text{if } y = a \\ 1 & \text{if } y \neq a \end{cases}\)

### Density estimation

\(\mathcal{A} = \text{density functions on } \mathcal{Y}\)

\(\ell(y, p) = - \log p(y)\)

Without \(X\): \(\mathcal{F} \subset \mathcal{A}\)
Statistical Learning
Statistical Learning

\[(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{iid}}{\sim} P^*\]

\[f^* = \arg \min_{f \in \mathcal{F}} \mathbb{E}[\ell(Y, f(X))]\]

\[d(\hat{f}, f^*) = \mathbb{E}[\ell(Y, \hat{f}(X)) - \ell(Y, f^*(X))]\]
(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} P^*

f^* = \arg \min_{f \in \mathcal{F}} \mathbb{E}[\ell(Y, f(X))]

d(\hat{f}, f^*) = \mathbb{E}[\ell(Y, \hat{f}(X)) - \ell(Y, f^*(X))]
Statistical Learning

\[(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} P^*\]

\[f^* = \arg \min_{f \in \mathcal{F}} \mathbb{E}[\ell(Y, f(X))]\]

\[d(\hat{f}, f^*) = \mathbb{E}[\ell(Y, \hat{f}(X)) - \ell(Y, f^*(X))] = O(n^{-?})\]
Statistical Learning

\[(X_1, Y_1), \ldots, (X_n, Y_n) \overset{iid}{\sim} P^*\]

\[f^* = \arg \min_{f \in \mathcal{F}} \mathbb{E}[\ell(Y, f(X))]\]

\[d(\hat{f}, f^*) = \mathbb{E}[\ell(Y, \hat{f}(X)) - \ell(Y, f^*(X))] = O(n^{-?})\]

• Two factors that determine rate of convergence:
  1. complexity of \(\mathcal{F}\)
  2. the margin condition
Definition of Stochastic Mixability

• Let $\eta \geq 0$. Then $(\ell, \mathcal{F}, P^*)$ is $\eta$-stochastically mixable if there exists an $f^* \in \mathcal{F}$ such that

$$E \left[ \frac{e^{-\eta \ell(Y,f(X))}}{e^{-\eta \ell(Y,f^*(X))}} \right] \leq 1 \quad \text{for all } f \in \mathcal{F}.$$ 

• Stochastically mixable: this holds for some $\eta > 0$
Immediate Consequences

\[
E \left[ \frac{e^{-\eta \ell(Y, f(X))}}{e^{-\eta \ell(Y, f^*(X))}} \right] \leq 1 \quad \text{for all } f \in \mathcal{F}
\]

- \( f^* \) minimizes risk over \( \mathcal{F} \): 
  \[
  f^* = \arg\min_{f \in \mathcal{F}} E[\ell(Y, f(X))]
  \]

- The larger \( \eta \), the stronger the property of being \( \eta \)-stochastically mixable
Density estimation example 1

• Log-loss: \( \ell(y, p) = -\log p(y) \), \( \mathcal{F} = \{ p_\theta \mid \theta \in \Theta \} \)

• Suppose \( p_{\theta^*} \in \mathcal{F} \) is the true density

• Then for \( \eta = 1 \) and any \( p_\theta \in \mathcal{F} \):

\[
\mathbb{E} \left[ \frac{e^{-\eta \ell(Y, p_\theta)}}{e^{-\eta \ell(Y, p_{\theta^*})}} \right] = \int \frac{p_\theta(y)}{p_{\theta^*}(y)} P^*(dy) = 1
\]
Density estimation example 2
Density estimation example 2

• Normal location family with fixed variance $\sigma^2$:
  $$\mathcal{F} = \{ \mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R} \} \quad P^* = \mathcal{N}(\mu^*, \tau^2)$$

