MACHINE LEARNING IN PHYSICS
FOUNDATIONS 3

HARRISON B. PROSPER

PHY6938

Recap: Risk Functional

Taking the limit of the empirical risk function

$$R(\omega) = \frac{1}{N}\sum_{i=1}^{N} L(y_i, f_i)$$

as $N \to \infty$ yields the **risk *functional***,*

$$R[f] = \int dx \int dy\, L(y, f)\, p(x, y)$$

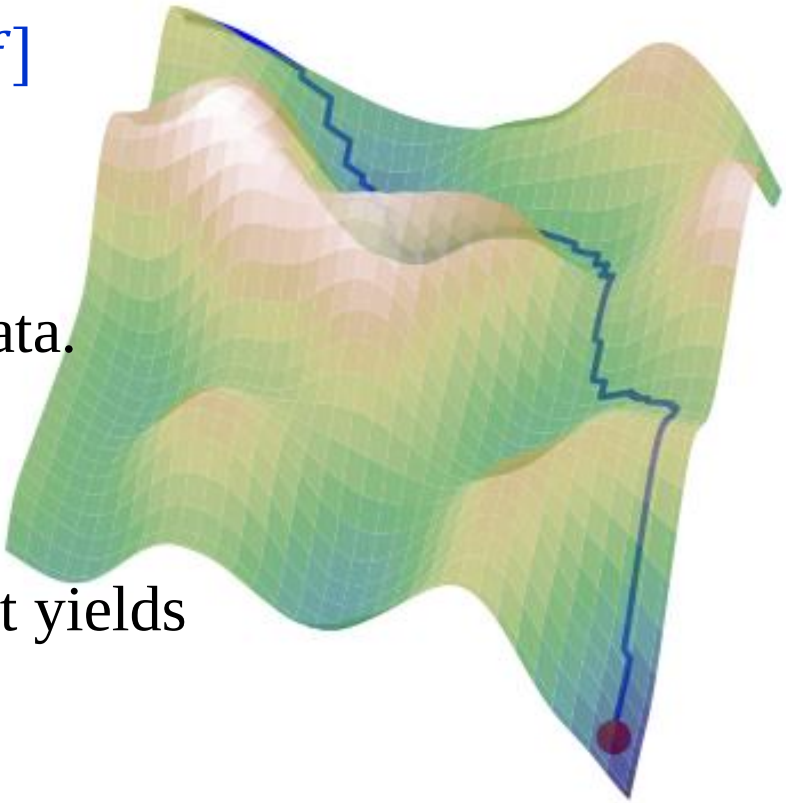
where $p(x, y)dxdy$ is the probability distribution of the data.

* A functional depends simultaneously on all the values of a function.

Recap: Risk Functional Landscape

$R(\omega)$ defines a "landscape" in the space of parameters.

The Goal: navigate to a good approximation of the lowest point of the landscape defined by $R[f]$ by navigating the landscape defined by $R(\omega)$, which, necessarily, is constructed with a *finite* amount of data.



This is what we mean when we say a model *generalizes*. The lowest point yields the best-fit function $f = f^*$.

FINDING THE BEST-FIT FUNCTION

$$f \equiv f^*$$

Finding the Best-Fit Function

Ideally, the quantity we would like to minimize is

$$R[f] = \int dx \int dy\, L(y, f)\, p(x, y)$$

to find the optimal function $f = f^*$.

We know the functional form of the loss function $L(y, f)$ because we choose it. But usually, we do not know the probability distribution, $p(x, y)dxdy$, of the data.

Nevertheless, we can still derive a very important result.

Finding the Best-Fit Function

To minimize

$$R[f] = \int dx \int dy\, L(y, f)\, p(x, y)$$

first note that $p(x, y) = p(y|x)\, p(x)$.

Therefore, we can write the functional $R[f]$ as

$$R[f] = \int dx\;\; p(x)\, \mathcal{L}(x, f)$$

where,

$$\mathcal{L}(x, f) = \int dy\, L(y, f)\, p(y|\, x)$$

Finding the Best-Fit Function

Now let's add an *arbitrary* function $\epsilon g(x)$ to the best-fit function f^*. Then

$$R[f^* + \epsilon g] = \int dx\ p(x)\, \mathcal{L}(x, f^* + \epsilon g)$$

$$\approx \int dx\ p(x) \left(\mathcal{L}(x, f^*) + \epsilon g \frac{\partial \mathcal{L}}{\partial f^*} \right)$$

$$= R[f^*] + \epsilon \int p(x) g(x) \frac{\partial \mathcal{L}}{\partial f^*}$$

Finding the Best-Fit Function

Rearranging we find

$$\frac{R[f^* + \epsilon g] - R[f^*]}{\epsilon} = \int p(x) g(x) \frac{\partial \mathcal{L}}{\partial f^*}$$

In the limit $\epsilon \to 0$, the lefthand side becomes the functional derivative $\delta R / \delta f$.

