James-Stein's Estimator and Stein's Paradox

John Meir Moussaffi
Department of Mathematics
Bar Ilan University

Introduction

Let $X = (X_1, \ldots, X_d) \sim N_d(\theta, \sigma^2 \mathcal{I}_d)$
$\theta = (\theta_1, \ldots, \theta_d)$ - unknown mean of X
$\sigma^2 \mathcal{I}_d$ - known (isotropic) covariance.

Problem: estimate θ based on observations $x = (x_1, \ldots, x_d)$ of X (assuming IID).

Naive approach (MLE):
calculate MSE of each dimension independently, i.e., $\hat{\theta}_i = x_i$.

$$\hat{\theta} = x$$

Definitions

An estimator $\hat{\theta}$ is **inadmissible** if there exists another estimator $\tilde{\theta}$ that dominates it. Or in other words:

$$\forall \theta, \quad R(\tilde{\theta}, \theta) \leq R(\hat{\theta}, \theta)$$

with strict inequality for some θ.
We'll use the mean squared error (MSE) for the risk.

$$R(\hat{\theta}, \theta) = \mathbb{E}_\theta \left[\left\| \hat{\theta}(X) - \theta \right\|^2 \right] = \mathrm{MSE}(\hat{\theta})$$

Stein's Estimator (& a bit of history)

In 1956, Charles Stein proved that the (then) usual estimator X is **inadmissible** when $d \geq 3$, under squared error loss.

James and Stein (1961) proposed a shrinkage estimator that strictly dominates it:

$$\hat{\theta}_{\mathrm{JS}} = \left(1 - \frac{(d-2)\sigma^2}{\|x\|^2}\right) x$$

Stein's Paradox (& Astrology)

Say we have observations of the movement of the stars, people's personality traits, height of people in Dublin, and wall street's stocks. James and Stein show that we could better estimate one of these using observations of another.

Seems almost trivial that admissibility would follow from x being IID, and it is true for $d = 1, 2$, but why not for $d \geq 3$?

Bias–Variance Tradeoff

Principle: introducing bias lets us control variance (lower total error).
Coupling: shrinkage uses the whole vector x to estimate each θ_i.

$$\underbrace{\mathbb{E}_\theta\left[\left\|\hat{\theta}(X)-\theta\right\|^2\right]}_{\textbf{MSE}} = \underbrace{\mathbb{E}_\theta\left[\left\|\hat{\theta}(X)-\mathbb{E}_\theta[\hat{\theta}(X)]\right\|^2\right]}_{\textbf{Variance}} + \underbrace{\left\|\mathbb{E}_\theta[\hat{\theta}(X)]-\theta\right\|^2}_{\textbf{Bias}^2}.$$

MLE: $\hat{\theta}=X \quad\Rightarrow\quad \mathbb{E}_\theta[\hat{\theta}]=\theta$ ($Bias^2=0$), $Var = d\sigma^2$.

James–Stein: $\hat{\theta}_{\mathrm{JS}} = \left(1-\dfrac{(d-2)\sigma^2}{\|X\|^2}\right)X$

$\Rightarrow \quad Bias > 0,\ Var = R(\hat{\theta}^{\mathrm{JS}},\theta) = p-(p-2)\cdot\mathbb{E}\left[\frac{1}{\|X\|^2}\right] < p$

From Shrinkage to Regularization

Shrinkage is essential in neural networks—it reins in weights, curbs overfitting, and stabilizes generalization. The idea is to bias estimates towards a target to reduce variance.

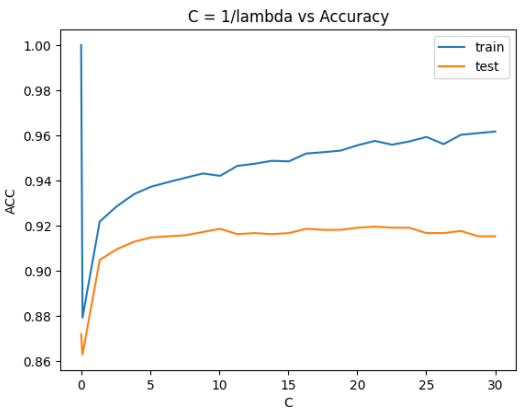
$$\min_{\theta} \; \frac{1}{n} \sum_{i=1}^{n} \ell\big(y_i, f_\theta(x_i)\big) \; + \; \lambda \, \Omega(\theta)$$

- **Ridge**: $\Omega(\theta) = \|\theta\|_2^2$ (*uniform shrink* toward 0)
- **Lasso**: $\Omega(\theta) = \|\theta\|_1$ (*sparse shrink*; some coefficients $\to 0$)
- and many more...

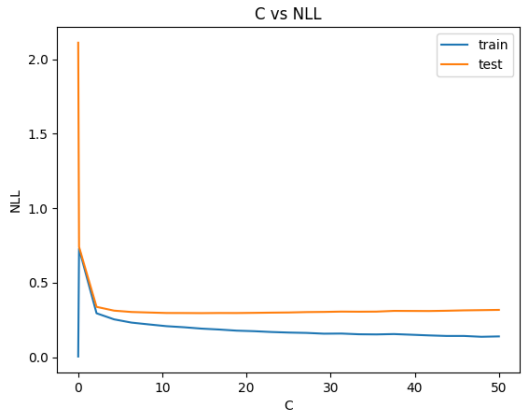
Example Application: Weight Decay

Weight Decay is a regularization method in neural networks, equivalent to ridge regression on the weights.
I applied weight decay to a neural network trained on the MNIST dataset to demonstrate how it improves generalization.
By comparing scores across different values of $C = 1/\lambda$, we can observed the impact of shrinkage on performance.

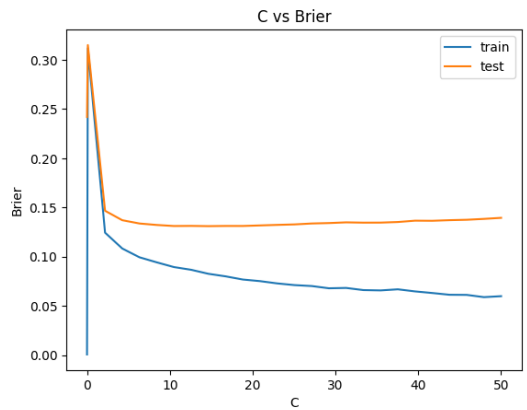
Model Applications



C vs ACC



C vs NLL



C vs Brier

Observation:
Visualization of how C (the shrinkage factor) affects the results, given by the metrics NLL, Brier, and Accuracy. This shows the improvements that shrinkage introduces. Also, a cool thing we can see here is the phenomenon of overfitting (diminishing returns in regularization).

END

Also here's the link for the colab notebook plus references if you're interested: https://colab.research.google.com/drive/1rNbf27Z5ZFcRj-kYDGa4uhxsC5DHnXff)