Iterative Convex Regularization

Lorenzo Rosasco

Universita' di Genova
Istituto Italiano di Tecnologia
Massachusetts Institute of Technology

UNIVERSITA' DI GENOVA

istituto italiano di tecnologia

Massachusetts Institute of Technology

LCSL Laboratory for Computational and Statistical Learning

Optimization and Statistical Learning Workshop, Les Houches, Montevideo, January 14

ongoing work with S. Villa IIT-MIT, B.C. Vu IIT-MIT

Early Stopping
~~Iterative Convex~~ Regularization

Lorenzo Rosasco

Universita' di Genova
Istituto Italiano di Tecnologia
Massachusetts Institute of Technology

UNIVERSITA' DI GENOVA

istituto italiano di tecnologia

Massachusetts Institute of Technology

LCSL Laboratory for Computational and Statistical Learning

Optimization and Statistical Learning Workshop, Les Houches, Montevideo, January 14

ongoing work with S. Villa IIT-MIT, B.C. Vu IIT-MIT

Plan

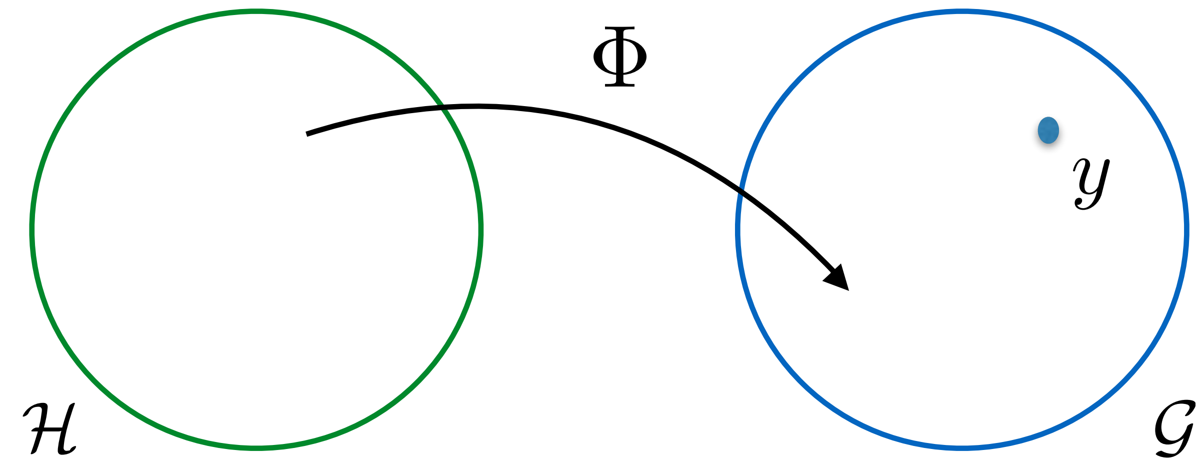
Optimization & Statistics/Estimation

- **part I: introduction to iterative regularization**
- part II: iterative convex regularization: problem and results

Linear Inverse Problems

$$\Phi w = y, \quad \Phi : \mathcal{H} \to \mathcal{G}$$

linear and bounded



Moore-Penrose Solution

$$w^\dagger = \arg\min_{\Phi w = y} R(w)$$

strongly convex lsc

Examples: *endless list here*

Data

$$\Phi w = y$$

Data Type I

$$\|y - \hat{y}\| \leq \delta$$

Data Type II

$$\left\|\Phi^* y - \hat{\Phi}^* \hat{y}\right\| \leq \delta$$

$$\left\|\Phi^* \Phi - \hat{\Phi}^* \hat{\Phi}\right\| \leq \eta \qquad \hat{\Phi} : \mathcal{H} \to \hat{\mathcal{G}}$$

- Data type I: Deterministic/stochastic noise […]
- Data type II: stochastic noise statistical Learning [R. et al. '05], also econometrics, discretized PDEs (?)

Learning* as an Inverse Problem

[De vito et al. ‘05]

$$Y_i = \langle w^\dagger, X_i \rangle + N_i, \quad i = 1, \ldots, n$$

Can be shown to fit Data Type II with

$$\Phi^* \Phi = \mathbb{E} X X^T, \qquad \hat{\Phi}^* \hat{\Phi} = \frac{1}{n} \sum_{i=1}^n X_i X_i^T$$

$$\delta, \eta \sim \frac{1}{\sqrt{n}}$$

$$\Phi^* y = \mathbb{E} X Y, \qquad \hat{\Phi}^* \hat{y} = \frac{1}{n} \sum_{i=1}^n X_i Y_i$$

Nonparametric extensions via RKHS theory:

Covariance operators become integral operators

*Random Design Regression

Tikhonov Regularization

$$\hat{w}_\lambda = \arg\min_{w \in \mathcal{H}} \left\| \hat{\Phi} w - \hat{y} \right\|^2 + \lambda R(w), \quad \lambda \geq 0$$

Computations

$\hat{w}_{t,\lambda}$

Variance

$$w_\lambda = \arg\min_{w \in \mathcal{H}} \left\| \Phi w - y \right\|^2 + \lambda R(w)$$