• $\eta$-stochastically mixable for $\eta = \sigma^2 / \tau^2$:
  $$\mathbb{E} \left[ \frac{e^{-\eta \ell(Y;p_\mu)}}{e^{-\eta \ell(Y;p_{\mu^*})}} \right] = \frac{1}{\sqrt{2\pi\tau^2}} \int e^{-\frac{\eta}{2\sigma^2} (y-\mu)^2 + \frac{\eta}{2\sigma^2} (y-\mu^*)^2 - \frac{1}{2\tau^2} (y-\mu^*)^2} \, dy$$
  $$= \frac{1}{\sqrt{2\pi\tau^2}} \int e^{-\frac{1}{2\tau^2} (y-\mu)^2} \, dy = 1$$
Density estimation example 2

• Normal location family with fixed variance $\sigma^2$:
  \[
  \mathcal{F} = \{ \mathcal{N}(\mu, \sigma^2) \mid \mu \in \mathbb{R} \} \quad P^* = \mathcal{N}(\mu^*, \tau^2)
  \]

• $\eta$-stochastically mixable for $\eta = \sigma^2 / \tau^2$:
  \[
  \mathbb{E} \left[ \frac{e^{-\eta l(Y, p_\mu)}}{e^{-\eta l(Y, p_{\mu^*})}} \right] = \frac{1}{\sqrt{2\pi \tau^2}} \int e^{-\frac{\eta}{2\sigma^2} (y-\mu)^2 + \frac{\eta}{2\sigma^2} (y-\mu^*)^2 - \frac{1}{2\tau^2} (y-\mu^*)^2} \, dy
  \]
  \[
  = \frac{1}{\sqrt{2\pi \tau^2}} \int e^{-\frac{1}{2\tau^2} (y-\mu)^2} \, dy = 1
  \]

• If $\hat{f}$ is empirical mean: \[
  \mathbb{E}[d(\hat{f}, f^*)] = \frac{\tau^2}{2\sigma^2 n} = \frac{\eta^{-1}}{2n}
  \]
Outline

• Part 1: Statistical learning
  • Stochastic mixability (definition)
  • Equivalence to margin condition

• Part 2: Sequential prediction

• Part 3: Convexity interpretation for stochastic mixability

• Part 4: Grünwald’s idea for adaptation to the margin
Margin condition

\[ c_0 V(f, f^*)^\kappa \leq d(f, f^*) \quad \text{for all } f \in \mathcal{F} \]

- where \( d(f, f^*) = \mathbb{E}[\ell(Y, f(X)) - \ell(Y, f^*(X))] \)
  \( V(f, f^*) = \mathbb{E} \left( \ell(Y, f(X)) - \ell(Y, f^*(X)) \right)^2 \)
- \( \kappa \geq 1, c_0 > 0 \)
- For 0/1-loss implies rate of convergence \( O(n^{-\kappa/(2\kappa-1)}) \) [Tsybakov, 2004]
- So smaller \( \kappa \) is better
Stochastic mixability ↔ margin

\[ c_0 V(f, f^\ast)^\kappa \leq d(f, f^\ast) \quad \text{for all } f \in \mathcal{F} \]

- **Thm [\kappa = 1]**: Suppose \( \ell \) takes values in \([0, V]\). Then \((\ell, \mathcal{F}, P^\ast)\) is stochastically mixable if and only if there exists \( c_0 > 0 \) such that the margin condition is satisfied with \( \kappa = 1 \).
Margin condition with $\kappa > 1$

\[ F_\epsilon = \{ f^* \} \cup \{ f \in F \mid d(f, f^*) \geq \epsilon \} \]

- **Thm [all $\kappa \geq 1$]:** Suppose $\ell$ takes values in $[0, V]$. Then the margin condition is satisfied if and only if there exists a constant $C > 0$ such that, for all $\epsilon > 0$, $(\ell, F_\epsilon, P^*)$ is $\eta$-stochastically mixable for $\eta = C\epsilon^{(\kappa-1)/\kappa}$. 
Outline

• Part 1: Statistical learning
• Part 2: Sequential prediction
• Part 3: Convexity interpretation for stochastic mixability
• Part 4: Grünwald’s idea for adaptation to the margin
Sequential Prediction with Expert Advice

- For rounds $t = 1, \ldots, n$:
  - K experts predict $\hat{f}_t^1, \ldots, \hat{f}_t^K$
  - Predict $(x_t, y_t)$ by choosing $\hat{f}_t$
  - Observe $(x_t, y_t)$

