Stability of Tikhonov Regularization

9.520 Class 07, March 2003

Alex Rakhlin

Plan

- Review of Stability Bounds
- Stability of Tikhonov Regularization Algorithms

Uniform Stability

Review notation: $S = \{z_1, ..., z_\ell\}$; $S^{i,z} = \{z_1, ..., z_{i-1}, z, z_{i+1}, ..., z_\ell\}$
$c(f, z) = V(f(\mathbf{x}), y)$, where $z = (\mathbf{x}, y)$.

An algorithm $\mathcal{A}$ has **uniform stability** β if

$$\forall (S, z) \in \mathcal{Z}^{\ell+1},\ \forall i,\ \sup_{u \in \mathcal{Z}} |c(f_S, u) - c(f_{S^{i,z}}, u)| \leq \beta.$$

Last class: Uniform stability of $\beta = O\left(\frac{1}{\ell}\right)$ implies good generalization bounds.

This class: Tikhonov Regularization has uniform stability of $\beta = O\left(\frac{1}{\ell}\right)$.

Reminder: The Tikhonov Regularization algorithm:

$$f_S = \arg\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} V(f(x_i), y_i) + \lambda \|f\|_K^2$$

Generalization Bounds Via Uniform Stability

If $\beta = \frac{k}{\ell}$ for some k, we have the following bounds from the last lecture:

$$P\left(|I[f_S] - I_S[f_S]| \geq \frac{k}{\ell} + \epsilon\right) \leq 2\exp\left(-\frac{\ell\epsilon^2}{2(k+M)^2}\right).$$

Equivalently, with probability $1-\delta$,

$$I[f_S] \leq I_S[f_S] + \frac{k}{\ell} + (2k+M)\sqrt{\frac{2\ln(2/\delta)}{\ell}}.$$

Lipschitz Loss Functions, I

We say that a loss function (over a possibly bounded domain $\mathcal{X}$) is Lipschitz with Lipschitz constant L if

$$\forall y_1, y_2, y' \in \mathcal{Y},\ |V(y_1, y') - V(y_2, y')| \leq L|y_1 - y_2|.$$

Put differently, if we have two functions f_1 and f_2, under an L-Lipschitz loss function,

$$\sup_{(\mathbf{x},y)} |V(f_1(\mathbf{x}), y) - V(f_2(\mathbf{x}), y)| \leq L|f_1 - f_2|_\infty.$$

Yet another way to write it:

$$|c(f_1, \cdot) - c(f_2, \cdot)|_\infty \leq L|f_1(\cdot) - f_2(\cdot)|_\infty$$

Lipschitz Loss Functions, II

If a loss function is L-Lipschitz, then closeness of two functions (in L_∞ norm) implies that they are close in loss.

The converse is false — it is possible for the difference in loss of two functions to be small, yet the functions to be far apart (in L_∞). Example: constant loss.

The hinge loss and the ϵ-insensitive loss are both L-Lipschitz with $L = 1$. The square loss function is L Lipschitz if we can bound the y values and the $f(x)$ values generated. The $0-1$ loss function is not L-Lipschitz at all — an arbitrarily small change in the function can change the loss by 1:

$$f_1 = 0,\ f_2 = \epsilon,\ V(f_1(x), 0) = 0,\ V(f_2(x), 0) = 1.$$

Lipschitz Loss Functions for Stability

Assuming L-Lipschitz loss, we transformed a problem of bounding

$$\sup_{u \in \mathcal{Z}} |c(f_S, u) - c(f_{S^{i,z}}, u)|$$

into a problem of bounding $|f_S - f_{S^{i,z}}|_\infty$.

As the next step, we bound the above L_∞ norm by the norm in the RKHS assosiated with a kernel K.