Bias

- *New Trade-Offs (?)*
- *Complexity of Model selection?*

$$w^\dagger = \arg\min_{\Phi w = y} R(w)$$

From Tikhonov Regularization

…to Landweber Regularization

$$R(w) = \|w\|^2$$

$$w^\dagger = \Phi^\dagger y$$

$$\sim (\Phi^*\Phi + \lambda I)^{-1}\Phi^* y$$

$$\sim \sum_{j=0}^{t} (I - \Phi^*\Phi)^j \Phi^* y$$

$$w_{t+1} = w_t + \Phi^*(\Phi w_t - y)$$

$$\hat{w}_{t+1} = \hat{w}_t + \hat{\Phi}^*(\hat{\Phi}\hat{w}_t - \hat{y})$$

Landweber Regularization aka Gradient Descent

$$R(w) = \|w\|^2$$

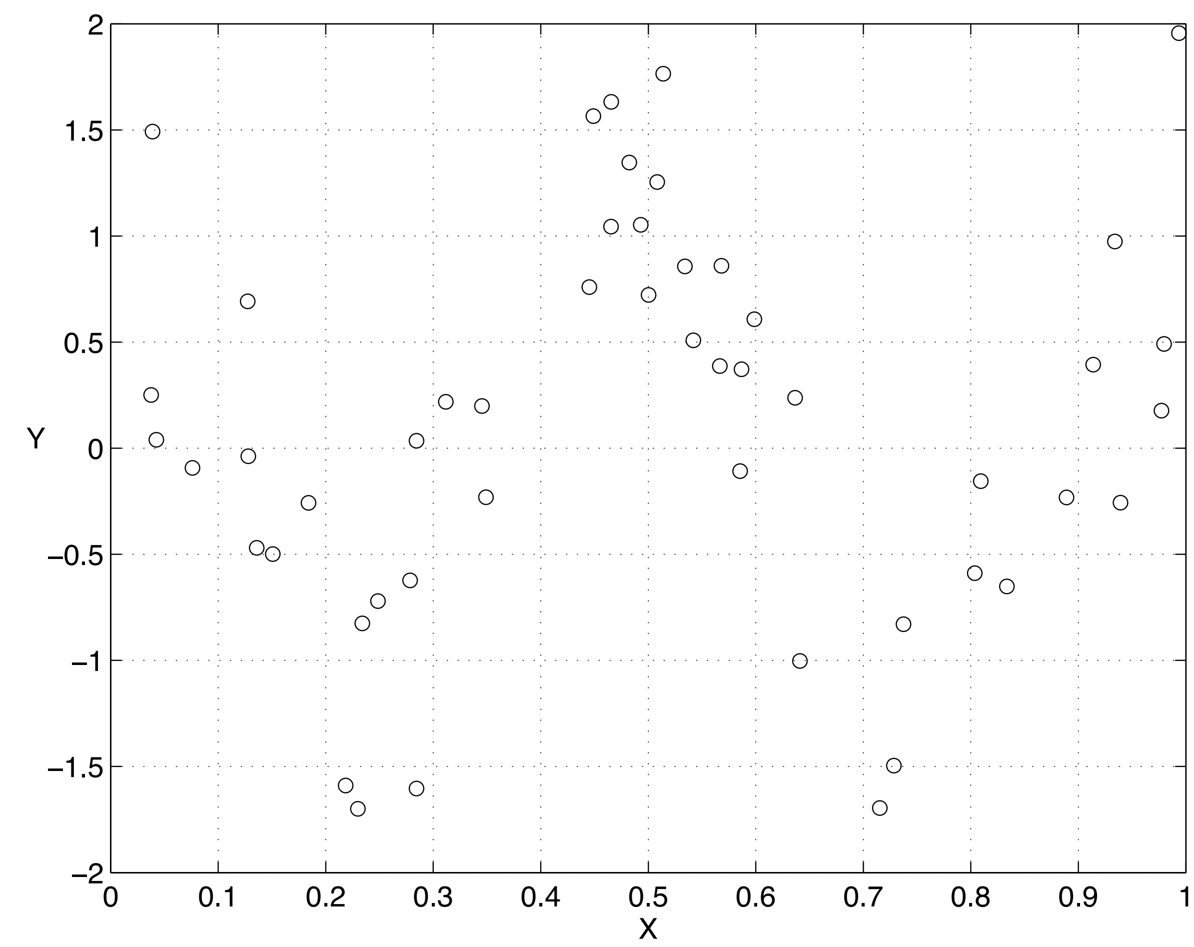
