Generalization properties of multiple passes stochastic gradient method

Silvia Villa
joint work with Lorenzo Rosasco

Laboratory for Computational and Statistical Learning, IIT and MIT
http://lcsl.mit.edu/data/silviavilla

– Computational and Statistical trade-offs in learning –
Paris 2016

Motivation

Large scale learning
=

Statistics + **Optimization**

Stochastic gradient method

single or multiple passes over the data?

Outline

- **Problem setting**
- **Tikhonov regularization for learning**
  - Assumptions: source condition
  - Theoretical results
- **Learning with the stochastic gradient method**
  - Algorithms
  - Theoretical results and discussion
  - Proof

Problem setting

- $\mathcal{H}$ separable Hilbert space
- ρ probability distribution on $\mathcal{H} \times \mathbb{R}$

Problem

$$\underset{w \in \mathcal{H}}{\text{minimize}} \; \mathcal{E}(w) = \int_{\mathcal{H} \times \mathbb{R}} (\langle w, x \rangle - y)^2 \, d\rho(x, y),$$

given $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ i.i.d. with respect to ρ.

Special cases of interest

Linear and functional regression

Let

$$y_i = \langle w_*, x_i \rangle + \delta_i, \qquad i = 1, \dots, n$$

with x_i, δ_i random iid and $w_*, x_i \in \mathcal{H}$.

- random design linear regression, $\mathcal{H} = \mathbb{R}^d$
- functional regression, $\mathcal{H}$ infinite dimensional Hilbert space

Special cases of interest

Learning with kernels

- $\Xi \times \mathbb{R}$ input/output space with probability μ.
- $\mathcal{H}_K$ RKHS with reproducing kernel K, $w(\xi) = \langle w, K(\xi, \cdot)\rangle_{\mathcal{H}}$

Problem

$$\text{minimize}_{w \in \mathcal{H}_K} \int_{\Xi \times \mathbb{R}} (w(\xi) - y)^2 d\mu(\xi, y)$$

Special cases of interest

Learning with kernels

- $\Xi \times \mathbb{R}$ input/output space with probability μ.
- $\mathcal{H}_K$ RKHS with reproducing kernel K, $w(\xi) = \langle w, K(\xi, \cdot)\rangle_{\mathcal{H}}$

Problem

$$\text{minimize}_{w \in \mathcal{H}_K} \int_{\Xi \times \mathbb{R}} (w(\xi) - y)^2 d\mu(\xi, y)$$

If ρ is the distribution of $(\xi, y) \mapsto (K(\xi, \cdot), y) = (x, y)$, then

Special cases of interest

Learning with kernels

- $\Xi \times \mathbb{R}$ input/output space with probability μ.
- $\mathcal{H}_K$ RKHS with reproducing kernel K, $w(\xi) = \langle w, K(\xi, \cdot)\rangle_{\mathcal{H}}$

Problem

$$\text{minimize}_{w \in \mathcal{H}_K} \int_{\Xi \times \mathbb{R}} (w(\xi) - y)^2 d\mu(\xi, y)$$

If ρ is the distribution of $(\xi, y) \mapsto (K(\xi, \cdot), y) = (x, y)$, then

$$\int_{\Xi \times \mathbb{R}} (w(\xi) - y)^2 d\mu(\xi, y) = \int_{\Xi \times \mathbb{R}} (\langle w, K(\xi, \cdot)\rangle - y)^2 d\mu = \int_{\mathcal{H}_K \times \mathbb{R}} (\langle w, x\rangle - y)^2 d\rho(w, y)$$

Outline

- Problem setting
- **Tikhonov regularization for learning**
  - Assumptions: source condition
  - Theoretical results
- Learning with the stochastic gradient method
  - Algorithms
  - Theoretical results and discussion
  - Proof

A classical approach: Tikhonov regularization of the empirical risk

$$\hat{w}_\lambda = \operatorname*{argmin}_{\mathcal{H}} \hat{\mathcal{E}}(w) + \lambda R(w)$$

- empirical risk,

$$\hat{\mathcal{E}}(w) = \frac{1}{n} \sum_{i=1}^{n} (\langle w, x_i \rangle - y_i)^2$$

A classical approach: Tikhonov regularization of the empirical risk

$$\hat{w}_\lambda = \operatorname*{argmin}_{\mathcal{H}} \hat{\mathcal{E}}(w) + \lambda R(w)$$

- empirical risk,

$$\hat{\mathcal{E}}(w) = \frac{1}{n}\sum_{i=1}^{n}(\langle w, x_i\rangle - y_i)^2$$

- regularizer, $R = \|\cdot\|_{\mathcal{H}}^2$

A classical approach: Tikhonov regularization of the empirical risk

$$\hat{w}_\lambda = \underset{\mathcal{H}}{\operatorname{argmin}}\, \hat{\mathcal{E}}(w) + \lambda R(w)$$

- empirical risk,

$$\hat{\mathcal{E}}(w) = \frac{1}{n}\sum_{i=1}^{n}(\langle w, x_i\rangle - y_i)^2$$

- regularizer, $R = \|\cdot\|_{\mathcal{H}}^2$
- regularization parameter, $\lambda > 0$

A classical approach: Tikhonov regularization of the empirical risk

$$\hat{w}_\lambda = \operatorname*{argmin}_{\mathcal{H}} \hat{\mathcal{E}}(w) + \lambda R(w)$$

- empirical risk,

$$\hat{\mathcal{E}}(w) = \frac{1}{n}\sum_{i=1}^{n}(\langle w, x_i\rangle - y_i)^2$$

- regularizer, $R = \|\cdot\|_{\mathcal{H}}^2$
- regularization parameter, $\lambda > 0$

What about statistics?

Assumptions

Boundedness.

There exist $\kappa > 0$ and $M > 0$ such that

$$|y| \leq M \quad \text{and} \quad \|x\|_{\mathcal{H}}^2 \leq \kappa \quad \text{a.s.}$$

Assumptions

Boundedness.

There exist $\kappa > 0$ and $M > 0$ such that

$$|y| \leq M \quad \text{and} \quad \|x\|_{\mathcal{H}}^2 \leq \kappa \quad \text{a.s.}$$

Existence a (minimal norm) solution

$$\mathcal{O} = \underset{\mathcal{H}}{\operatorname{argmin}}\, \mathcal{E} \neq \varnothing, \qquad w^\dagger = \underset{\mathcal{O}}{\operatorname{argmin}}\, \|w\|_{\mathcal{H}}$$

More assumptions needed for finite sample error bounds...