For our derivations, we need to make another assumption: there exists a κ satisfying

$$\forall \mathbf{x} \in \mathcal{X},\ \sqrt{K(\mathbf{x}, \mathbf{x})} \leq \kappa.$$

Relationship Between L_∞ and L_K

Using the reproducing property and the Cauchy-Schwartz inequality, we can derive the following:

$$
\begin{aligned}
\forall \mathbf{x}\ |f(\mathbf{x})| &= |\langle K(\mathbf{x},\cdot), f(\cdot)\rangle_K| \\
&\leq ||K(\mathbf{x},\cdot)||_K ||f||_K \\
&= \sqrt{\langle K(x,\cdot), K(x,\cdot)\rangle} ||f||_K \\
&= \sqrt{K(\mathbf{x},\mathbf{x})} ||f||_K \\
&\leq \kappa ||f||_K
\end{aligned}
$$

Since above inequality holds for all $\mathbf{x}$, we have $|f|_\infty \leq ||f||_K$.

Hence, if we can bound the RKHS norm, we can bound the L_∞ norm. Note that the converse is not true.

Note that we now transformed the problem to bounding $||f_S - f_{S^{i,z}}||_K$.

A Key Lemma

We will prove the following lemma about **Tikhonov regularization**:

$$||f_S - f_{S^{i,z}}||_K^2 \leq \frac{L|f_S - f_{S^{i,z}}|_\infty}{\lambda \ell}$$

This theorem says that when we replace a point in the training set, the change in the RKHS norm (squared) of the difference between the two functions cannot be too large compared to the change in L_∞.

We will first explore the implications of this lemma, and defer its proof until later.

Bounding β, I

Using our lemma and the relation between L_K and L_∞,

$$\begin{aligned} ||f_S - f_{S^{i,z}}||_K^2 &\leq \frac{L|f_S - f_{S^{i,z}}|_\infty}{\lambda \ell} \\ &\leq \frac{L\kappa ||f_S - f_{S^{i,z}}||_K}{\lambda \ell} \end{aligned}$$

Dividing through by $||f_S - f_{S^{i,z}}||_K$, we derive

$$||f_S - f_{S^{i,z}}||_K \leq \frac{\kappa L}{\lambda \ell}.$$

Bounding β, II

Using again the relationship between L_K and L_∞, and the L Lipschitz condition,

$$\begin{aligned} \sup|V(f_S(\cdot),\cdot) - V(f_{S^{z,i}}(\cdot),\cdot)| &\leq L|f_S - f_{S^{z,i}}|_\infty \\ &\leq L\kappa||f_S - f_{S^{z,i}}||_K \\ &\leq \frac{L^2\kappa^2}{\lambda\ell} \\ &= \beta \end{aligned}$$

Divergences

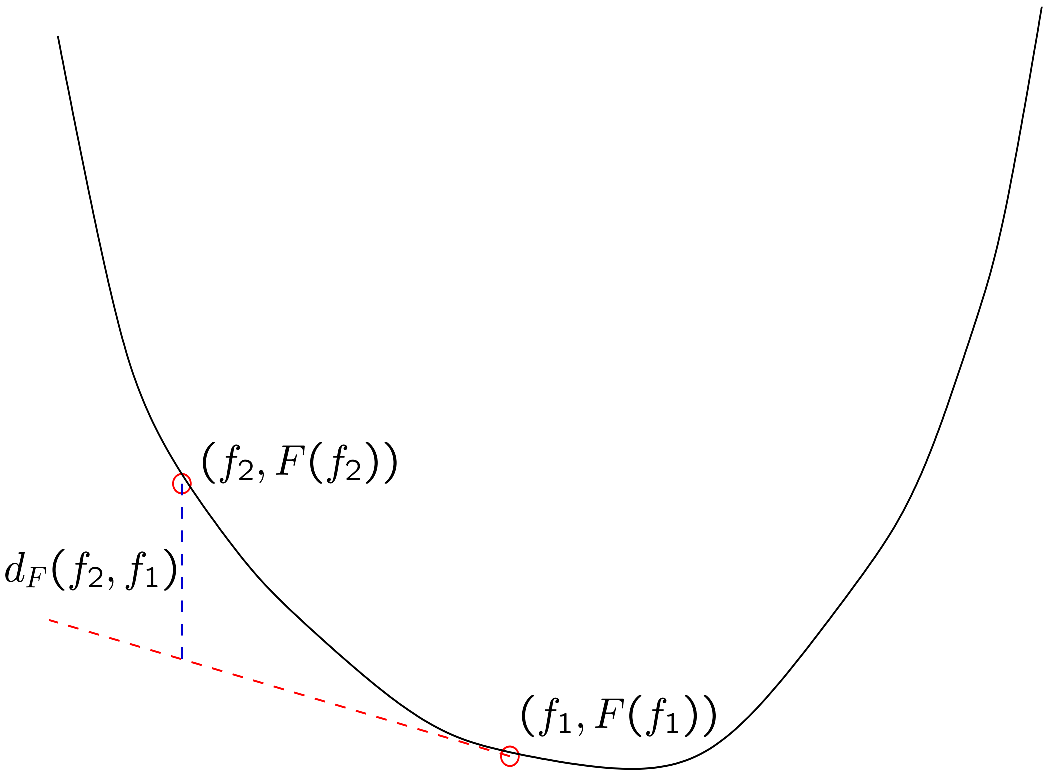
Suppose we have a convex, differentiable function F, and we know $F(f_1)$ for some f_1. We can "guess" $F(f_2)$ by considering a linear approximation to F at f_1:

$$\widehat{F}(f_2) = F(f_1) + \langle f_2 - f_1, \nabla F(f_1) \rangle.$$

The Bregman divergence is the error in this linearized approximation:

$$d_F(f_2, f_1) = F(f_2) - F(f_1) - \langle f_2 - f_1, \nabla F(f_1) \rangle.$$

Divergences Illustrated



Divergences Cont'd

We will need the following key facts about divergences:

- $d_F(f_2, f_1) \geq 0$
- If f_1 minimizes F, then the gradient is zero, and $d_F(f_2, f_1) = F(f_2) - F(f_1)$.
- If $F = A + B$, where A and B are also convex and differentiable, then $d_F(f_2, f_1) = d_A(f_2, f_1) + d_B(f_2, f_1)$ (the derivatives add).

The Tikhonov Functionals

We shall consider the Tikhonov functional

$$T_S(f) = \frac{1}{\ell}\sum_{i=1}^{\ell} V(f(\mathbf{x_i}), y_i) + \lambda ||f||_K^2,$$

as well as the component functionals

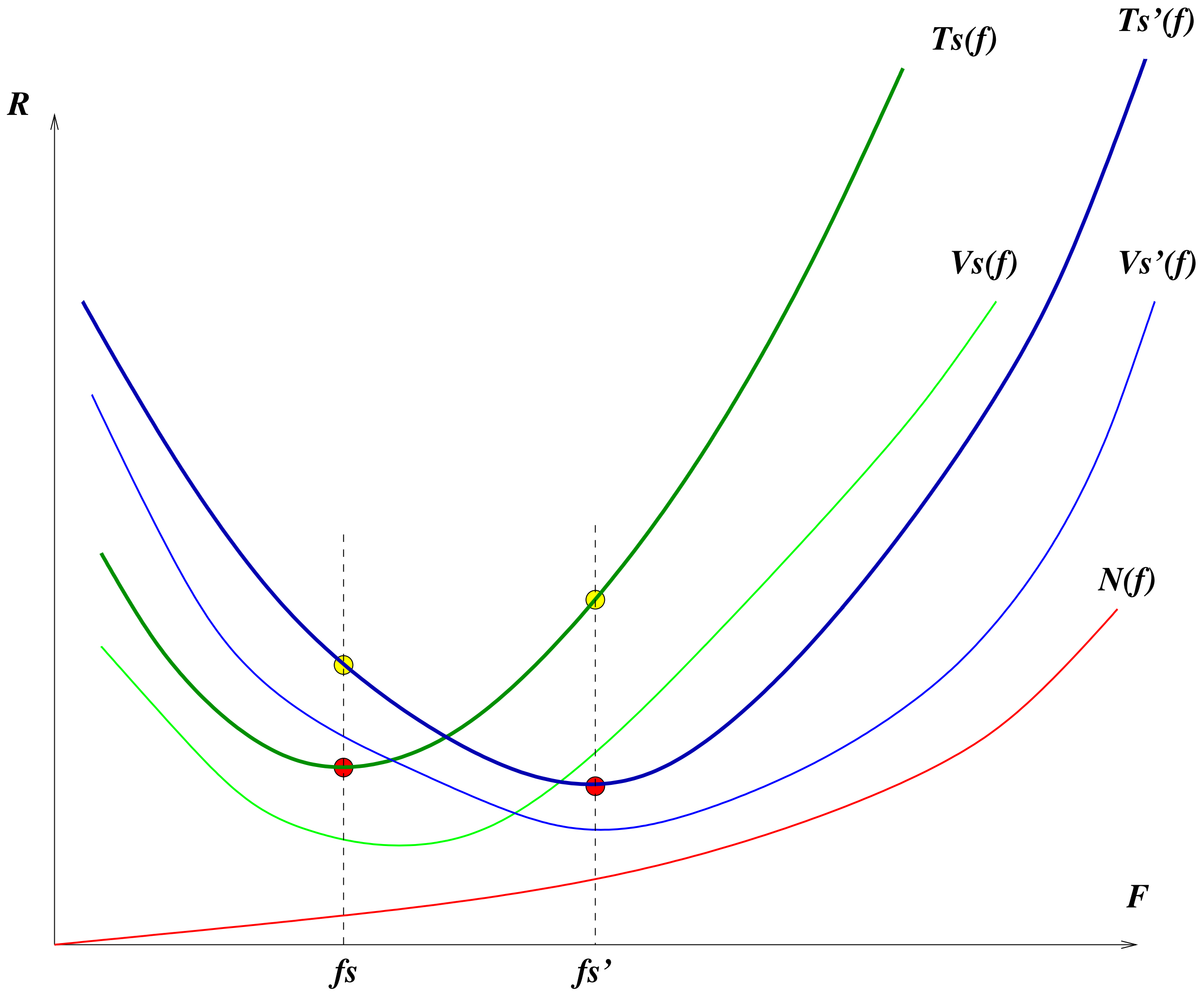
$$V_S(f) = \frac{1}{\ell}\sum_{i=1}^{\ell} V(f(\mathbf{x_i}), y_i)$$

and

