TTIC 31020 Introduction to Machine Learning

Recitation Week 5

LINGYU GAO

lygao@ttic.edu

01

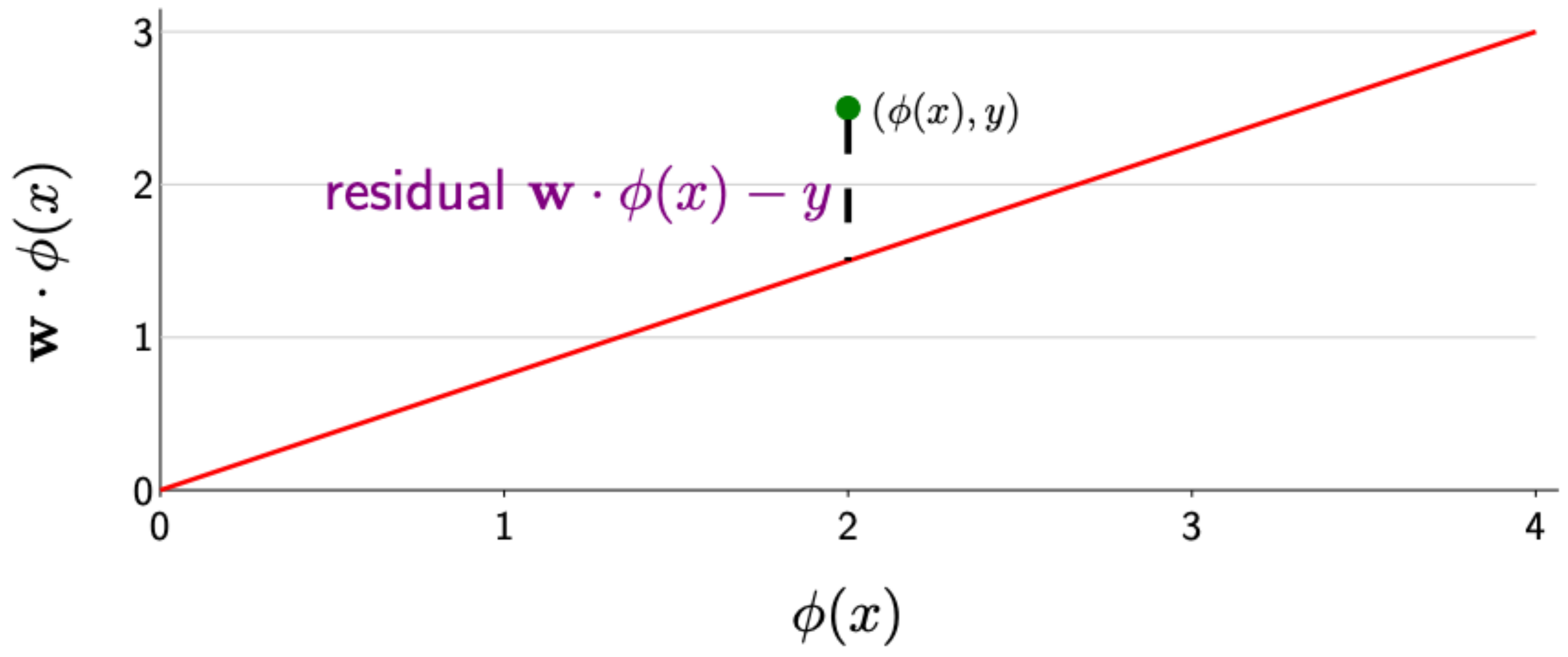
Loss Functions for Binary Classification

Some prediction tasks

Linear Regression

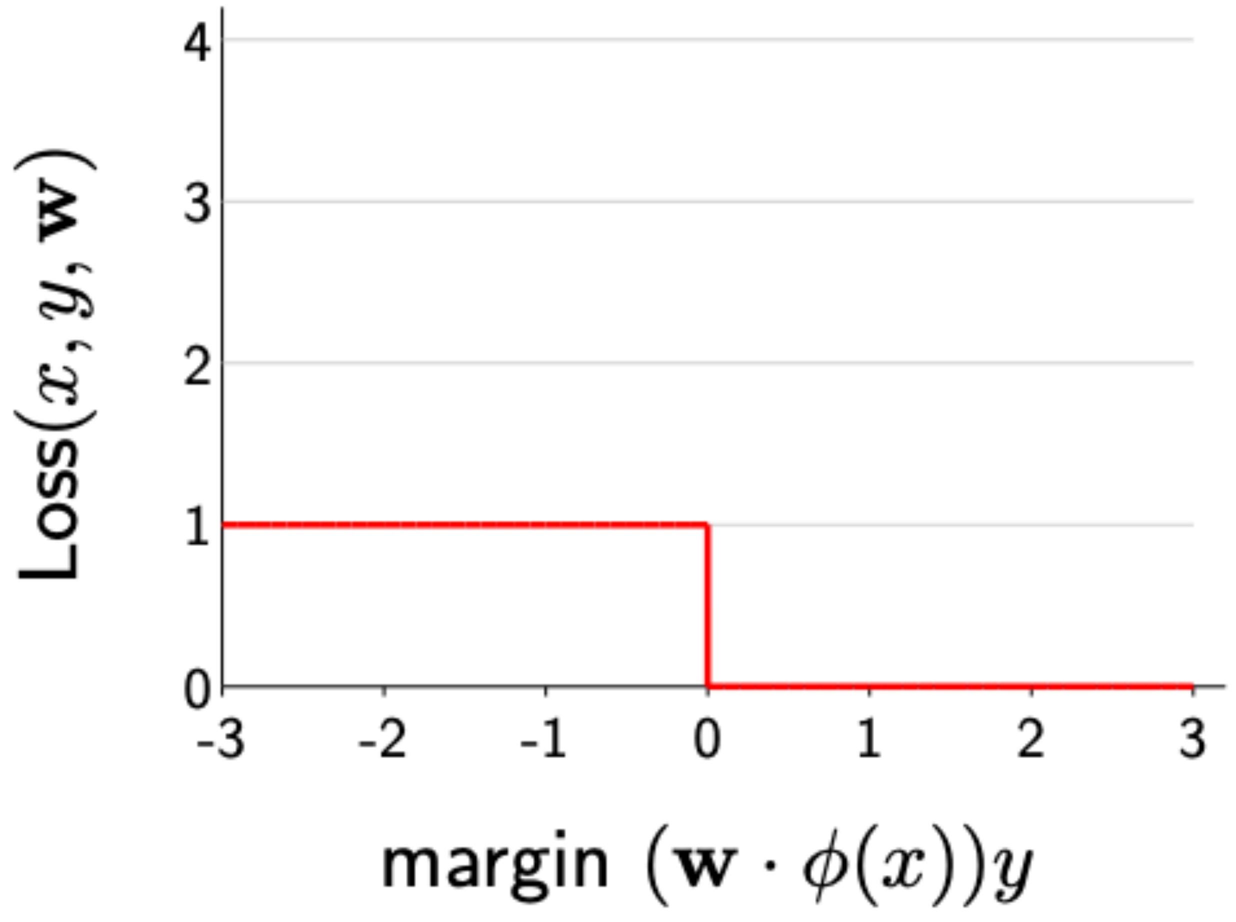
$$x \longrightarrow \boxed{f} \longrightarrow y \in \mathbb{R}$$