$$w^\dagger = \Phi^\dagger y$$

$$\sim (\Phi^*\Phi + \lambda I)^{-1}\Phi^* y$$

$$\sim \sum_{j=0}^{t} (I - \Phi^*\Phi)^j \Phi^* y$$

$$w_{t+1} = w_t + \Phi^*(\Phi w_t - y)$$

$$\hat{w}_{t+1} = \hat{w}_t + \hat{\Phi}^*(\hat{\Phi}\hat{w}_t - \hat{y})$$



Landweber Regularization aka Gradient Descent

$$R(w) = \|w\|^2$$

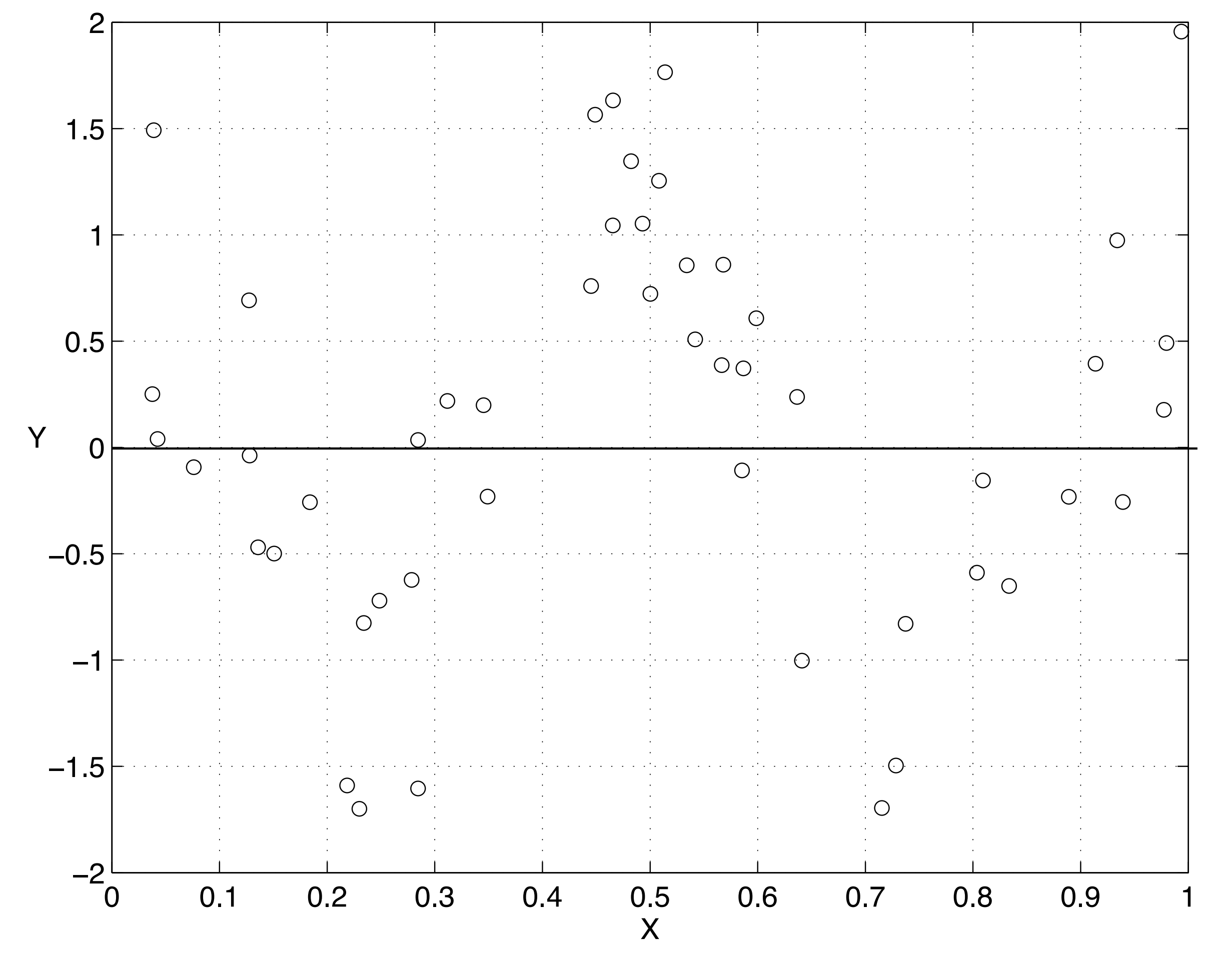
$$w^\dagger = \Phi^\dagger y$$

$$\sim (\Phi^*\Phi + \lambda I)^{-1}\Phi^* y$$

$$\sim \sum_{j=0}^{t} (I - \Phi^*\Phi)^j \Phi^* y$$

$$w_{t+1} = w_t + \Phi^*(\Phi w_t - y)$$