- Regret $= \frac{1}{n} \sum_{t=1}^{n} \ell(y_t, \hat{f}_t(x_t)) - \min_k \frac{1}{n} \sum_{t=1}^{n} \ell(y_t, \hat{f}_t^k(x_t))$

- Game-theoretic (minimax) analysis: want to guarantee small regret against adversarial data
Sequential Prediction with Expert Advice

- For rounds $t = 1, \ldots, n$:
  - $K$ experts predict $\hat{f}_t^1, \ldots, \hat{f}_t^K$
  - Predict $(x_t, y_t)$ by choosing $\hat{f}_t$
  - Observe $(x_t, y_t)$

- Regret
  \[
  \frac{1}{n} \sum_{t=1}^{n} \ell(y_t, \hat{f}_t(x_t)) - \min_{k} \frac{1}{n} \sum_{t=1}^{n} \ell(y_t, \hat{f}_t^k(x_t))
  \]

- Worst-case regret $= O(1/n)$ iff the loss is mixable!  [Vovk, 1995]
Mixability

- A loss $\ell: \mathcal{Y} \times \mathcal{A} \rightarrow [0, \infty]$ is $\eta$-mixable if for any distribution $\pi$ on $\mathcal{A}$ there exists an action $a_\pi \in \mathcal{A}$ such that

$$\mathbb{E}_{A \sim \pi} \left[ \frac{e^{-\eta \ell(y,A)}}{e^{-\eta \ell(y,a_\pi)}} \right] \leq 1$$

for all $y$.

- Vovk: fast $O(1/n)$ rates if and only if loss is mixable.
(Stochastic) Mixability

• A loss $\ell : \mathcal{Y} \times \mathcal{A} \to [0, \infty]$ is $\eta$-mixable if for any distribution $\pi$ on $\mathcal{A}$ there exists an action $a_\pi \in \mathcal{A}$ such that

$$\mathbb{E}_{A \sim \pi} \left[ \frac{e^{-\eta \ell(y,A)}}{e^{-\eta \ell(y,a_\pi)}} \right] \leq 1 \quad \text{for all } y.$$ 

• $(\ell, \mathcal{F}, P^*)$ is $\eta$-stochastically mixable if

$$\mathbb{E}_{X,Y \sim P^*} \left[ \frac{e^{-\eta \ell(Y,f(X))}}{e^{-\eta \ell(Y,f^*(X))}} \right] \leq 1 \quad \text{for all } f \in \mathcal{F}.$$
(Stochastic) Mixability

- A loss $\ell : \mathcal{Y} \times \mathcal{A} \rightarrow [0, \infty]$ is $\eta$-mixable if for any distribution $\pi$ on $\mathcal{A}$ there exists an action $a_\pi \in \mathcal{A}$ such that

$$\ell(y, a_\pi) \leq -\frac{1}{\eta} \ln \int e^{-\eta \ell(y,a)} \pi(da) \quad \text{for all } y.$$
(Stochastic) Mixability

- A loss $\ell: \mathcal{Y} \times \mathcal{A} \rightarrow [0, \infty]$ is $\eta$-mixable if for any distribution $\pi$ on $\mathcal{A}$ there exists an action $a_\pi \in \mathcal{A}$ such that

$$\ell(y, a_\pi) \leq -\frac{1}{\eta} \ln \int e^{-\eta \ell(y, a)} \pi(da) \quad \text{for all } y.$$ 

- **Thm:** $(\ell, \mathcal{F}, P^*)$ is $\eta$-stochastically mixable iff for any distribution $\pi$ on $\mathcal{F}$ there exists $f^* \in \mathcal{F}$ such that

$$\mathbb{E}[\ell(Y, f^*(X))] \leq \mathbb{E}\left[-\frac{1}{\eta} \ln \int e^{-\eta \ell(Y, f(X))} \pi(df)\right]$$
Equivalence of Stochastic Mixability and Ordinary Mixability
Equivalence of Stochastic Mixability and Ordinary Mixability

\[ \mathcal{F}_{\text{full}} = \{ \text{all functions from } \mathcal{X} \text{ to } A \} \]