$$f_{\mathbf{w}}(x) = \mathbf{w} \cdot \phi(x)$$



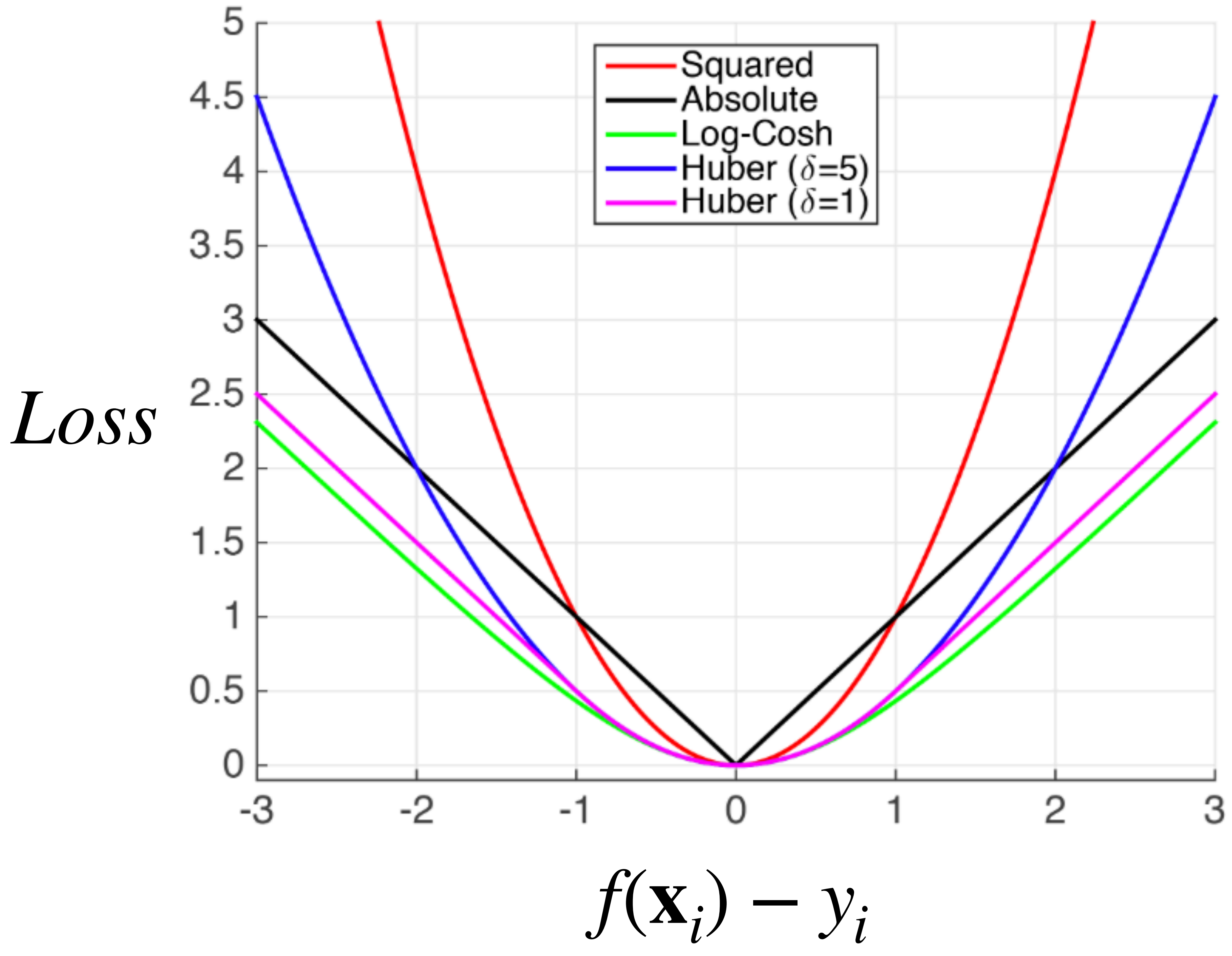
Binary Classification

$$x \longrightarrow \boxed{f} \longrightarrow y \in \{+1, -1\}$$



Loss Function → parameter estimation

Regression



Squared Loss

$$(f(\mathbf{x}_i) - y_i)^2$$

Absolute Loss

$$|f(\mathbf{x}_i) - y_i|$$

Smooth Absolute Loss (Huber Loss)