By assumption, $\delta R / \delta f$ is zero at $f = f^*$. Therefore,

$$\int p(x) g(x) \frac{\partial \mathcal{L}}{\partial f} = 0$$

when $f = f^*$.

Finding the Best-Fit Function

We want the expression

$$\int p(x) g(x) \frac{\partial \mathcal{L}}{\partial f} = 0$$

to hold for any function $g(x)$ and $\forall\, x.$

This can happen if only if

$$\frac{\partial \mathcal{L}}{\partial f} = 0$$

Assuming the integral and partial derivative operations commute, and noting that $\mathcal{L}(x, f) = \int dy\, L(y, f)\, p(y|\, x)$, we arrive at the

Very **I**mportant **R**esult:

$$\boxed{\int dy\, \frac{\partial L}{\partial f} p(y \mid x) = 0}$$

Finding the Best-Fit Function

Points to Note:

$$\boxed{\int dy \, \frac{\partial L}{\partial f} p(y \mid x) = 0}$$

1. The result is independent of the details of $f(x, \omega)$. However,…
2. The function $f(x, \omega)$ must have sufficient *capacity*: i.e., there must exist an approximation $\hat{f}(x, \widehat{\omega})$ found by minimizing $R[\omega]$ that is arbitrarily close to the optimal function f^*.
3. Moreover, it must be possible to find that function.

MINIMIZATION IN PRACTICE

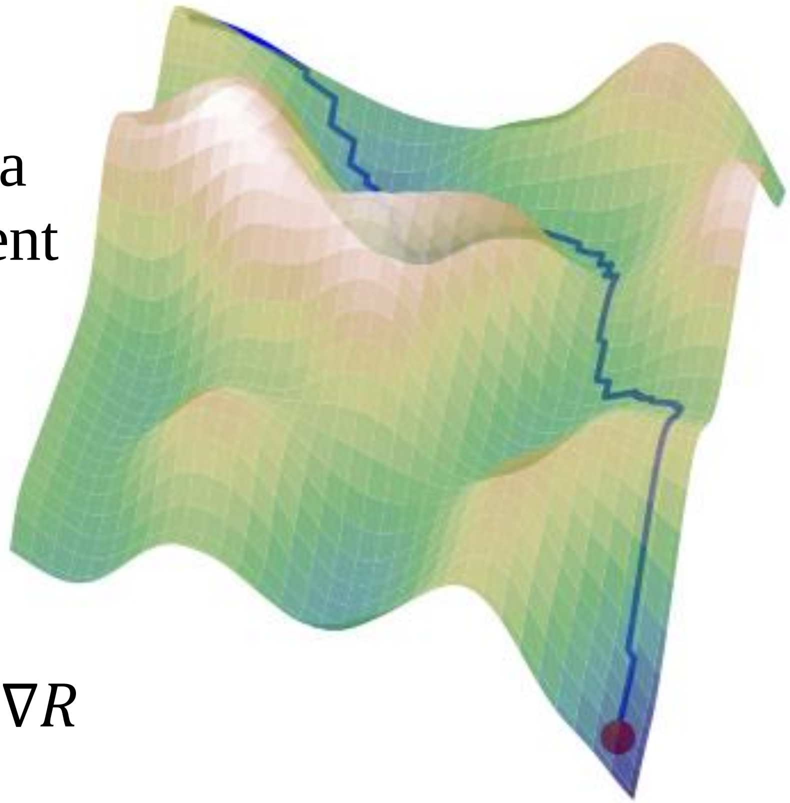
Minimization in Practice

The minimization of $R(\omega)$ is typically done by moving in the direction of *steepest descent*. The algorithms used are variations of **S**tochastic **G**radient **D**escent.

Algorithm

1. At the current point ω_j, compute a *noisy* approximation of the gradient of $R(\omega) \approx \frac{1}{n}\sum_{i=1}^{n} L(y_i, f_i)$ by using a **batch** of **training** data, where $\boldsymbol{n} \ll \boldsymbol{N}$.
2. Move to the next position ω_{j+1} in the landscape using