- **Thm**: Suppose \( \ell \) is a proper loss and \( \mathcal{X} \) is discrete. Then \( \ell \) is \( \eta \)-mixable if and only if \(( \ell, \mathcal{F}_{\text{full}}, P^* )\) is \( \eta \)-stochastically mixable for all \( P^* \).
Equivalence of Stochastic Mixability and Ordinary Mixability

\[ \mathcal{F}_{\text{full}} = \{ \text{all functions from } \mathcal{X} \text{ to } A \} \]

- **Thm**: Suppose \( \ell \) is a proper loss and \( \mathcal{X} \) is discrete. Then \( \ell \) is \( \eta \)-mixable if and only if \((\ell, \mathcal{F}_{\text{full}}, P^*)\) is \( \eta \)-stochastically mixable for all \( P^* \).

- Proper losses are e.g. 0/1-loss, log-loss, squared loss.

- Thm generalizes to other losses that satisfy two technical conditions.
Summary

- **Stochastic mixability** fast rates of convergence in different settings:
  - statistical learning (margin condition)
  - sequential prediction (mixability)
Outline

• Part 1: Statistical learning

• Part 2: Sequential prediction

• Part 3: Convexity interpretation for stochastic mixability

• Part 4: Grünwald’s idea for adaptation to the margin
Density estimation example 1

- Log-loss: $\ell(y, p) = -\log p(y)$, $\mathcal{F} = \{p_\theta \mid \theta \in \Theta\}$
- Suppose $p_{\theta^*} \in \mathcal{F}$ is the true density
- Then for $\eta = 1$ and any $p_\theta \in \mathcal{F}$:

$$
\mathbb{E} \left[ \frac{e^{-\eta \ell(Y,p_\theta)}}{e^{-\eta \ell(Y,p_{\theta^*})}} \right] = \int \frac{p_\theta(y)}{p_{\theta^*}(y)} P^*(dy) = 1
$$
Log-loss example 3 (convex $\mathcal{F}$)

- **Log-loss:** $\ell(y, p) = -\log p(y)$, $\mathcal{F} = \{p_\theta \mid \theta \in \Theta\}$

- **Suppose model misspecified:** $p_{\theta^*} = \arg \min_{p_\theta \in \mathcal{F}} \mathbb{E}[−\log p_\theta(Y)]$
  is not the true density

- **Thm [Li, 1999]:** Suppose $\mathcal{F}$ is convex. Then

  $$\int \frac{p_\theta(y)}{p_{\theta^*}(y)} P^*(dy) \leq 1 \quad \text{for all } p_\theta \in \mathcal{F}$$

- **Convexity is common condition for convergence of minimum description length and Bayesian methods**
Log-loss and convexity for $\eta = 1$
Log-loss and convexity for $\eta = 1$

**Thm:** $(\ell, \mathcal{F}, P^*)$ is $\eta$-stochastically mixable iff for any distribution $\pi$ on $\mathcal{F}$ there exists $f^* \in \mathcal{F}$ such that

$$
\mathbb{E}[\ell(Y, f^*(X))] \leq \mathbb{E}
\left[
-\frac{1}{\eta} \ln \int e^{-\eta \ell(Y, f(X))} \pi(\mathrm{d}f)
\right]
$$
Log-loss and convexity for $\eta = 1$

- Thm: $(\ell, \mathcal{F}, P^*)$ is $\eta$-stochastically mixable iff for any distribution $\pi$ on $\mathcal{F}$ there exists $f^* \in \mathcal{F}$ such that

$$E[\ell(Y, f^*(X))] \leq E[-\frac{1}{\eta} \ln \int e^{-\eta \ell(Y, f(X))} \pi(df)]$$

- Corollary: For log-loss, 1-stochastic mixability means

$$\min_{p \in \mathcal{F}} E[-\ln p(Y)] = \min_{p \in \text{co}(\mathcal{F})} E[-\ln p(Y)],$$

where $\text{co}(\mathcal{F})$ denotes the convex hull of $\mathcal{F}$. 
Log-loss and convexity for $\eta = 1$