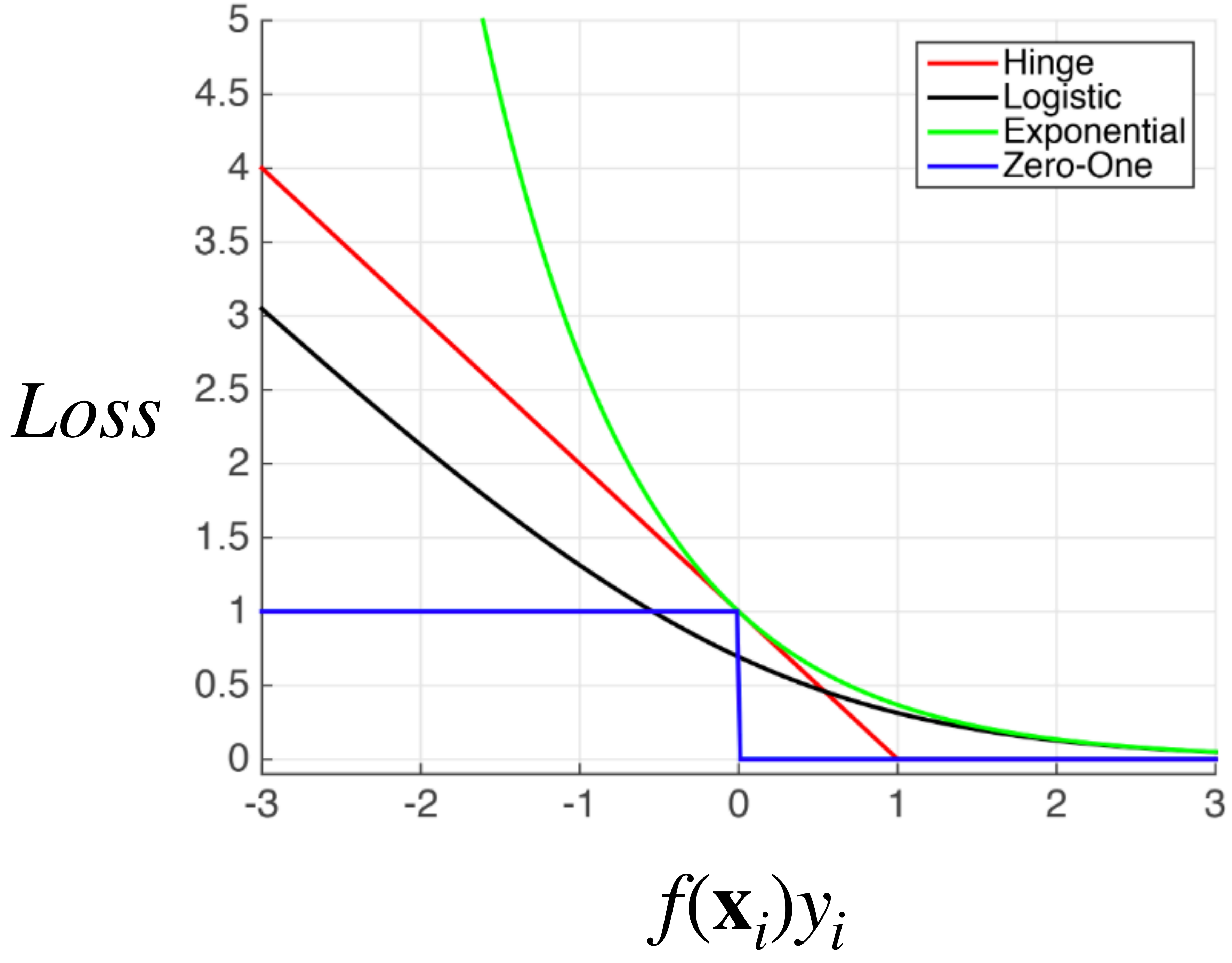
$$\frac{1}{2}\left(f(\mathbf{x}_i) - y_i\right)^2 \text{ if } |f(\mathbf{x}_i) - y_i| < \delta\text{, otherwise } \delta(|f(\mathbf{x}_i) - y_i| - \frac{\delta}{2})$$

Log-Cosh Loss

$$\log(cosh(f(\mathbf{x}_i) - y_i)), \quad cosh(x) = \frac{e^x + e^{-x}}{2}$$

Loss Function → parameter estimation

Binary Classification



Zero-one Loss $\mathbb{1}[f_{\mathbf{w}}(\mathbf{x}_i) \neq y_i]$

Exponential Loss $e^{-f_{\mathbf{w}}(\mathbf{x}_i)y_i}$

Log Loss $\log(1 + e^{-f_{\mathbf{w}}(\mathbf{x}_i)y_i})$

Hinge Loss $\max\left[1 - f_{\mathbf{w}}(\mathbf{x}_i)y_i,\ \ 0\right]^p$

02

MLE and MAP

MLE: Maximum Likelihood Estimation

Lec.3

likelihood prior

posterior $$P(\theta \mid \mathbf{X}) = \frac{P(\mathbf{X} \mid \theta) P(\theta)}{P(\mathbf{X})}$$

marginal likelihood
(normalization constant)

conditional likelihood in slides of lecture $P(\mathbf{y} \mid \mathbf{X}; \mathbf{w}, \sigma)$

MLE of Logistic Regression

Why gradient descent?