Source condition

Boundedness assumption implies

$$\begin{aligned} T\colon \mathcal{H} &\to \mathcal{H} \\ w &\mapsto \int_{\mathcal{H}} \langle w, x\rangle\, x\, d\rho_{\mathcal{H}}(x), \qquad \rho_{\mathcal{H}} \text{ marginal of } \rho \end{aligned}$$

is well defined.

Source condition

Boundedness assumption implies

$$\begin{aligned} T\colon \mathcal{H} &\to \mathcal{H} \\ w &\mapsto \int_{\mathcal{H}} \langle w, x\rangle\, x\, d\rho_{\mathcal{H}}(x), \qquad \rho_{\mathcal{H}} \text{ marginal of } \rho \end{aligned}$$

is well defined.

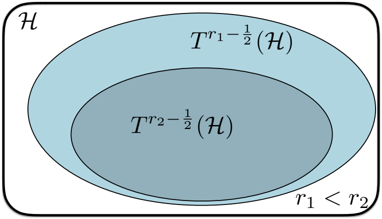
Source condition

Let $r \in [1/2, +\infty[$ and assume that

$$\exists h \in \mathcal{H} \quad \text{such that} \quad \mathbf{w}^{\dagger} = \mathbf{T}^{\mathbf{r}-1/2}\mathbf{h}. \tag{SC}$$

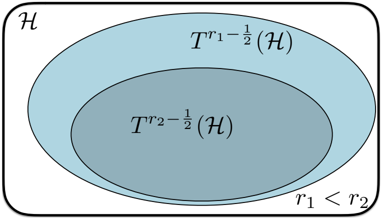
Source condition: remarks

- If $r = 1/2$, no assumption
- if $r > 1/2$, (SC) implies $w^\dagger$ is in a subspace of $\mathcal{H}$



Source condition: remarks

- If $r = 1/2$, no assumption
- if $r > 1/2$, (SC) implies $w^\dagger$ is in a subspace of $\mathcal{H}$



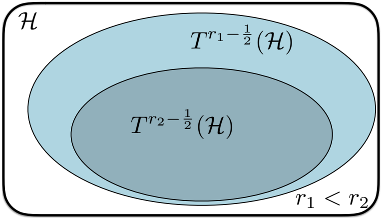
Spectral point of view

- If $(\sigma_i, v_i)_{i\in I}$ is the eigenbasis of T, then

$$\|w^\dagger\|^2_{\mathcal{H}} = \sum |\langle w^\dagger, v_i\rangle|^2 < +\infty$$

Source condition: remarks

- If $r = 1/2$, no assumption
- if $r > 1/2$, (SC) implies $w^\dagger$ is in a subspace of $\mathcal{H}$



Spectral point of view

- If $(\sigma_i, v_i)_{i\in I}$ is the eigenbasis of T, then

$$\|w^\dagger\|^2_{\mathcal{H}} = \sum |\langle w^\dagger, v_i\rangle|^2 < +\infty$$

- If $h = T^{1/2-r} w^\dagger$, then

$$\|\mathbf{h}\|^2_{\mathcal{H}} = \sum |\langle \mathbf{w}^\dagger, \mathbf{v_i}\rangle|^2 / \sigma_{\mathbf{i}}^{\mathbf{2r-1}} < +\infty$$

Results: error bounds

Theorem

Assume boundedness and (SC) for some $r \in]1/2, +\infty[$, Let

$$\hat{\lambda} = \mathbf{n}^{-\frac{1}{2r+1}}.$$

then with high probability,

$$\|\hat{\mathbf{w}}_{\hat{\lambda}} - \mathbf{w}^{\dagger}\|_{\mathcal{H}} = \begin{cases} O\left(n^{-\frac{r-1/2}{2r+1}}\right) & \text{if } r \leq 3/2 \\ O\left(n^{-1/2}\right) & \text{if } r > 3/2 \end{cases}$$

Proof: bias-variance trade-off

Set

$$w_\lambda = \operatorname*{argmin}_{\mathcal{H}} \mathcal{E}(w) + \lambda \|w\|_{\mathcal{H}}^2,$$

and decompose the error

$$\|\hat{w}_\lambda - w^\dagger\|_{\mathcal{H}} \leq \underbrace{\|\hat{w}_\lambda - w_\lambda\|_{\mathcal{H}}}_{\text{Variance}} + \underbrace{\|w_\lambda - w^\dagger\|_{\mathcal{H}}}_{\text{Bias}}$$

Remarks

- The bounds are minimax [Blanchard-Muecke 2016]: if

$$\mathcal{P}_r = \{\rho \,|\, \text{Boundedness and (SC) are satisfied}\},$$

then

$$\min_{\hat{w}\in\mathcal{H}} \max_{\rho\in\mathcal{P}_r} \mathsf{E}\|\hat{w} - w^\dagger\|_{\mathcal{H}} \geq Cn^{-\frac{r-1/2}{2r+1}}$$

Remarks

- The bounds are minimax [Blanchard-Muecke 2016]: if

$$\mathcal{P}_r = \{\rho \,|\, \text{Boundedness and (SC) are satisfied}\},$$

  then

$$\min_{\hat{w}\in\mathcal{H}} \max_{\rho\in\mathcal{P}_r} \mathsf{E}\|\hat{w} - w^\dagger\|_{\mathcal{H}} \geq Cn^{-\frac{r-1/2}{2r+1}}$$

- saturation for $r > 3/2$

Remarks

- The bounds are minimax [Blanchard-Muecke 2016]: if

$$\mathcal{P}_r = \{\rho \,|\, \text{Boundedness and (SC) are satisfied}\},$$

then

$$\min_{\hat{w}\in\mathcal{H}} \max_{\rho\in\mathcal{P}_r} \mathsf{E}\|\hat{w} - w^\dagger\|_{\mathcal{H}} \geq Cn^{-\frac{r-1/2}{2r+1}}$$

- saturation for $r > 3/2$
- Adaptivity via Lepskii/Balancing principle [De Vito-Pereverzev-Rosasco 2010]

Remarks

- The bounds are minimax [Blanchard-Muecke 2016]: if

$$\mathcal{P}_r = \{\rho \,|\, \text{Boundedness and (SC) are satisfied}\},$$

  then

$$\min_{\hat{w} \in \mathcal{H}} \max_{\rho \in \mathcal{P}_r} \mathsf{E}\|\hat{w} - w^\dagger\|_{\mathcal{H}} \geq C n^{-\frac{r-1/2}{2r+1}}$$

- saturation for $r > 3/2$
- Adaptivity via Lepskii/Balancing principle [De Vito-Pereverzev-Rosasco 2010]
- Improved bounds under further assumptions on the decay of the eigenvalues of T [De Vito-Caponnetto 2006 ...]