$$\hat{w}_{t+1} = \hat{w}_t + \hat{\Phi}^*(\hat{\Phi}\hat{w}_t - \hat{y})$$



Landweber Regularization aka Gradient Descent

$$R(w) = \|w\|^2$$

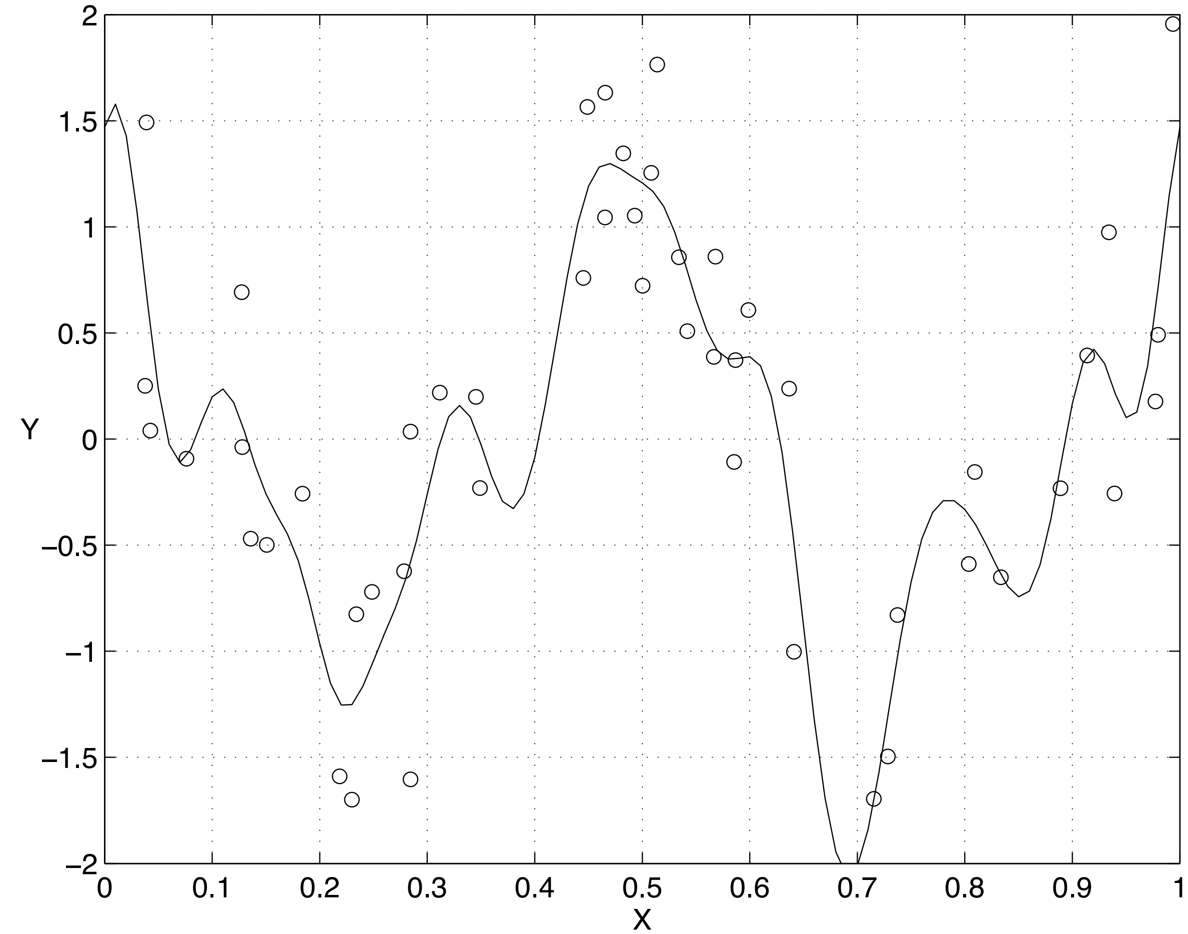
$$w^\dagger = \Phi^\dagger y$$

$$\sim (\Phi^*\Phi + \lambda I)^{-1}\Phi^* y$$

$$\sim \sum_{j=0}^{t} (I - \Phi^*\Phi)^j \Phi^* y$$

$$w_{t+1} = w_t + \Phi^*(\Phi w_t - y)$$

$$\hat{w}_{t+1} = \hat{w}_t + \hat{\Phi}^*(\hat{\Phi}\hat{w}_t - \hat{y})$$



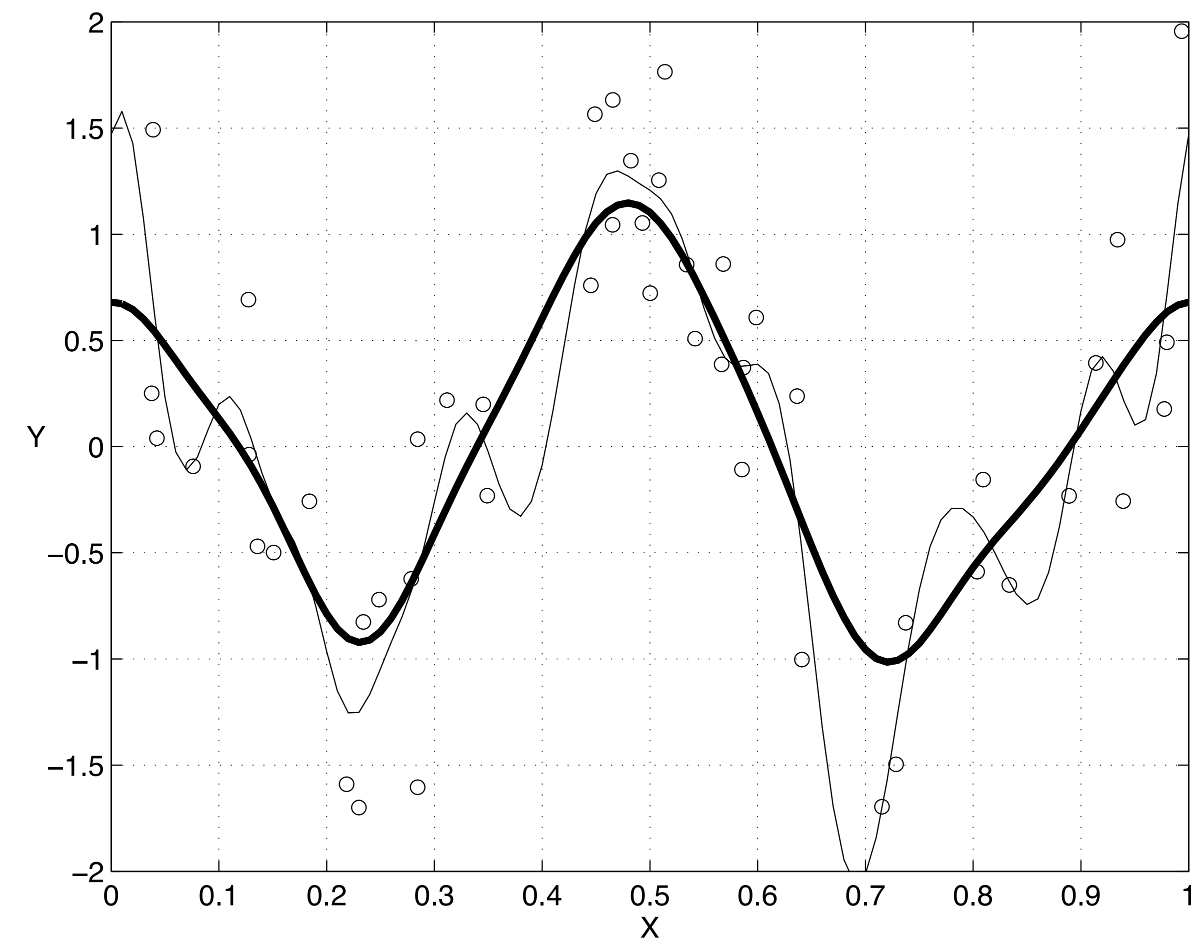
Landweber Regularization aka Gradient Descent