No close-formed solution

Why there's no close-formed solution?

Sigmoid function is non-linear

Is the negative log loss function convex?

Yes, the Hessian matrix is positive-definite

Question Formalization:

Gradient and Hessian of log-likelihood for logistic regression

a. Let $\sigma(a) = \frac{1}{1+e^{-a}}$ be the sigmoid function. Show that

$$\frac{d\sigma(a)}{da} = \sigma(a)(1-\sigma(a))$$

b. Using the previous result and the chain rule of calculus, derive an expression for the gradient of the log likelihood

c. The Hessian can be written as $\mathbf{H} = \mathbf{X}^T\mathbf{S}\mathbf{X}$, where $\mathbf{S} \triangleq \mathrm{diag}(\mu_1(1-\mu_1), \ldots, \mu_n(1-\mu_n))$. Show that $\mathbf{H}$ is positive definite. (You may assume that $0 < \mu_i < 1$, so the elements of $\mathbf{S}$ will be strictly positive, and that $\mathbf{X}$ is full rank.)

Logistic Regression → Multinomial Logistic Regression

$$
\begin{aligned}
g(\mathbf{w}) &= \frac{\partial}{\partial \mathbf{w}} NLL(\mathbf{w}) \\
&= \sum_{n=1}^{N} \frac{\partial}{\partial \mathbf{w}} [y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i)] \\
&= \sum_{n=1}^{N} y_i \frac{1}{\sigma} \sigma(1 - \sigma) - \mathbf{x}_i + (1 - y_i) \frac{-1}{1 - \sigma} \sigma(1 - \sigma) - \mathbf{x}_i \\
&= \sum_{n=1}^{N} (\sigma(\mathbf{w}^T \mathbf{x}_i) - y_i) \mathbf{x}_i
\end{aligned}
$$

For an arbitrary non-zero vector$\mathbf{u}$(with proper shape):

$$\mathbf{u}^T \mathbf{X}^T \mathbf{S} \mathbf{X} \mathbf{u} = (\mathbf{X}\mathbf{u})^T \mathbf{S} (\mathbf{X}\mathbf{u})$$

Since $\mathbf{S}$ is positive definite, for arbitrary non-zero $\mathbf{v}$:

$$\mathbf{v}^T \mathbf{S} \mathbf{v} > 0$$

Assume $\mathbf{X}$ is a full-rank matrix, $\mathbf{X}\mathbf{u}$ is not zero, thus:

$$(\mathbf{X}\mathbf{u})^T \mathbf{S} (\mathbf{X}\mathbf{u}) = \mathbf{u}^T (\mathbf{X}^T \mathbf{S} \mathbf{X}) \mathbf{u} > 0$$

So $\mathbf{X}^T \mathbf{S} \mathbf{X}$ is positive definite.

Second Order Conditions for Convexity

Proposition 1.29 *Let $D \subset \mathbb{R}^n$ be an open convex set and let $f : D \longrightarrow \mathbb{R}$ be twice continuously differentiable in D. Then f is convex if and only if the Hessian matrix of f is positive semidefinite throughout D.*

Proof: By Taylor's Theorem we have

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle + \frac{1}{2} \langle \boldsymbol{y} - \boldsymbol{x}, \nabla^2 f(\boldsymbol{x} + \lambda (\boldsymbol{y} - \boldsymbol{x}))(\boldsymbol{y} - \boldsymbol{x}) \rangle \, ,$$

for some $\lambda \in [0, 1]$. Clearly, if the Hessian is positive semi-definite, we have

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \nabla f(\boldsymbol{x}), \boldsymbol{y} - \boldsymbol{x} \rangle \, ,$$

which in view of the definition of the excess function, means that $E(\boldsymbol{x}, \boldsymbol{y}) \geq 0$ which implies that f is convex on D.

Conversely, suppose that the Hessian is *not* positive semi-definite at some point $\boldsymbol{x} \in D$. Then, by the continuity of the Hessian, there is a $\boldsymbol{y} \in D$ so that, for all $\lambda \in [0, 1]$,

$$\langle \boldsymbol{y} - \boldsymbol{x}, \nabla^2 f(\boldsymbol{x} + \lambda(\boldsymbol{y} - \boldsymbol{x})) \, (\boldsymbol{y} - \boldsymbol{x}) \rangle < 0 \, ,$$

which, in light of the second order Taylor expansion implies that $E(\boldsymbol{x}, \boldsymbol{y}) < 0$ and so f cannot be convex. □

Definition 1.22 *A real symmetric $n \times n$ matrix A is said to be*

(a) Positive definite *provided $\boldsymbol{x}^\top A \boldsymbol{x} > 0$ for all $x \in \mathbb{R}^n$, $\boldsymbol{x} \neq 0$.*

(b) Negative definite *provided $\boldsymbol{x}^\top A \boldsymbol{x} < 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \neq 0$.*

(c) Positive semidefinite *provided $\boldsymbol{x}^\top A \boldsymbol{x} \geq 0$ for all $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x} \neq 0$.*

(d) Negative semidefinite *provided $\boldsymbol{x}^\top A \boldsymbol{x} \leq 0$ for all $x \in \mathbb{R}^n$, $\boldsymbol{x} \neq 0$.*

(e) Indefinite *provided $\boldsymbol{x}^\top A \boldsymbol{x}$ takes on values that differ in sign.*

MAP: Maximum A Posteriori

$$\max_\theta \log p(\theta \mid \{x_i, y_i\}) = \max_\theta \log p(\theta) + \underbrace{\log p(\{x_i, y_i\} \mid \theta)}_{\text{MLE loss}}$$

We may have some belief about the value of the parameters before seeing any data

- Prior over the hypotheses: $p(\theta)$
- Posterior over the hypotheses: $p(\theta \mid \{x_i, y_i\})$
- Likelihood: $p(\{x_i, y_i\} | \theta)$

When MLE is the same with MAP?