Remarks

- The bounds are minimax [Blanchard-Muecke 2016]: if

$$\mathcal{P}_r = \{\rho \,|\, \text{Boundedness and (SC) are satisfied}\},$$

then

$$\min_{\hat{w} \in \mathcal{H}} \max_{\rho \in \mathcal{P}_r} \mathrm{E} \|\hat{w} - w^\dagger\|_{\mathcal{H}} \geq C n^{-\frac{r-1/2}{2r+1}}$$

- saturation for $r > 3/2$
- Adaptivity via Lepskii/Balancing principle [De Vito-Pereverzev-Rosasco 2010]
- Improved bounds under further assumptions on the decay of the eigenvalues of T [De Vito-Caponnetto 2006 ...]

BUT...

Remarks

- The bounds are minimax [Blanchard-Muecke 2016]: if

$$\mathcal{P}_r = \{\rho \,|\, \text{Boundedness and (SC) are satisfied}\},$$

then

$$\min_{\hat{w}\in\mathcal{H}} \max_{\rho\in\mathcal{P}_r} \mathrm{E}\|\hat{w} - w^\dagger\|_{\mathcal{H}} \geq Cn^{-\frac{r-1/2}{2r+1}}$$

- saturation for $r > 3/2$
- Adaptivity via Lepskii/Balancing principle [De Vito-Pereverzev-Rosasco 2010]
- Improved bounds under further assumptions on the decay of the eigenvalues of T [De Vito-Caponnetto 2006 ...]

BUT...

...what about the **optimization error**?

A new perspective [Bottou-Bousquet 2008]

Let $\hat{w}_{\lambda,t}$ the t-th iteration of some algorithm for regularized empirical risk minimization,

A new perspective [Bottou-Bousquet 2008]

Let $\hat{w}_{\lambda,t}$ the t-th iteration of some algorithm for regularized empirical risk minimization, then

$$\|\hat{w}_{\lambda,t} - w^\dagger\|_{\mathcal{H}} \leq \underbrace{\|\hat{w}_{\lambda,t} - \hat{w}_\lambda\|_{\mathcal{H}}}_{\text{Optimization}} + \underbrace{\underbrace{\|\hat{w}_\lambda - w_\lambda\|_{\mathcal{H}}}_{\text{Variance}} + \underbrace{\|w_\lambda - w^\dagger\|_{\mathcal{H}}}_{\text{Bias}}}_{\text{Statistics}}$$

A new perspective [Bottou-Bousquet 2008]

Let $\hat{w}_{\lambda,t}$ the t-th iteration of some algorithm for regularized empirical risk minimization, then

$$\|\hat{w}_{\lambda,t} - w^\dagger\|_{\mathcal{H}} \leq \underbrace{\|\hat{w}_{\lambda,t} - \hat{w}_\lambda\|_{\mathcal{H}}}_{\text{Optimization}} + \underbrace{\underbrace{\|\hat{w}_\lambda - w_\lambda\|_{\mathcal{H}}}_{\text{Variance}} + \underbrace{\|w_\lambda - w^\dagger\|_{\mathcal{H}}}_{\text{Bias}}}_{\text{Statistics}}$$

A new trade-off

$\Rightarrow$ Optimization accuracy tailored to statistical accuracy

Optimization error

Recently a *lot* of interest in methods to solve

$$\min_{w \in \mathcal{H}} \sum_{i=1}^{n} V_i(w)$$

Optimization error

Recently a *lot* of interest in methods to solve

$$\min_{w \in \mathcal{H}} \sum_{i=1}^{n} V_i(w)$$

- Let $V_i(w) = V(\langle w, x_i \rangle_{\mathcal{H}}, y_i) + \lambda \|w\|_{\mathcal{H}}^2$ for some loss function V

Optimization error

Recently a *lot* of interest in methods to solve

$$\min_{w\in\mathcal{H}} \sum_{i=1}^{n} V_i(w)$$

- Let $V_i(w) = V(\langle w, x_i\rangle_{\mathcal{H}}, y_i) + \lambda\|w\|^2_{\mathcal{H}}$ for some loss function V
- Large scale setting, n *"large"*

Optimization error

Recently a *lot* of interest in methods to solve

$$\min_{w \in \mathcal{H}} \sum_{i=1}^{n} V_i(w)$$

- Let $V_i(w) = V(\langle w, x_i \rangle_{\mathcal{H}}, y_i) + \lambda \|w\|_{\mathcal{H}}^2$ for some loss function V
- Large scale setting, n *"large"*
- Focus on *batch and stochastic* **first order methods**

A new perspective [Bottou-Bousquet 2008]

$$\|\hat{w}_{\lambda,t} - w^\dagger\|_{\mathcal{H}} \leq \underbrace{\|\hat{w}_{\lambda,t} - \hat{w}_\lambda\|_{\mathcal{H}}}_{\text{Optimization}} + \underbrace{\underbrace{\|\hat{w}_\lambda - w_\lambda\|_{\mathcal{H}}}_{\text{Variance}} + \underbrace{\|w_\lambda - w^\dagger\|_{\mathcal{H}}}_{\text{Bias}}}_{\text{Statistics}}$$

This suggests

- Approach 1: combine statistics with optimization
- Approach 2: use a different decomposition

Outline

- Problem setting
- Tikhonov regularization for learning
  - Assumptions: source condition
  - Theoretical results
- **Learning with the stochastic gradient method**
  - Algorithms
  - Theoretical results and discussion
  - Proof

Another approach: stochastic gradient method

Directly minimize the expected risk $\mathcal{E}$

$$\hat{\mathbf{w}}_{\mathbf{t}} = \hat{\mathbf{w}}_{\mathbf{t-1}} - \gamma_{\mathbf{t}}(\langle \hat{\mathbf{w}}_{\mathbf{t-1}}, \mathbf{x}_{\mathbf{t}} \rangle - \mathbf{y}_{\mathbf{t}})\mathbf{x}_{\mathbf{t}}$$

Another approach: stochastic gradient method

Directly minimize the expected risk $\mathcal{E}$

$$\hat{\mathbf{w}}_{\mathbf{t}} = \hat{\mathbf{w}}_{\mathbf{t-1}} - \gamma_{\mathbf{t}}(\langle \hat{\mathbf{w}}_{\mathbf{t-1}}, \mathbf{x}_{\mathbf{t}} \rangle - \mathbf{y}_{\mathbf{t}})\mathbf{x}_{\mathbf{t}}$$

