Risk: % you will get it wrong. Expected value of loss.

$\log p_\theta(y|x) = \mathcal{N}(f_\theta(x), \Sigma_\theta(x)) = -\frac{1}{2}(f_\theta(x)-y)\Sigma_\theta(x)^{-1}(f_\theta(x)-y) - \frac{1}{2}\log|\Sigma_\theta(x)| + \text{const}$

$(L(\theta, x, y))$ if $\Sigma_\theta(x) = I \rightarrow -\frac{1}{2}\|f_\theta(x)-y\|^2 + \text{const} \approx MSE$

overfitting - low emp. risk, high risk

underfitting - high emp. risk, high risk

$p(\mathcal{D}) = \prod_{i=1}^{N} p(x_i)p(y_i|x_i)$

expected value of error w.r.t. data dist (theoretical): $\mathbb{E}_{\mathcal{D}\sim p(\mathcal{D})}[\|f_\mathcal{D}(x) - f(x)\|^2]$

"how wrong will our predictions be, on average." $= \sum_\mathcal{D} p(\mathcal{D})\|f_\mathcal{D}(x) - f(x)\|^2$

Let $\bar{f}(x) = \mathbb{E}_{\mathcal{D}\sim p(\mathcal{D})}[f_\mathcal{D}(x)]$, $\mathbb{E}_{\mathcal{D}\sim p(\mathcal{D})}[\|f_\mathcal{D}(x) - f(x)\|^2]$

$= \mathbb{E}_{\mathcal{D}\sim p(\mathcal{D})}[\|(f_\mathcal{D}(x) - \bar{f}(x)) + (\bar{f}(x) - f(x))\|^2]$

$= \underbrace{\mathbb{E}_{\mathcal{D}\sim p(\mathcal{D})}[\|f_\mathcal{D}(x) - \bar{f}(x)\|^2]}_{\text{Variance}} + \underbrace{\mathbb{E}_{\mathcal{D}\sim p(\mathcal{D})}[\|\bar{f}(x) - f(x)\|^2]}_{\|\bar{f}(x) - f(x)\|^2 \rightarrow \text{Bias}^2}$

"how much does our prediction change w/ this dataset"

error never goes away.

"Bias + Variance Tradeoff"

Regularization: something we add to the loss function to reduce variance.

Bayesian perspective: Given $\mathcal{D}$, what is most likely θ?

$p(\theta|\mathcal{D}) = \frac{p(\theta,\mathcal{D})}{p(\mathcal{D})} \propto p(\theta,\mathcal{D}) = p(\mathcal{D}|\theta)p(\theta)$ ← prior, how likely θ is before $\mathcal{D}$.

New loss: $-\left(\sum_{i=1}^{N}\log p(y_i|x_i;\theta)\right) - \log p(\theta)$ ← choose this

Regularized Lin. Reg: $\log p(\theta) = -\lambda\|\theta\|^2 + \text{const}$ $\left(\lambda = \frac{1}{2\sigma^2}\right)$, same for Reg. Log. Reg. (based on $\mathcal{N}(0,\sigma^2)$)

L1 Reg. MAE - sparsity

L2 Reg. MSE - smooth

Dropout: used for NNs

Numerical perspective: regularizer makes underdetermined problems well-determined.

Optimization Perspective: regularizer makes loss landscape easier to search.

Regularizer introduce hyperparameters that we have to select for success

ML Workflow

Train	- optimize to find θ, hyperparams.
Valid.	- tune hyperparameters, select model class, which features to use
Test	- "final exam", report final performance.