- Bayesian rule:

$$p(\theta \mid \{x_i, y_i\}) = \frac{p(\theta)p(\{x_i, y_i\}|\theta)}{p(\{x_i, y_i\})}$$

posterior ∝ likelihood × prior

Prior is uniform!

$\log P(\theta)$ **constant**

03

Regularization and Prior

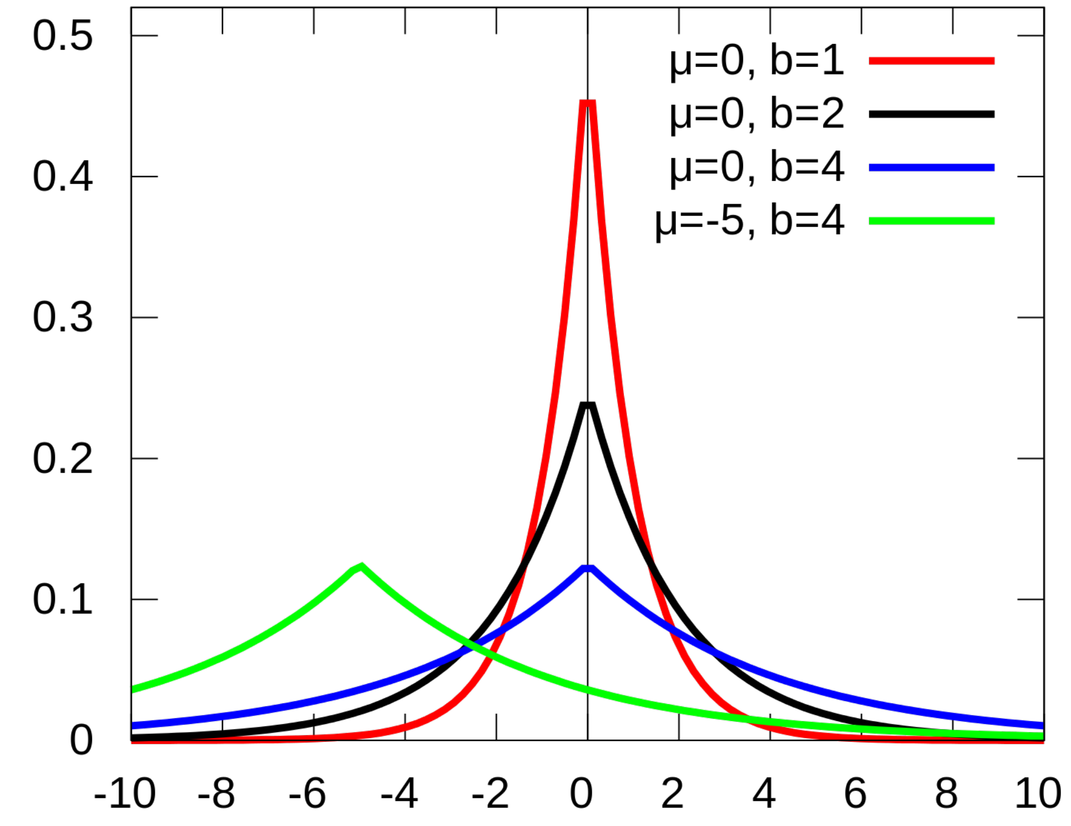
MAP: Maximum A Posteriori

$$\max_{\theta} \log p(\theta \mid \{x_i, y_i\}) = \max_{\theta} \underbrace{\log p(\theta)}_{\text{Regularization}} + \underbrace{\log p(\{x_i, y_i\} \mid \theta)}_{\text{MLE loss}}$$