$$R(w) = \|w\|^2$$

$$w^\dagger = \Phi^\dagger y$$

$$\sim (\Phi^*\Phi + \lambda I)^{-1}\Phi^* y$$

$$\sim \sum_{j=0}^{t} (I - \Phi^*\Phi)^j \Phi^* y$$



$$w_{t+1} = w_t + \Phi^*(\Phi w_t - y)$$

$$\hat{w}_{t+1} = \hat{w}_t + \hat{\Phi}^*(\hat{\Phi}\hat{w}_t - \hat{y})$$

Landweber Regularization aka Gradient Descent

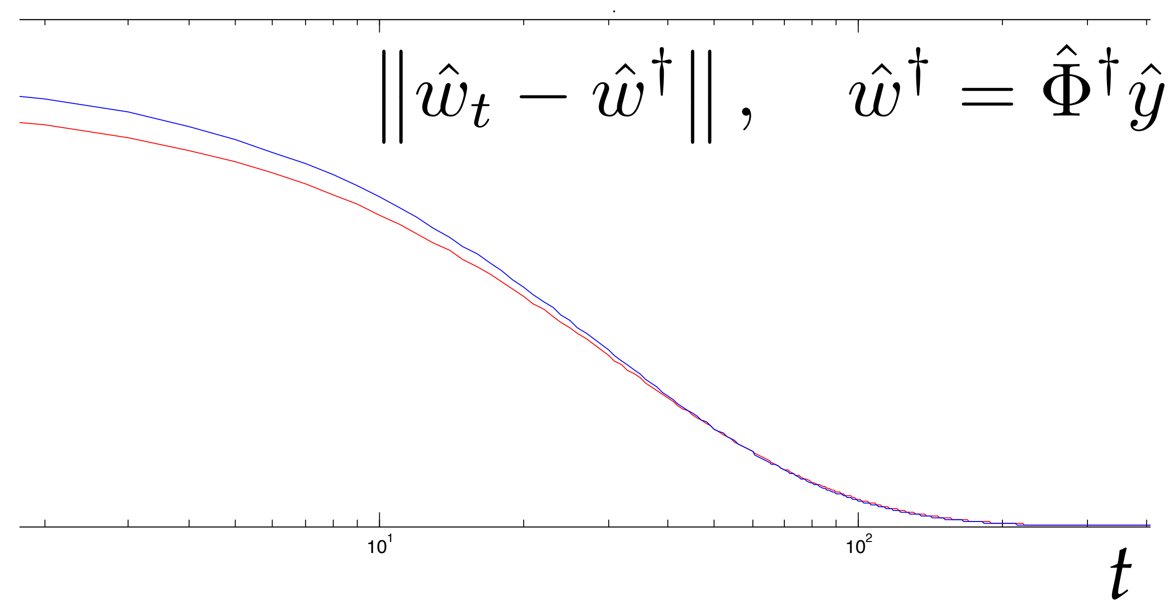
$$R(w) = \|w\|^2$$

$$w^\dagger = \Phi^\dagger y$$

$$\sim (\Phi^*\Phi + \lambda I)^{-1}\Phi^* y$$

$$\sim \sum_{j=0}^{t} (I - \Phi^*\Phi)^j \Phi^* y$$

$$w_{t+1} = w_t + \Phi^*(\Phi w_t - y)$$



$$\hat{w}_{t+1} = \hat{w}_t + \hat{\Phi}^*(\hat{\Phi}\hat{w}_t - \hat{y})$$

Landweber Regularization aka Gradient Descent

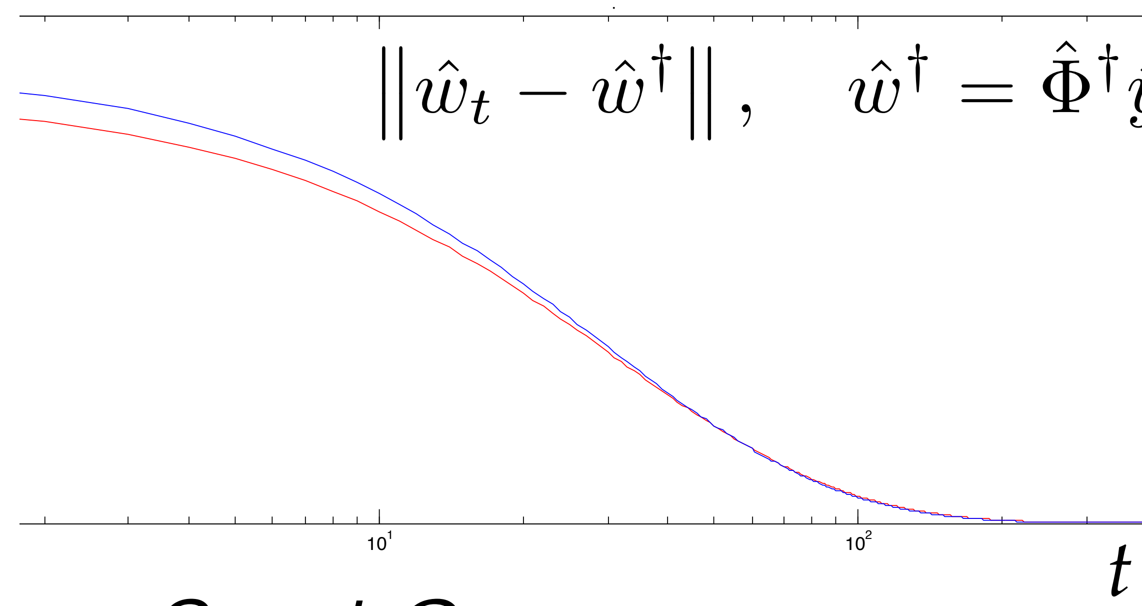
$$R(w) = \|w\|^2$$

$$w^\dagger = \Phi^\dagger y$$

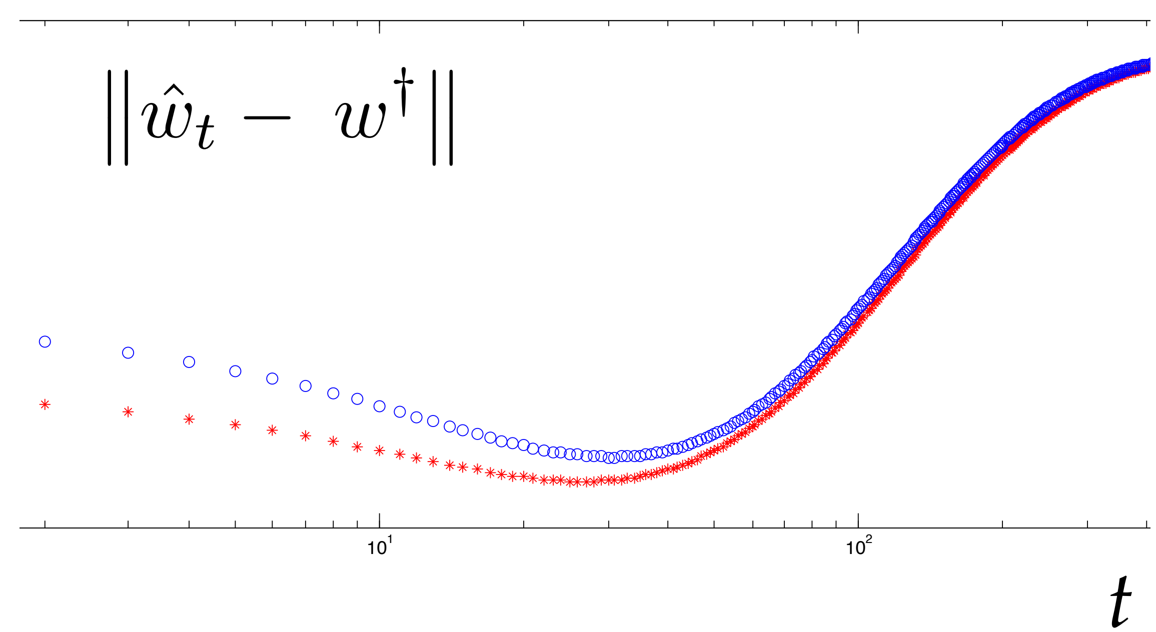
$$\sim (\Phi^*\Phi + \lambda I)^{-1}\Phi^* y$$

$$\sim \sum_{j=0}^{t} (I - \Phi^*\Phi)^j \Phi^* y$$

$$w_{t+1} = w_t + \Phi^*(\Phi w_t - y)$$



Semi-Convergence



$$\hat{w}_{t+1} = \hat{w}_t + \hat{\Phi}^*(\hat{\Phi}\hat{w}_t - \hat{y})$$

Remarks I

$$R(w) = \|w\|^2$$

Data type I:

- **History**: iteration+semiconvergence [Landweber ’50] …[…Nemirovski’86…]
- Other iterative approaches— some **acceleration: nu-method**/Chebyshev method [Brakhage ‘87, Nemirovski Polyak'84], **conjugate gradient** [Nemirovski’86…]…

- **Deterministic** noise [Engl et al. ’96], **stocastic** noise […,Buhlmann, Yu ’02 (L2 Boosting),Bissantz et al. ’07]
- Extensions to **noise in the operator** [Nemirovski’86,…]
- **Nonlinear** problems [Kaltenbacher et al. ’08]
- **Banach** Spaces [Schuster et al. ‘12]