- each iteration corresponds to one sample/gradient evaluation

Another approach: stochastic gradient method

Directly minimize the expected risk $\mathcal{E}$

$$\hat{\mathbf{w}}_{\mathbf{t}} = \hat{\mathbf{w}}_{\mathbf{t-1}} - \gamma_{\mathbf{t}}(\langle \hat{\mathbf{w}}_{\mathbf{t-1}}, \mathbf{x}_{\mathbf{t}} \rangle - \mathbf{y}_{\mathbf{t}})\mathbf{x}_{\mathbf{t}}$$

- each iteration corresponds to one sample/gradient evaluation
- no explicit regularization

Another approach: stochastic gradient method

Directly minimize the expected risk $\mathcal{E}$

$$\hat{\mathbf{w}}_{\mathbf{t}} = \hat{\mathbf{w}}_{\mathbf{t-1}} - \gamma_{\mathbf{t}}(\langle \hat{\mathbf{w}}_{\mathbf{t-1}}, \mathbf{x}_{\mathbf{t}} \rangle - \mathbf{y}_{\mathbf{t}})\mathbf{x}_{\mathbf{t}}$$

- each iteration corresponds to one sample/gradient evaluation
- no explicit regularization
- several theoretical results

Another approach: stochastic gradient method

Directly minimize the expected risk $\mathcal{E}$

$$\hat{\mathbf{w}}_{\mathbf{t}} = \hat{\mathbf{w}}_{\mathbf{t-1}} - \gamma_{\mathbf{t}}(\langle \hat{\mathbf{w}}_{\mathbf{t-1}}, \mathbf{x}_{\mathbf{t}} \rangle - \mathbf{y}_{\mathbf{t}})\mathbf{x}_{\mathbf{t}}$$

- each iteration corresponds to one sample/gradient evaluation
- no explicit regularization
- several theoretical results

BUT...

Another approach: stochastic gradient method

Directly minimize the expected risk $\mathcal{E}$

$$\hat{\mathbf{w}}_{\mathbf{t}} = \hat{\mathbf{w}}_{\mathbf{t-1}} - \gamma_{\mathbf{t}}(\langle \hat{\mathbf{w}}_{\mathbf{t-1}}, \mathbf{x}_{\mathbf{i_t}} \rangle - \mathbf{y}_{\mathbf{i_t}})\mathbf{x}_{\mathbf{i_t}}$$

- each iteration corresponds to one sample/gradient evaluation
- no explicit regularization
- several theoretical results

BUT...

. . . in practice, **MULTIPLE PASSES** over the data are used

Multiple passes stochastic gradient learning

Rewrite the iteration

Let $\hat{w}_0 = 0$ and $\gamma > 0$. For $t \in \mathbb{N}$, iterate:

$$
\begin{array}{l}
\hat{v}_0 = \hat{w}_t \\
\text{for } i = 1, \dots, n \\
\quad \hat{v}_i = \hat{v}_{i-1} - (\gamma/n)(\langle \hat{v}_{i-1}, x_i \rangle - y_i)x_i \\
\hat{w}_{t+1} = \hat{v}_n
\end{array}
$$

Multiple passes stochastic gradient learning

Rewrite the iteration

Let $\hat{w}_0 = 0$ and $\gamma > 0$. For $t \in \mathbb{N}$, iterate:

$$
\left\lfloor
\begin{array}{l}
\hat{v}_0 = \hat{w}_t \\
\text{for } i = 1, \dots, n \\
\left\lfloor \; \hat{v}_i = \hat{v}_{i-1} - (\gamma/n)(\langle \hat{v}_{i-1}, x_i \rangle - y_i) x_i \right. \\
\hat{w}_{t+1} = \hat{v}_n
\end{array}
\right.
$$

- **incremental** gradient method for the **empirical** risk $\hat{\mathcal{E}}$ [Bertsekas -Tsitsiklis 2000]

Multiple passes stochastic gradient learning

Rewrite the iteration

Let $\hat{w}_0 = 0$ and $\gamma > 0$. For $t \in \mathbb{N}$, iterate:

$$
\left\lfloor
\begin{array}{l}
\hat{v}_0 = \hat{w}_t \\
\text{for } i = 1, \ldots, n \\
\left\lfloor \; \hat{v}_i = \hat{v}_{i-1} - (\gamma/n)(\langle \hat{v}_{i-1}, x_i \rangle - y_i) x_i \right. \\
\hat{w}_{t+1} = \hat{v}_n
\end{array}
\right.
$$

- **incremental** gradient method for the **empirical** risk $\hat{\mathcal{E}}$ [Bertsekas -Tsitsiklis 2000]
- each *inner* step is one *pass* of stochastic gradient method

Multiple passes stochastic gradient learning

Rewrite the iteration

Let $\hat{w}_0 = 0$ and $\gamma > 0$. For $t \in \mathbb{N}$, iterate:

$$
\begin{array}{l}
\hat{v}_0 = \hat{w}_t \\
\text{for } i = 1, \ldots, n \\
\quad \hat{v}_i = \hat{v}_{i-1} - (\gamma/n)(\langle \hat{v}_{i-1}, x_i \rangle - y_i) x_i \\
\hat{w}_{t+1} = \hat{v}_n
\end{array}
$$

- **incremental** gradient method for the **empirical** risk $\hat{\mathcal{E}}$ [Bertsekas -Tsitsiklis 2000]
- each *inner* step is one *pass* of stochastic gradient method
- t is the number of *passes* over the data (**epochs**)

Main question

How many passes we need to approximately minimize the expected risk $\mathcal{E}$?

Multiple passes stochastic gradient learning

We have

$$\hat{w}_t \to \operatorname*{argmin}_{\mathcal{H}} \hat{\mathcal{E}},$$

Multiple passes stochastic gradient learning

We have

$$\hat{w}_t \to \operatorname*{argmin}_{\mathcal{H}} \hat{\mathcal{E}},$$

but we would like

$$\hat{w}_t \to \operatorname*{argmin}_{\mathcal{H}} \mathcal{E}.$$

Multiple passes stochastic gradient learning

We have

$$\hat{w}_t \to \operatorname*{argmin}_{\mathcal{H}} \hat{\mathcal{E}},$$

but we would like

$$\hat{w}_t \to \operatorname*{argmin}_{\mathcal{H}} \mathcal{E}.$$

What's the catch?