$\log p(\theta)$

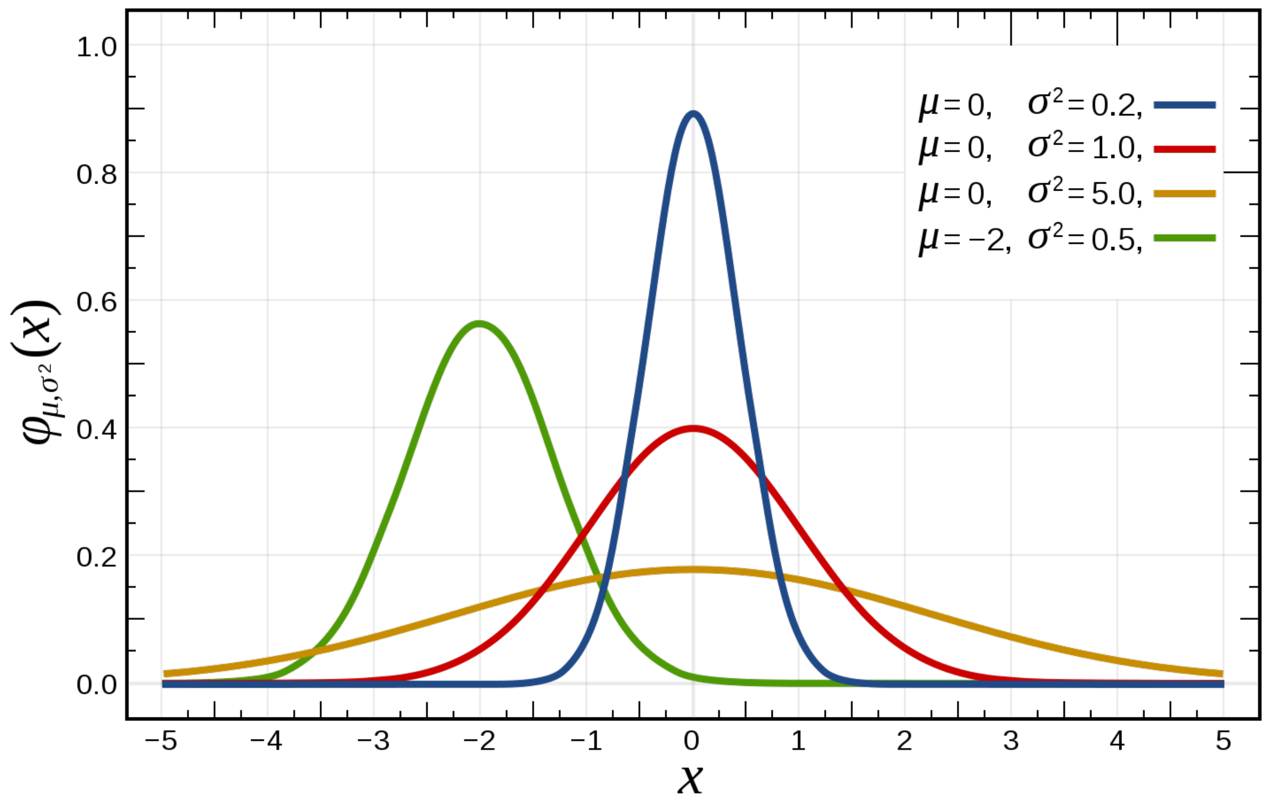
Laplace Distribution

$$f(x \mid \mu, b) = \frac{1}{2b} \exp\left(-\frac{|x-\mu|}{b}\right)$$



Gaussian Distribution

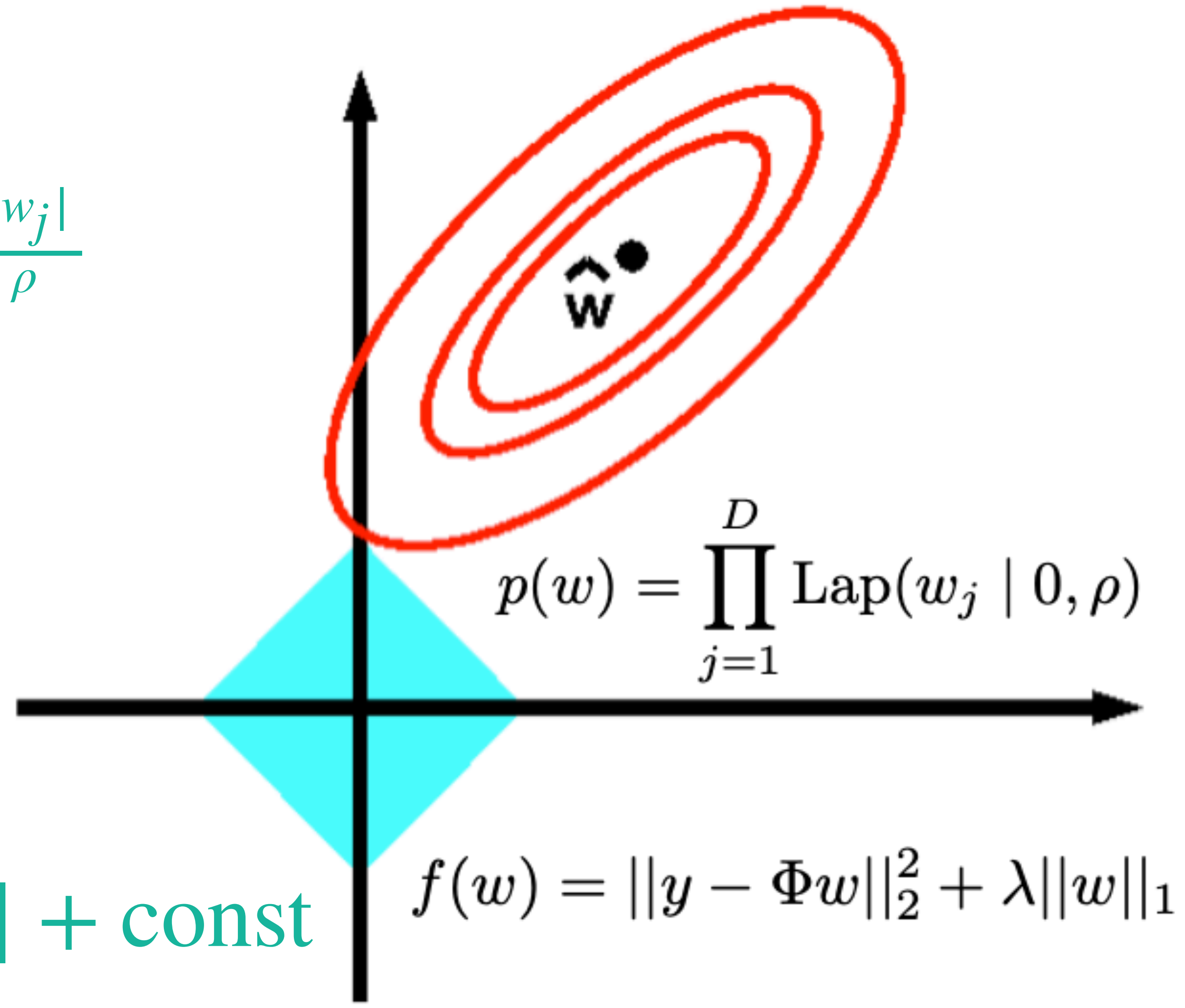
$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