**Corollary:** For log-loss, 1-stochastic mixability means

$$\min_{p \in \mathcal{F}} \mathbb{E}[-\ln p(Y)] = \min_{p \in \text{co}(\mathcal{F})} \mathbb{E}[-\ln p(Y)],$$

where $\text{co}(\mathcal{F})$ denotes the convex hull of $\mathcal{F}$. 
Convexity interpretation with pseudo-likelihoods

- **Pseudo-likelihoods**: \( p_{f,\eta}(Y|X) = e^{-\eta \ell(Y,f(X))} \)

\[ \mathcal{P}_\mathcal{F}(\eta) = \{ p_{f,\eta}(Y|X) \mid f \in \mathcal{F} \} \]

- **Corollary**: \((\ell, \mathcal{F}, P^*)\) is \(\eta\)-stochastically mixable iff

\[ \min_{p \in \mathcal{P}_\mathcal{F}(\eta)} \mathbb{E}\left[-\frac{1}{\eta} \ln p(Y|X)\right] = \min_{p \in \text{co}(\mathcal{P}_\mathcal{F}(\eta))} \mathbb{E}\left[-\frac{1}{\eta} \ln p(Y|X)\right] \]
Outline

- Part 1: Statistical learning
- Part 2: Sequential prediction
- Part 3: Convexity interpretation for stochastic mixability
- Part 4: Grünwald’s idea for adaptation to the margin
Adapting to the margin / $\eta$

- Penalized empirical risk minimization:

$$
\hat{f} = \arg \min_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \ell(Y_i, f(X_i)) + \lambda \cdot \text{pen}(f) \right\}
$$

- Optimal $\lambda \propto 1/\eta$ depends on $\eta$ / the margin

- Single model: take $\text{pen}(f) = \text{const.}$ no need to know $\lambda$

- Model selection: $\mathcal{F} = \bigcup_{m} \mathcal{F}_m$, $\text{pen}(f) = \text{pen}(m) \neq \text{const.}$
Convexity testing
Convexity testing

• **Corollary:** \((\ell, \mathcal{F}, P^*)\) is \(\eta\)-stochastically mixable iff

\[
\min_{p \in \mathcal{P}_\mathcal{F}(\eta)} \mathbb{E}[ -\frac{1}{\eta} \ln p(Y|X) ] = \min_{p \in \text{co}(\mathcal{P}_\mathcal{F}(\eta))} \mathbb{E}[ -\frac{1}{\eta} \ln p(Y|X) ]
\]
Convexity testing

- **Corollary:** \((\ell, \mathcal{F}, P^*)\) is \(\eta\)-stochastically mixable iff
  \[
  \min_{p \in \mathcal{P}(\mathcal{F}(\eta))} \mathbb{E}\left[-\frac{1}{\eta} \ln p(Y|X)\right] = \min_{p \in \text{co}(\mathcal{P}(\mathcal{F}(\eta)))} \mathbb{E}\left[-\frac{1}{\eta} \ln p(Y|X)\right]
  \]

- \([\text{Gr"unwald, 2011}]\): pick the largest \(\eta\) such that
  \[
  \min_{p \in \mathcal{P}(\mathcal{F}(\eta))} \frac{1}{n} \sum_{i=1}^{n} -\frac{1}{\eta} \ln p(Y_i|X_i) \geq \min_{p \in \text{co}(\mathcal{P}(\mathcal{F}(\eta)))} \frac{1}{n} \sum_{i=1}^{n} -\frac{1}{\eta} \ln p(Y_i|X_i) - \text{something}
  \]
  where “something” depends on concentration inequalities and penalty function.
Summary

• **Stochastic mixability** → fast rates of convergence in different settings:
  - statistical learning (margin condition)
  - sequential prediction (mixability)

• Convexity interpretation

• Idea for adaptation to the margin
References

Slides and NIPS 2012 paper: [www.timvanerven.nl](http://www.timvanerven.nl)


- J. Li, *Estimation of Mixture Models* (PhD thesis), Yale University, 1999