Stability and early stopping

Consider the gradient descent iteration for the expected risk.

$$
\begin{array}{l}
w_0 = 0 \in \mathcal{H} \\
\left\lfloor \begin{array}{l}
v_0 = w_t \\
\text{for } i = 1, \ldots, n \\
\left\lfloor \; v_i = v_{i-1} - (\gamma/n) \int_{\mathcal{H}} (\langle v_{i-1}, x \rangle - y) x \, d\rho(x, y) \right. \\
w_{t+1} = v_n
\end{array} \right.
\end{array}
$$

Note: step-size γ/n.

Stability and early stopping

Consider the gradient descent iteration for the expected risk.

$$
\begin{aligned}
& w_0 = 0 \in \mathcal{H} \\
& \quad v_0 = w_t \\
& \quad \text{for } i = 1, \ldots, n \\
& \quad \quad v_i = v_{i-1} - (\gamma/n) \int_{\mathcal{H}} (\langle v_{i-1}, x \rangle - y) x \, d\rho(x, y) \\
& \quad w_{t+1} = v_n
\end{aligned}
$$

Note: step-size γ/n.

$$w_t \longrightarrow w_{t+1} \quad \longrightarrow \ldots$$

Stability and early stopping

Consider the gradient descent iteration for the expected risk.

$$
\begin{aligned}
& w_0 = 0 \in \mathcal{H} \\
& \quad v_0 = w_t \\
& \quad \text{for } i = 1, \dots, n \\
& \quad\quad v_i = v_{i-1} - (\gamma/n) \int_{\mathcal{H}} (\langle v_{i-1}, x\rangle - y) x \, d\rho(x, y) \\
& \quad w_{t+1} = v_n
\end{aligned}
$$

Note: step-size γ/n.

$$
w_t \longrightarrow w_{t+1} \quad \longrightarrow \dots \quad \longrightarrow \quad \longrightarrow \operatorname*{argmin}_{\mathcal{H}} \mathcal{E}
$$

Stability and early stopping

Consider the gradient descent iteration for the expected risk.

$$
\begin{aligned}
& w_0 = 0 \in \mathcal{H} \\
& \quad v_0 = w_t \\
& \quad \text{for } i = 1, \ldots, n \\
& \quad \quad v_i = v_{i-1} - (\gamma/n) \int_{\mathcal{H}} (\langle v_{i-1}, x\rangle - y) x \, d\rho(x, y) \\
& \quad w_{t+1} = v_n
\end{aligned}
$$

Note: step-size γ/n.

$$
w_t \longrightarrow w_{t+1} \quad \longrightarrow \ldots \quad \longrightarrow \quad \longrightarrow \operatorname*{argmin}_{\mathcal{H}} \mathcal{E}
$$

$$
\hat{w}_t \longrightarrow \hat{w}_{t+1} \quad \longrightarrow \ldots
$$

Stability and early stopping

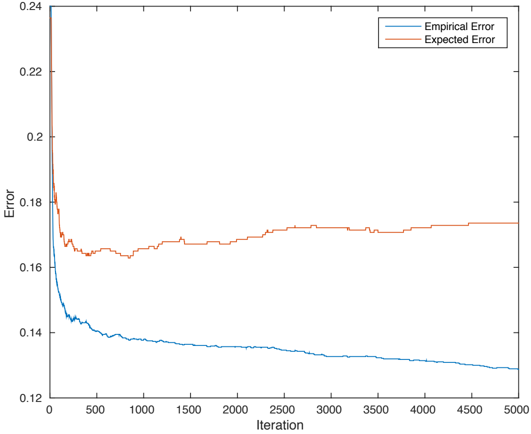
Consider the gradient descent iteration for the expected risk.

$$
\begin{array}{l}
w_0 = 0 \in \mathcal{H} \\
\left\lfloor \begin{array}{l}
v_0 = w_t \\
\text{for } i = 1, \dots, n \\
\left\lfloor \; v_i = v_{i-1} - (\gamma/n) \int_{\mathcal{H}} (\langle v_{i-1}, x \rangle - y) x \, d\rho(x,y) \right. \\
w_{t+1} = v_n
\end{array} \right.
\end{array}
$$

Note: step-size γ/n.

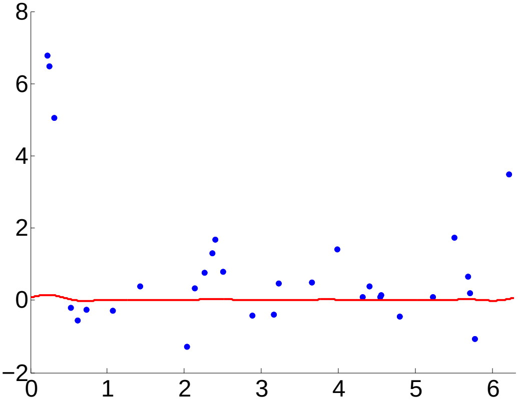
$$
\begin{array}{l}
w_t \longrightarrow w_{t+1} \quad \longrightarrow \dots \quad \longrightarrow \quad \longrightarrow \operatorname*{argmin}_{\mathcal{H}} \mathcal{E} \\
\hat{w}_t \longrightarrow \hat{w}_{t+1} \quad \longrightarrow \dots \\
\qquad\qquad\qquad\qquad \searrow \\
\qquad\qquad\qquad\qquad\quad \searrow \\
\qquad\qquad\qquad\qquad\qquad \operatorname*{argmin}_{\mathcal{H}} \hat{\mathcal{E}}
\end{array}
$$

Early stopping - semi-convergence



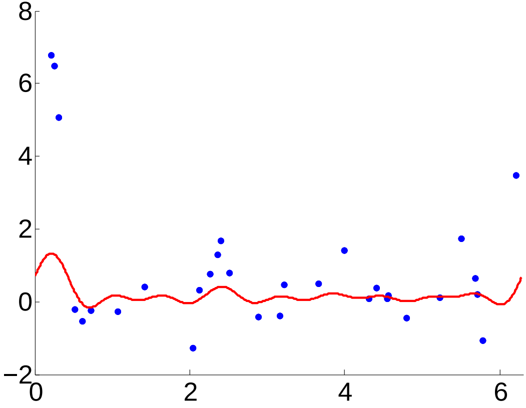
Early stopping - example

First epoch:



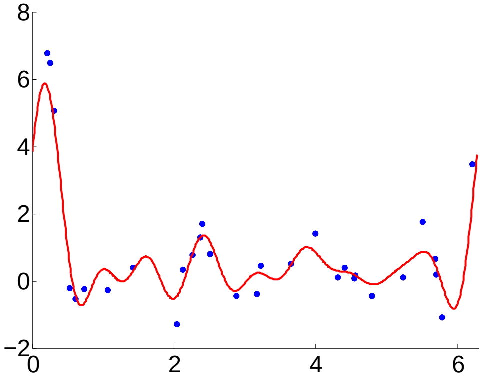
Early stopping - example

10-th epoch:



Early stopping - example

100-th epoch:



Main results: consistency

Theorem

Assume boundedness. Let $\gamma \in \left]0, \kappa^{-1}\right[$. Let $t_*(n)$ be such that

$$\mathbf{t_*(n) \to +\infty} \text{ and } \mathbf{t_*(n)\,(\log n/n)^{1/3} \to 0}$$

Assume $\mathcal{O} = \operatorname{argmin}_{\mathcal{H}} \mathcal{E} \neq \varnothing$. Then

$$\|\hat{w}_{t_*(n)} - w^\dagger\| \to 0 \quad \rho - \text{a.s.}$$

Main results: consistency

Theorem

Assume boundedness. Let $\gamma \in \left]0, \kappa^{-1}\right[$. Let $t_*(n)$ be such that

$$\mathbf{t_*(n) \to +\infty} \text{ and } \mathbf{t_*(n)\,(\log n/n)^{1/3} \to 0}$$

Assume $\mathcal{O} = \operatorname{argmin}_{\mathcal{H}} \mathcal{E} \neq \varnothing$. Then

$$\|\hat{w}_{t_*(n)} - w^\dagger\| \to 0 \quad \rho - \text{a.s.}$$

- universal step-size fixed a priori
- early stopping needed for consistency
- multiple passes are needed

First comparison with one pass stochastic gradient

Consistency in the following cases:

- **Multiple passes Stochastic Gradient method:**

$$\gamma_t = \gamma/n, \qquad t_*(n) \sim (n/\log n)^{1/3}$$

First comparison with one pass stochastic gradient

Consistency in the following cases:

- **Multiple passes Stochastic Gradient method:**

$$\gamma_t = \gamma/n, \qquad t_*(n) \sim (n/\log n)^{1/3}$$

- **One pass Stochastic Gradient method:**

$$\gamma_t = \gamma/\sqrt{n}, \qquad t_*(n) = 1\,(+\text{ averaging})$$

[$\sim$ Ying-Pontil 2008 and Dieuleveut-Bach 2014]

First comparison with one pass stochastic gradient

Consistency in the following cases:

- **Multiple passes Stochastic Gradient method:**

$$\gamma_t = \gamma/n, \qquad t_*(n) \sim (n/\log n)^{1/3}$$

- **One pass Stochastic Gradient method:**

$$\gamma_t = \gamma/\sqrt{n}, \qquad t_*(n) = 1\,(+\text{ averaging})$$

[$\sim$ Ying-Pontil 2008 and Dieuleveut-Bach 2014]

Why multiple passes make sense?

Main results: error bounds

Theorem

Assume boundedness and (SC) with $r \in \,]1/2, +\infty[$. If

$$\mathbf{t_*(n)} = \left\lceil \mathbf{n^{1/(2r+1)}} \right\rceil$$

then with high probability,

$$\|\hat{w}_t - w^\dagger\|_{\mathcal{H}} = \mathbf{O}\left(\mathbf{n}^{-\frac{\mathbf{r-1/2}}{\mathbf{2r+1}}}\right)$$

Main results: error bounds

Theorem

Assume boundedness and (SC) with $r \in]1/2, +\infty[$. If

$$\mathbf{t}_*(\mathbf{n}) = \left\lceil \mathbf{n}^{\mathbf{1}/(\mathbf{2r}+\mathbf{1})} \right\rceil$$

then with high probability,

$$\|\hat{w}_t - w^\dagger\|_{\mathcal{H}} = \mathbf{O}\left(\mathbf{n}^{-\frac{\mathbf{r}-\mathbf{1}/\mathbf{2}}{\mathbf{2r}+\mathbf{1}}}\right)$$

- optimal capacity independent rates for $\|\hat{w}_t - w^\dagger\|_{\mathcal{H}}$
- no saturation w.r.t. r
- the stopping rule depends on the source condition $\Rightarrow$ a balancing principle can be used

Stochastic gradient method (one pass)

$$\hat{\mathbf{w}}_{\mathbf{t}} = \hat{\mathbf{w}}_{\mathbf{t-1}} - \gamma_{\mathbf{t}}(\langle \hat{\mathbf{w}}_{\mathbf{t-1}}, \mathbf{x}_{\mathbf{t}} \rangle - \mathbf{y}_{\mathbf{t}})\mathbf{x}_{\mathbf{t}} + \lambda_{\mathbf{t}} \hat{\mathbf{w}}_{\mathbf{t-1}})$$

- classically studied in stochastic optimization, for strongly convex functions ($\lambda_t = 0$) (Robbins-Monro), in the finite dimensional setting

Stochastic gradient method (one pass)

$$\hat{\mathbf{w}}_{\mathbf{t}} = \hat{\mathbf{w}}_{\mathbf{t-1}} - \gamma_{\mathbf{t}}(\langle \hat{\mathbf{w}}_{\mathbf{t-1}}, \mathbf{x}_{\mathbf{t}} \rangle - \mathbf{y}_{\mathbf{t}})\mathbf{x}_{\mathbf{t}} + \lambda_{\mathbf{t}} \hat{\mathbf{w}}_{\mathbf{t-1}})$$

- classically studied in stochastic optimization, for strongly convex functions ($\lambda_t = 0$) (Robbins-Monro), in the finite dimensional setting
- In a RKHS, square loss first in [Smale-Yao 2006]

Stochastic gradient method (one pass) in RKHS - square loss

Assume **Source Condition**

- Let $r \in]1/2, +\infty[$ $\gamma_t = n^{-2r/(2r+1)}$, $\lambda_t = 0$, then [Ying-Pontil 2008]

$$\mathsf{E}\|w_t - w^\dagger\|_{\mathcal{H}} \leq ct^{-\frac{r-1/2}{2r+1}}$$

- Let $r \in]1/2, 1]$, let $\gamma_t = n^{-2r/(2r+1)}$, $\lambda_t = n^{-1/(2r+1)}$, then [Tarrès-Yao 2011]

$$\|w_t - w^\dagger\|_{\mathcal{H}} = O\left(t^{-\frac{r-1/2}{2r+1}}\right) \quad \text{with h. p.}$$

- Optimal rates, capacity dependent bounds in expectation on the risk [Dieuleveut-Bach 2014] (saturation for $r > 1$).