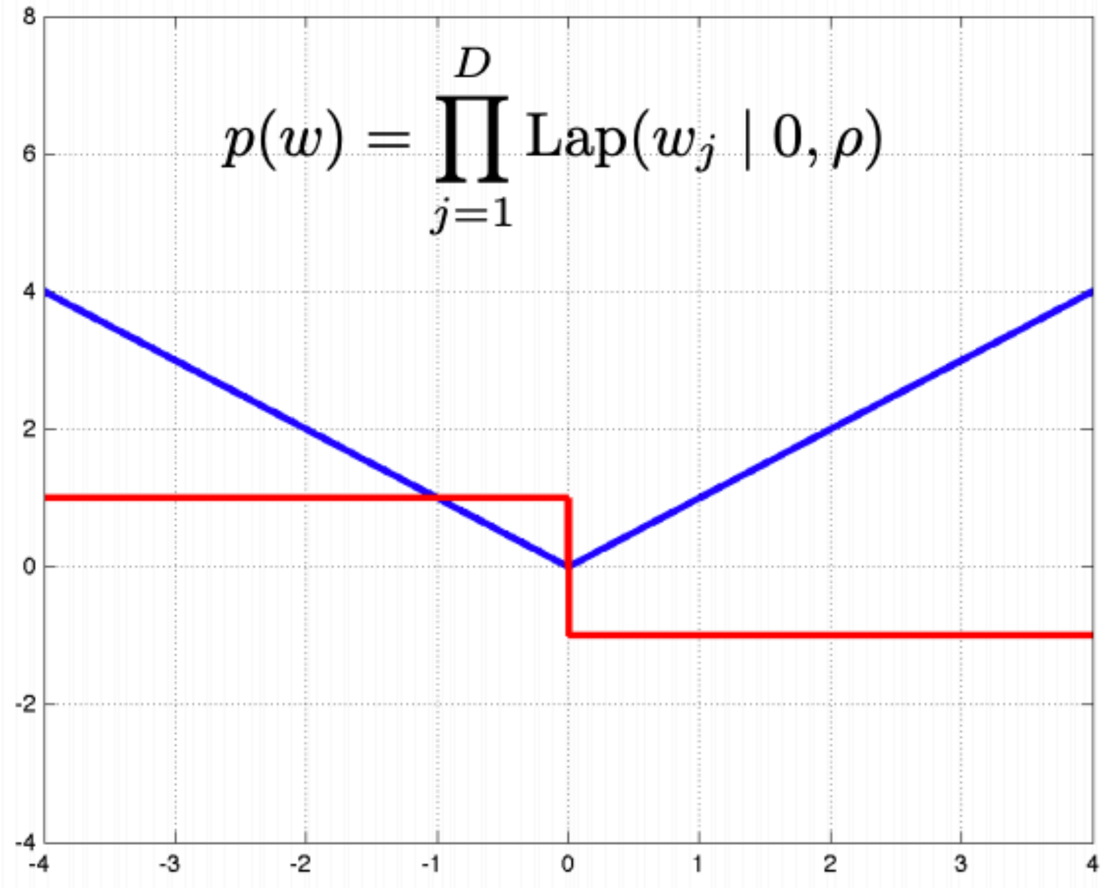
$$\max_{\theta} \log p(\theta \mid \{x_i, y_i\}) = \max_{\theta} \underbrace{\log p(\theta)}_{\text{Regularization}} + \underbrace{\log p(\{x_i, y_i\} \mid \theta)}_{\text{MLE loss}}$$