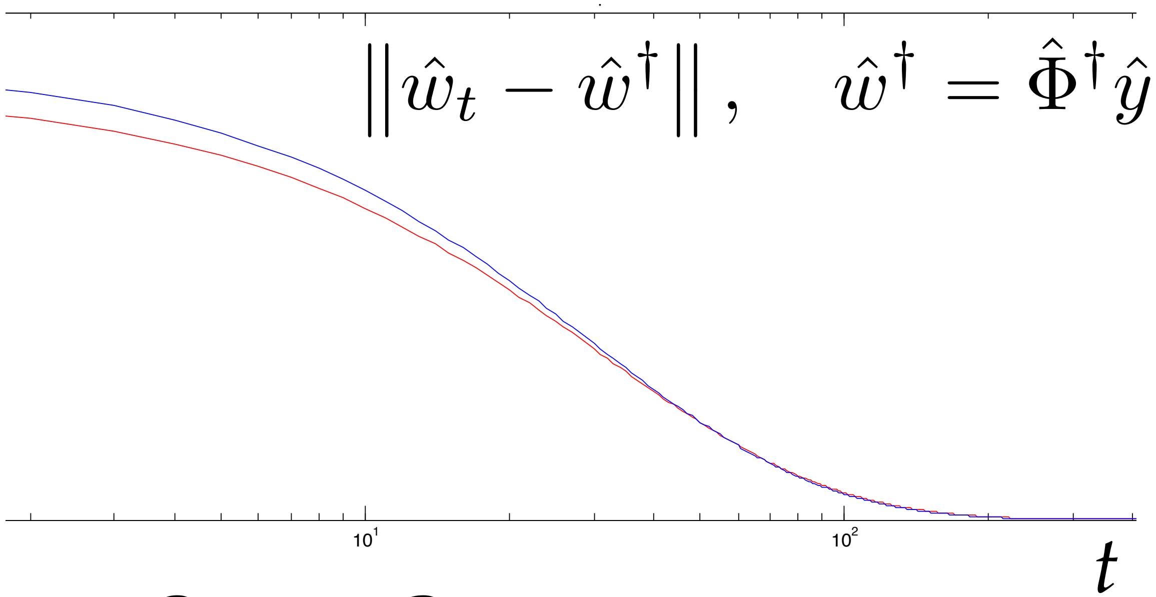
Remarks II

$$R(w) = \|w\|^2$$

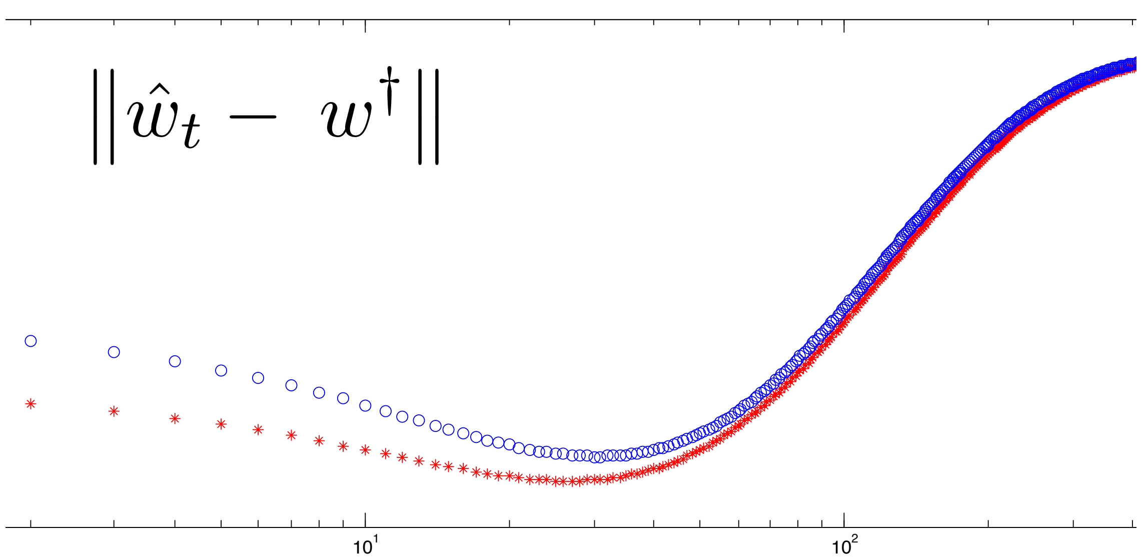
Data type II:

- **Deterministic noise** Landweber and nu-method [De Vito et al. ‘06]
- **Stochastic noise/learning Landweber** and **nu-method** [Ong Canu ’04, R et al ’04, Yao et al.’05, Bauer et al. ’06, Caponetto Yao ’07, Raskutti et al.’13]
- …also **conjugate gradient** [Blanchard Cramer ‘10]
- …and **incremental gradient** aka multiple passes SGD [R et al.’14]
- …and **(convex) loss,** subgradient method [Lin, R, Zhou ‘15]
- **Works really well** in practice [Huang et al. ’14, Perronnin et al. ‘13]
- **Regularization “path”** is for free

Remarks III



Semi-Convergence



$$\hat{w}_{t+1} = \hat{w}_t + \hat{\Phi}^*(\hat{\Phi}\hat{w}_t - \hat{y})$$

Take home message

Computations/iterations

control

stability/regularization

New trade-offs?

Can we derive iterative regularization for any (strongly) convex regularization?