$$\omega_{j+1} = \omega_j - \eta \nabla R$$



Minimization in Practice

Why does the algorithm

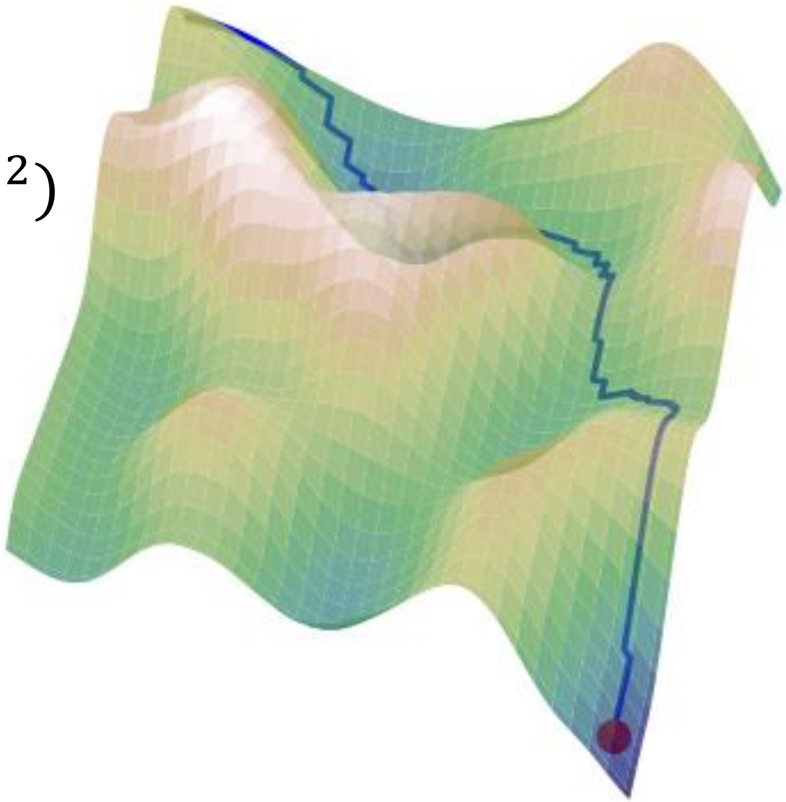
$$\omega_{j+1} = \omega_j - \eta \nabla R$$

work?

Consider

$$R(\omega_{j+1}) = R(\omega_j - \eta \nabla R)$$

$$= R(\omega_j) - \eta \nabla R \cdot \nabla R + O(\eta^2)$$



If the $O(\eta^2)$ can be neglected, and given that the $O(\eta)$ term is always negative, it follows that

$$R(\omega_{j+1}) < R(\omega_j).$$

COMMON LOSS FUNCTIONS

Common Loss Functions

Quadratic loss (typically used for regression)

$$L(y, f) = (y - f)^2$$

Binary cross entropy (typically used for classification)

$$L(y, f) = -[y \log f + (1 - y) \log(1 - f)]$$

Exponential loss

$$L(y, f) = \exp(-wyf/2)$$

Quantile loss ($0 \leq \tau \leq 1$)

$$L(y, f) = \begin{cases} \tau(y - f) & y \geq f \\ (1 - \tau)(f - y) & y < f \end{cases}$$

Common Loss Functions

Quadratic loss: $L(y, f) = (y - f)^2$

$$\int \frac{\partial L}{\partial f}\, p(y|x)\, dy = 0$$

Solution

$$\boxed{f(x, \omega^*) = \int y\, p(y \mid x)\, dy}$$

Very Important Point (VIP): As noted, the result is independent of the details of f. The result depends solely on the form of the loss function and the probability distribution, $p(x, y)$, of the data.

Common Loss Functions

Binary cross entropy loss:

$$L(y, f) = -[y \log f + (1-y)\log(1-f)]$$

$$\int \frac{\partial L}{\partial f}\, p(y|x)dy = 0$$

Solution

$$\boxed{f(x, \omega^*) = p(y = 1 \mid x) = \frac{p(x|y=1)\epsilon}{p(x|y=1)\epsilon + p(x|y=0)}}$$

where $y \in [0,1]$ and $\epsilon = \frac{\pi(y=1)}{\pi(y=0)}$ is the ratio of data sample sizes for the two classes of objects labeled by $y \in [0,1]$.

Common Loss Functions

Exponential loss:

$$L(y, f) = \exp(-wyf/2)$$

$$\int \frac{\partial L}{\partial f} p(y|x)\, dy = 0$$

Solution

$$\boxed{f(x, \omega^*) = \frac{1}{w} \log\left(\frac{p(x|y=1)}{p(x|y=-1)} \epsilon\right)}$$

where $y \in [-1, 1]$ and $\epsilon = \frac{\pi(y=1)}{\pi(y=-1)}$ is the ratio of data sample sizes for the two classes of objects labeled by $y \in [-1, 1]$.

Summary

Supervised Learning

- Given a data set $D = \{(x, y)\}_{i=1}^{N}$, a model $f(x, \omega)$, and a loss function $L(y, f)$, the optimal function $f^* = f(x, \omega^*)$ satisfies:

$$\int dy \, \frac{\partial L}{\partial f} p(y \mid x) = 0$$

Stochastic Gradient Descent

- This is the method of choice for minimizing the empirical risk $R(\omega)$:

$$\omega_{j+1} = \omega_j - \eta \nabla R$$

- Batches of data are used, which introduces noise into ∇R.