Laplacian prior
L_1 regularization
Lasso regression

$$p(w) = \prod_{j=1}^{D} \frac{1}{2\rho} e^{-\frac{|w_j|}{\rho}}$$

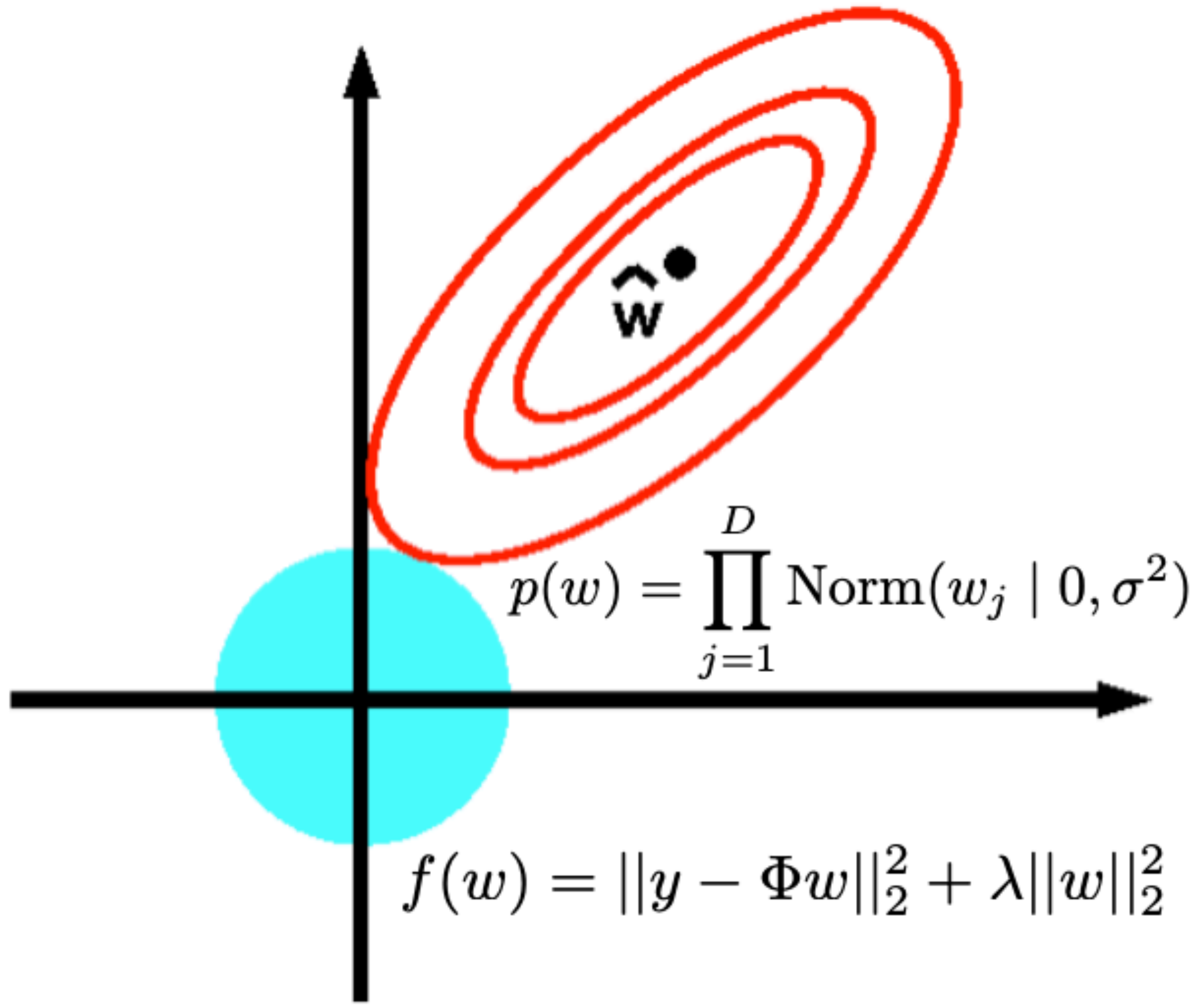


$$\log p(w) = -\frac{1}{\rho} \sum_{i=1}^{d} |w_j| + \text{const}$$

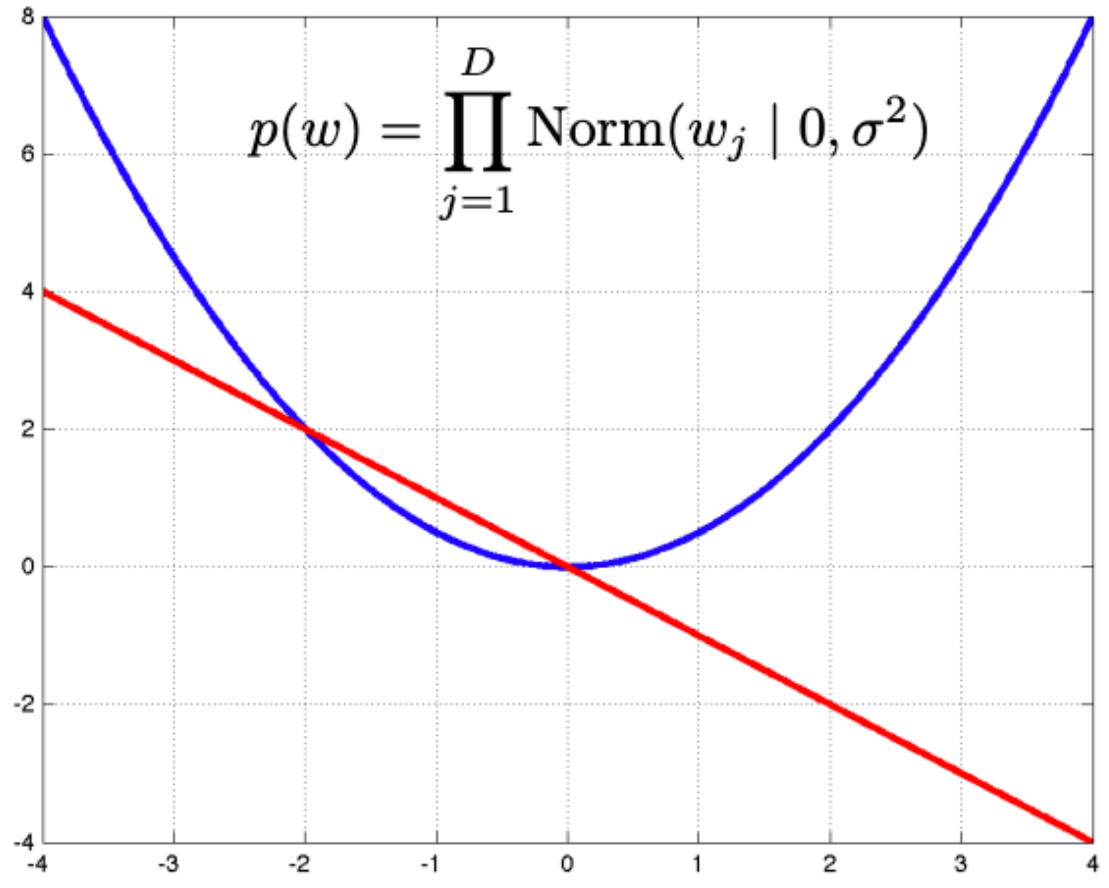


Gaussian prior
L_2 regularization
Ridge regression

$$p(w) = \prod_{j=1}^{D} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{w_j^2}{2\sigma^2}}$$



$$\log p(w) = -\frac{1}{2\sigma^2} \sum_{i=1}^{d} w_i^2 + \text{const}$$



Thank you!

Materials Used in This Slides Are Collected From Multiple Resources on the Internet.