Comparison between multiple passes and one pass

There are **two** regimes with **optimal error bounds**

- **Multiple passes Stochastic Gradient method:**

$$\gamma_t = \gamma n^{-1} \qquad t_*(n) \sim n^{1/(2r+1)}$$

Comparison between multiple passes and one pass

There are **two** regimes with **optimal error bounds**

- **Multiple passes Stochastic Gradient method:**

$$\gamma_t = \gamma n^{-1} \qquad t_*(n) \sim n^{1/(2r+1)}$$

- **One pass Stochastic Gradient method:**

$$\gamma_t = \gamma n^{-2r/(2r+1)} \qquad t_*(n) = 1$$

Comparison between multiple passes and one pass

There are **two** regimes with **optimal error bounds**

- **Multiple passes Stochastic Gradient method:**

$$\gamma_t = \gamma n^{-1} \qquad t_*(n) \sim n^{1/(2r+1)}$$

- **One pass Stochastic Gradient method:**

$$\gamma_t = \gamma n^{-2r/(2r+1)} \qquad t_*(n) = 1$$

...in practice

model selction is needed and multiple passes stochastic gradient is a natural approach

Comparison with gradient descent learning

Let $\hat{w}_0 = 0$ and $\gamma > 0$. Iterate:

$$\hat{w}_{t+1} = \hat{w}_t - \gamma/n \sum_{i=1}^{n} (\langle \hat{w}_{t-1}, x_i \rangle - y_i) x_i$$

Comparison with gradient descent learning

Let $\hat{w}_0 = 0$ and $\gamma > 0$. Iterate:

$$\hat{w}_{t+1} = \hat{w}_t - \gamma/n \sum_{i=1}^{n} (\langle \hat{w}_{t-1}, x_i \rangle - y_i) x_i$$

Multiple passes Stochastc gradient vs. Gradient descent: same computational and statistical properties [Bauer-Pereverzev-Rosasco 2007],[Caponnetto-Yao 2008],[Raskutti-Wainwright-Yu 2013]

Proof: Bias-Variance trade-off

Define

$$
\begin{array}{l}
w_0 = 0 \in \mathcal{H} \\
\left\lfloor \begin{array}{l}
v_0 = w_t \\
\text{for } i = 1, \ldots, n \\
\left\lfloor \; v_i = v_{i-1} - (\gamma/n) \int_{\mathcal{H}} (\langle v_{i-1}, x \rangle - y) x \, d\rho(x, y) \right. \\
w_{t+1} = v_n
\end{array} \right.
\end{array}
$$

(w_t) is the nt-th gradient descent iteration with step-size γ/n on the risk.

Proof: Bias-Variance trade-off

Define

$$
\begin{array}{l}
w_0 = 0 \in \mathcal{H} \\
\left\lfloor \begin{array}{l}
v_0 = w_t \\
\text{for } i = 1, \ldots, n \\
\left\lfloor \; v_i = v_{i-1} - (\gamma/n) \int_{\mathcal{H}} (\langle v_{i-1}, x \rangle - y) x \, d\rho(x, y) \right. \\
w_{t+1} = v_n
\end{array} \right.
\end{array}
$$

(w_t) is the nt-th gradient descent iteration with step-size γ/n on the risk. Then

$$
\|\hat{w}_t - w^\dagger\|_{\mathcal{H}} \leq \underbrace{\|\hat{w}_t - w_t\|_{\mathcal{H}}}_{\textbf{Variance}} + \underbrace{\|w_t - w^\dagger\|_{\mathcal{H}}}_{\textbf{Bias}=\textbf{Optimization}}
$$

Variance - Step 1

$\hat{w}_t$ can be written as a **perturbed gradient descent** iteration on the empirical risk

$$\hat{w}_{t+1} = (I - \gamma \hat{T})\hat{w}_t + \gamma \hat{g} + \gamma^2 (\hat{A}\hat{w}_t - \hat{b})$$

with

- $\hat{T} = (1/n) \sum_{i=1}^n x_i \otimes x_i$, where $(x \otimes x)w = \langle w, x \rangle x$

Variance - Step 1

$\hat{w}_t$ can be written as a **perturbed gradient descent** iteration on the empirical risk

$$\hat{w}_{t+1} = (I - \gamma \hat{T})\hat{w}_t + \gamma \hat{g} + \gamma^2 (\hat{A}\hat{w}_t - \hat{b})$$

with

- $\hat{T} = (1/n)\sum_{i=1}^n x_i \otimes x_i$, where $(x \otimes x)w = \langle w, x \rangle\, x$
- $\hat{g} = (1/n)\sum_{i=1}^n y_i x_i$

Variance - Step 1

$\hat{w}_t$ can be written as a **perturbed gradient descent** iteration on the empirical risk

$$\hat{w}_{t+1} = (I - \gamma \hat{T})\hat{w}_t + \gamma \hat{g} + \gamma^2(\hat{A}\hat{w}_t - \hat{b})$$

with

- $\hat{T} = (1/n)\sum_{i=1}^n x_i \otimes x_i$, where $(x \otimes x)w = \langle w, x \rangle x$
- $\hat{g} = (1/n)\sum_{i=1}^n y_i x_i$
- $\hat{A} = \frac{1}{n^2}\sum_{k=2}^{n}\left[\prod_{i=k+1}^{n}\left(I - \frac{\gamma}{n} x_i \otimes x_i\right)\right](x_k \otimes x_k)\sum_{j=1}^{k-1} x_j \otimes x_j$

Variance - Step 1

$\hat{w}_t$ can be written as a **perturbed gradient descent** iteration on the empirical risk

$$\hat{w}_{t+1} = (I - \gamma \hat{T})\hat{w}_t + \gamma \hat{g} + \gamma^2 (\hat{A}\hat{w}_t - \hat{b})$$

with

- $\hat{T} = (1/n)\sum_{i=1}^n x_i \otimes x_i$, where $(x \otimes x)w = \langle w, x \rangle\, x$
- $\hat{g} = (1/n)\sum_{i=1}^n y_i x_i$
- $\hat{A} = \dfrac{1}{n^2}\displaystyle\sum_{k=2}^{n}\left[\prod_{i=k+1}^{n}\left(I - \frac{\gamma}{n} x_i \otimes x_i\right)\right](x_k \otimes x_k)\sum_{j=1}^{k-1} x_j \otimes x_j$
- $\hat{b} = \dfrac{1}{n^2}\displaystyle\sum_{k=2}^{n}\left[\prod_{i=k+1}^{n}\left(I - \frac{\gamma}{n} x_i \otimes x_i\right)\right](x_k \otimes x_k)\sum_{j=1}^{k-1} y_j x_j$