$$N(f) = ||f||_K^2.$$

Hence, $T_S(f) = V_S(f) + \lambda N(f)$. If the loss function is convex (in the first variable), then all three functionals are convex.

A Picture of Tikhonov Regularization



Proving the Lemma, I

Let f_S be the minimizer of T_S, and let $f_{S^{i,z}}$ be the minimizer of $T_{S^{i,z}}$, the perturbed data set with $(\mathbf{x}_i, y_i)$ replaced by a new point $z = (\mathbf{x}, y)$. Then

$$\begin{aligned}
\lambda(d_N(f_{S^{i,z}}, f_S) + d_N(f_S, f_{S^{i,z}})) &\leq \\
d_{T_S}(f_{S^{i,z}}, f_S) + d_{T_{S^{i,z}}}(f_S, f_{S^{i,z}}) &= \\
\frac{1}{\ell}(c(f_{S^{i,z}}, z_i) - c(f_S, z_i) + c(f_S, z) - c(f_{S^{i,z}}, z)) &\leq \\
\frac{2L|f_S - f_{S^{i,z}}|_\infty}{\ell}.
\end{aligned}$$

We conclude that

$$d_N(f_{S^{i,z}}, f_S) + d_N(f_S, f_{S^{i,z}}) \leq \frac{2L|f_S - f_{S^{i,z}}|_\infty}{\lambda\ell}$$

Proving the Lemma, II

But what is $d_N(f_{S^{i,z}}, f_S)$?