Plan

- part I: introduction to iterative regularization
- **part II: iterative convex regularization: problem and results**

$$\hat{w}_{t+1} = \hat{w}_t + \hat{\Phi}^*(\hat{\Phi}\hat{w}_t - \hat{y})$$

How can I tell the iteration which regularization I want to use?

$$w^\dagger = \arg\min_{\Phi w = y} R(w)$$

Iterative Regularization and Early Stopping

$$w_t = A(w_0, \ldots, w_{t-1}, \Phi, y)$$

Convergence

Exact

$$\left\|w_t - w^\dagger\right\| \to 0, \quad t \to \infty$$

Noisy

$$\exists\, t^\dagger = t^\dagger(w^\dagger, \delta, \eta) \quad \textbf{s.t.} \quad \left\|\hat{w}_{t^\dagger} - w^\dagger\right\| \to 0, \quad (\delta, \eta) \to 0$$

Error Bounds

$$\exists\, t^\dagger = t^\dagger(w^\dagger, \delta, \eta) \quad \textbf{s.t.} \quad \left\|\hat{w}_{t^\dagger} - w^\dagger\right\| \leq \varepsilon(w^\dagger, \delta, \eta)$$

adaptivity, e.g. via discrepancy or Lepskii principles

Dual Forward Backward (DFB)

$$w^\dagger = \operatorname*{arg\,min}_{\Phi w = y} R(w) \qquad R = F + \frac{\alpha}{2}\left\|\cdot\right\|^2, \quad \alpha \geq 0$$

convex lsc

$$(\forall t \in \mathbb{N}) \quad \left\lfloor \begin{array}{l} w_t = \operatorname{prox}_{\alpha^{-1}F}\left(-\alpha^{-1}\Phi^* v_t\right) \\ v_{t+1} = v_t + \gamma_t(\Phi w_t - y). \end{array} \right. \qquad \gamma_t = \alpha$$

- Analogous iteration for **noisy data**
- Special case of **dual forward backward splitting** [Combettes et al. '10]…
- …also a form of **augmented Lagrangian method/ADMM** [see Beck Teboulle '14]
- …also can be shown to be equivalent to **linearized Bregmanized operator splitting** [Burger, Osher et al. …]
- Reduces to **Landweber** iteration if we consider only the squared norm

Analysis for **Data Type I**

[R.Villa Vu et al.'14]

$$\left\|\hat{w}_t - w^\dagger\right\| \leq \left\|\hat{w}_t - w_t\right\| + \left\|w_t - w^\dagger\right\|$$

$$\left\|v^\dagger\right\| / (\alpha\sqrt{t})$$

Theorem. If there exists $v^\dagger \in \mathcal{G}$ such that

$$\Phi^* v^\dagger \in \partial R(w^\dagger)$$

the DFB sequence $(w_t)_t$ for $v_0 = 0$ satisfies

$$\|w_t - w^\dagger\| \leq \frac{\|v^\dagger\|}{\alpha\sqrt{t}}$$

Proof idea $\quad \frac{\alpha}{2}\|w_t - w^\dagger\|^2 \leq D(v_t) - D(v^\dagger)$

Analysis for **Data Type I**

[R.Villa Vu et al.'14]

$$\left\|\hat{w}_t - w^\dagger\right\| \leq \|\hat{w}_t - w_t\| + \left\|w_t - w^\dagger\right\|$$

$$c\delta t \qquad\qquad \left\|v^\dagger\right\| / (\alpha\sqrt{t})$$

Theorem. Let $(w_t)_t$,$(\hat{w}_t)_t$ be the DFB sequences for $\hat{v}_0 = v_0 = 0$. Then it holds

$$\|\hat{w}_t - w_t\| \leq \frac{2t\delta}{\|\Phi\|}$$

Analysis for **Data Type I**

[R.Villa Vu et al.'14]

$$\left\|\hat{w}_t - w^\dagger\right\| \leq \|\hat{w}_t - w_t\| + \left\|w_t - w^\dagger\right\|$$

$$c\delta t \qquad \left\|v^\dagger\right\| / (\alpha\sqrt{t})$$

$$t^\dagger = c\delta^{-2/3} \Rightarrow \left\|\hat{w}_{t^\dagger} - w^\dagger\right\| \leq c\delta^{1/3}$$

Analysis for **Data Type II**

[R. Villa Vu et al.'14]

$$\left\|\hat{w}_t - w^\dagger\right\| \leq \|\hat{w}_t - w_t\| + \left\|w_t - w^\dagger\right\|$$

$$(\delta + \eta)(1 + c)^t \qquad \left\|v^\dagger\right\| / (\alpha\sqrt{t})$$

$$\hat{t} = c \log \sqrt{1/(\delta+\eta)} \;\Rightarrow\; \|\hat{w}_{\hat{t}} - w^\dagger\| \leq \frac{c}{\sqrt{\log(1/\sqrt{\delta+\eta})}}$$

Remarks

Data Type I

- General **convex** setting— only **weak convergence** [Burger, Osher et al. ~'09'10], no stability results, no strong convergence.
- **Sparsity** based regularization [Osher et al. '14]

Data Type II

- **No previous results**, either convergence or error bounds.
- Directly give **results for statistical learning**.
- **Acceleration** possible, but stability harder to prove (e.g. via dual FISTA, Chambolle Pock…)
- **Polynomial estimates** of variance under stronger conditions (satisfied in certa smooth cases, e.g. Landweber)
- Connections to **regularization path**, e.g. Lasso path/Lars Results…

Work in Progress

- Purely convex case: exact penalization result for atomic norms?
- Analysis under partial smoothness
- Sharper Bounds (high-finite dimension)
- Truly ill-posed problems
- (more) Experiments

Conclusions

- Iterative Regularization **viable alternative** to Tikhonov regularization for large problems
- *Old (?) Trade offs* in ML: **computational regularization??**
- A whole **new** playground - loss, iterations, randomization