Variance - Step 1

w_t is a **perturbed gradient descent** iteration with step γ on the risk

$$w_{t+1} = (I - \gamma T)w_t + \gamma g + \gamma^2(Aw_t - b)$$

with

- $T = \mathrm{E}[x \otimes x]$,
- $g = \mathrm{E}[g_\rho(x)x]$

Variance - Step 1

w_t is a **perturbed gradient descent** iteration with step γ on the risk

$$w_{t+1} = (I - \gamma T)w_t + \gamma g + \gamma^2(Aw_t - b)$$

with

- $T = \mathrm{E}[x \otimes x]$,
- $g = \mathrm{E}[g_\rho(x)x]$
- $A = \frac{1}{n^2}\sum_{k=2}^{n}\left[\prod_{i=k+1}^{n}\left(I - \frac{\gamma}{n}T\right)\right] T \sum_{j=1}^{k-1} T$

Variance - Step 1

w_t is a **perturbed gradient descent** iteration with step γ on the risk

$$w_{t+1} = (I - \gamma T)w_t + \gamma g + \gamma^2 (A w_t - b)$$

with

- $T = \mathrm{E}[x \otimes x]$,
- $g = \mathrm{E}[g_\rho(x)x]$
- $A = \dfrac{1}{n^2} \displaystyle\sum_{k=2}^{n} \left[\prod_{i=k+1}^{n} \left(I - \frac{\gamma}{n} T \right) \right] T \sum_{j=1}^{k-1} T$
- $b = \dfrac{1}{n^2} \displaystyle\sum_{k=2}^{n} \left[\prod_{i=k+1}^{n} \left(I - \frac{\gamma}{n} T \right) \right] T \sum_{j=1}^{k-1} g$

Variance - Step 2

$$\|\hat{w}_t - w_t\|_{\mathcal{H}} \leq \gamma \sum_{k=0}^{t-1} \left\| (I - \gamma \hat{T} + \gamma^2 \hat{A})^{t-k+1} \right\|$$

$$\left\| (T - \hat{T}) w_k + \gamma(\hat{A} - A) w_k + (\hat{g} - g) - \gamma(\hat{b} - b) \right\|$$

$$\leq \gamma \sum_{k=0}^{t-1} (\|T - \hat{T}\| + \gamma \underbrace{\|\hat{A} - A\|}_{\text{sum of martingales}}) \underbrace{\|w_k\|_{\mathcal{H}}}_{\textit{bounded}} + \|\hat{g} - g\|_{\mathcal{H}} + \gamma \underbrace{\|\hat{b} - b\|_{\mathcal{H}}}_{\text{sum of martingales}}$$

+ Pinelis concentration inequality

$$\leq c_1 \frac{\log(16/\delta) t}{\sqrt{n}}$$

Bias

- Convergence results for the gradient descent applied to $\mathcal{E}$
- Standard approach based on spectral calculus (square loss used here!)
- Convergence depends on the step-size $\gamma \in \left]0, n\kappa^{-1}\right[$ and the source condition

$$\|w_t - w^\dagger\|_{\mathcal{H}} \leq c \left(\frac{r - 1/2}{\gamma t}\right)^{r-1/2}$$

Bias-Variance trade-off - again

Then, with probability greater than $1-\delta$,

$$\|\hat{w}_t - w^\dagger\|_{\mathcal{H}} \leq \log\left(\frac{16}{\delta}\right) c_1 \mathbf{t}\mathbf{n}^{-\mathbf{1/2}} + c_2 \mathbf{t}^{\mathbf{1/2-r}}$$

Contributions and future work

Contributions

- first results on generalization properties of multiple passes stochastic gradient method
- results support commonly used heuristics, e.g. early stopping

Contributions and future work

Contributions

- first results on generalization properties of multiple passes stochastic gradient method
- results support commonly used heuristics, e.g. early stopping

Future work

- extension to other losses and sampling techniques [Lin-Rosasco 2016]
- capacity dependent bounds and optimal bounds for the risk
- unified analysis for one pass and multiple passes

References

L. Rosasco, S. Villa
Learning with incremental iterative Regularization, NIPS 2015

Finite sample bounds for the risk

Corollary

Assume boundedness and source condition with $r \in]1/2, +\infty[$. Then, Choosing $\mathbf{t_*(n)} = \lceil \mathbf{n^{1/2(r+1)}} \rceil$, with high probability

$$\mathcal{E}(\hat{w}_t) - \inf_{\mathcal{H}} \mathcal{E} = \mathbf{O}\left(\mathbf{n}^{-\frac{\mathbf{r}}{\mathbf{r+1}}}\right)$$

Finite sample bounds for the risk

Corollary

Assume boundedness and source condition with $r \in]1/2, +\infty[$. Then, Choosing $\mathbf{t_*(n) = \lceil n^{1/2(r+1)} \rceil}$, with high probability

$$\mathcal{E}(\hat{w}_t) - \inf_{\mathcal{H}} \mathcal{E} = \mathbf{O}\left(\mathbf{n}^{-\frac{\mathbf{r}}{\mathbf{r+1}}}\right)$$

The rates are not optimal...
but valid under more general source condition (even in the nonattainable case)

Incremental gradient in a RKHS

$\mathcal{H}$ RKHS of functions from $\mathcal{X}$ to $\mathcal{Y}$ with kernel $K\colon \mathcal{X} \times \mathcal{X} \to R$. Let $\hat{w}_0 = 0$, then

$$\hat{w}_t = \sum_{k=1}^{n} (\alpha_t)_k K_{x_k}$$

where $\alpha_t = ((\alpha_t)_1, \dots, (\alpha_t)_n) \in \mathbb{R}^n$ satisfy

$$\alpha_{t+1} = c_t^n$$

$$c_t^0 = \alpha_t, \quad (c_t^i)_k = \begin{cases} (c_t^{i-1})_k - \dfrac{\gamma}{n} \left(\sum_{j=1}^{n} K(x_i, x_j)(c_t^{i-1})_j - y_i \right), & k = i \\ (c_t^{i-1})_k, & k \neq i \end{cases}$$