We will express our functions as the sum of orthogonal eigenfunctions in the RKHS:

$$\begin{aligned} f_S(\mathbf{x}) &= \sum_{n=1}^{\infty} c_n \phi_n(\mathbf{x}) \\ f_{S^{i,z}}(\mathbf{x}) &= \sum_{n=1}^{\infty} c'_n \phi_n(\mathbf{x}) \end{aligned}$$

Once we express a function in this form, we recall that

$$||f||_K^2 = \sum_{n=1}^{\infty} \frac{c_n^2}{\lambda_n}$$

Proving the Lemma, III

Using this notation, we reexpress the divergence in terms of the c_i and c'_i:

$$\begin{aligned}
d_N(f_{S^{i,z}}, f_S) &= ||f_{S^{i,z}}||_K^2 - ||f_S||_K^2 - \langle f_{S^{i,z}} - f_S, \nabla ||f_S||_K^2 \rangle \\
&= \sum_{n=1}^{\infty} \frac{c_n'^2}{\lambda_n} - \sum_{n=1}^{\infty} \frac{c_n^2}{\lambda_n} - \sum_{i=1}^{\infty} (c_n' - c_n)(\frac{2c_n}{\lambda_n}) \\
&= \sum_{n=1}^{\infty} \frac{c_n'^2 + c_n^2 - 2c_n' c_n}{\lambda_n} \\
&= \sum_{n=1}^{\infty} \frac{(c_n' - c_n)^2}{\lambda_n} \\
&= ||f_{S^{i,z}} - f_S||_K^2
\end{aligned}$$

We conclude that

$$d_N(f_{S^{i,z}}, f_S) + d_N(f_S, f_{S^{i,z}}) = 2||f_{S^{i,z}} - f_S||_K^2$$

Proving the Lemma, IV

Combining these results proves our Lemma:

$$
\begin{aligned}
||f_{S^{i,z}} - f_S||_K^2 &= \frac{d_N(f_{S^{i,z}}, f_S) + d_N(f_S, f_{S^{i,z}})}{2} \\
&\leq \frac{2L|f_S - f_{S^{i,z}}|_\infty}{\lambda \ell}
\end{aligned}
$$

Bounding the Loss, I

We have shown that Tikhonov regularization with an L-Lipschitz loss is β-stable with $\beta = \frac{L^2\kappa^2}{\lambda\ell}$. If we want to actually apply the theorems and get the generalization bound, we need to bound the loss.

Let C_0 be the maximum value of the loss when we predict a value of zero. If we have bounds on $\mathcal{X}$ and $\mathcal{Y}$, we can find C_0.

Bounding the Loss, II

Noting that the "all 0" function $\vec{\mathbf{0}}$ is always in the RKHS, we see that

$$\begin{aligned} \lambda ||f_S||_K^2 &\leq T(f_S) \\ &\leq T(\vec{\mathbf{0}}) \\ &= \frac{1}{\ell} \sum_{i=1}^{\ell} V(\vec{\mathbf{0}}(\mathbf{x}_i), y_i) \\ &\leq C_0. \end{aligned}$$

Therefore,

$$\begin{aligned} & ||f_S||_K^2 \leq \frac{C_0}{\lambda} \\ \Longrightarrow \quad & |f_S|_\infty \leq \kappa ||f_S||_K \leq \kappa \sqrt{\frac{C_0}{\lambda}} \end{aligned}$$

Since the loss is L-Lipschitz, a bound on $|f_S|_\infty$ implies boundedness of the loss function.

A Note on λ

We have shown that Tikhonov regularization is uniformly stable with

$$\beta = \frac{L^2\kappa^2}{\lambda \ell}.$$

If we keep λ fixed as we increase ℓ, the generalization bound will tighten as $O\left(\frac{1}{\sqrt{\ell}}\right)$. However, keeping λ fixed is equivalent to keeping our hypothesis space fixed. As we get more data, we want λ to get smaller. If λ gets smaller too fast, the bounds become trivial.

Tikhonov vs. Ivanov

It is worth noting that Ivanov regularization

$$\begin{aligned} \hat{f}_{H,S} &= \arg\min_{f \in \mathcal{H}} \frac{1}{\ell} \sum_{i=1}^{\ell} V(f(x_i), y_i) \\ \text{s.t.} \quad & \|f\|_K^2 \leq \tau \end{aligned}$$

is **not** uniformly stable with $\beta = O\left(\frac{1}{n}\right)$, essentially because the constraint bounding the RKHS norm may not be tight. This is an important distinction between Tikhonov and Ivanov regularization.