Gradient Descent Method in Learning

online vs. batch[a]

Yuan Yao

Department of Mathematics,
University of California, Berkeley

[a]Some joint work with Andrea Caponnetto, Lorenzo Rosasco, Steve Smale, Pierre Tarrès, with help from Yiming Ying and D.-X. Zhou, etc.

Outline of the talk

- Batch vs Online learning

Outline of the talk

- Batch vs Online learning
- Gradient Descent Method in both settings

Outline of the talk

- Batch vs Online learning
- Gradient Descent Method in both settings
- How to do regularizations

Outline of the talk

- Batch vs Online learning
- Gradient Descent Method in both settings
- How to do regularizations
- Lower and Upper Bounds

Outline of the talk

- Batch vs Online learning
- Gradient Descent Method in both settings
- How to do regularizations
- Lower and Upper Bounds
- “Random Connections” with Random Projections

Outline of the talk

- Batch vs Online learning
- Gradient Descent Method in both settings
- How to do regularizations
- Lower and Upper Bounds
- “Random Connections” with Random Projections
- Future Directions

Batch vs. Online

Given a sequence of examples $(z_i)_{i\in\mathbb{N}} \in (\mathcal{X} \times \mathcal{Y})^\infty$

- Batch Learning: truncation set $\mathbf{z}_T = (z_i)_{i=1}^T$, find a mapping

$$\mathbf{z}_T \mapsto f_{\mathbf{z}_T} \in \mathscr{H}$$

- Online Learning: a Markov Decision Process

$$f_{t+1} = T_t(f_t, z_{t+1})$$

where f_t only depends on $z_1, \ldots, z_t$.

Why Online?

- Low computational cost:
  online needs $\geq O(t)$ steps
  batch typically needs $\geq O(T^3)$ (inverting a matrix)
- Fast convergence: order optimality
- Temporal dependence of samples:
  Markov Chain sampling: large-scale biological networks
  Mixing processes: exponential-mixing and polynomial-mixing
  Games: competitive (non-statistical) analysis, etc.

Where we start...

$$\min_{f \in \mathscr{H}} \frac{1}{T} \sum_{i=1}^{T} V(y_i, f(x_i)) + \lambda \|f\|_{\mathscr{H}}^2$$

where we choose $V(y, f(x))$:

- L_2 loss: for order optimality analysis
- L_1 loss (soft margin): for sparsity, e.g. Basis Pursuit and SVM regression

continued...

and $\mathscr{H} = \mathscr{H}_K$ a RKHS such that the gradient map takes a simple form

$$\begin{aligned} \mathrm{grad} V : \mathscr{H}_K &\rightarrow \mathscr{H}_K \\ f &\mapsto V'_f(y, f(x)) K_x. \end{aligned}$$

continued...

and $\mathscr{H} = \mathscr{H}_K$ a RKHS such that the gradient map takes a simple form

$$\begin{aligned} \mathrm{grad} V : \mathscr{H}_K &\to \mathscr{H}_K \\ f &\mapsto V'_f(y, f(x))K_x. \end{aligned}$$

Note: when V is non-differentiable, V'_f is understood to be a *subgradient*. Singularities of V are designed to obtain *sparse* solutions.

RKHS

- $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is a *Mercer* kernel, i.e. a *continuous*, symmetric and positive definite function
- $\mathscr{H}_K = \overline{\text{span}\{K_x : x \in \mathcal{X}\}}$ where the closure is w.r.t. the inner product as the linear extension of $\langle K_x, K_{x'} \rangle_K = K(x, x')$
- *Reproducing* property: $f(x) = \langle f, K_x \rangle_K$
- $\mathscr{H}_K$ is a subspace (closed iff finite dimension) in $\mathscr{L}^2_{\rho_X} \cap \mathscr{C}(\mathcal{X})$
- $\mathscr{H}_K$ can be dense in $\mathscr{L}^2_{\rho_X}$, e.g. Gaussian kernel $K(x,t) = e^{-a\|x-t\|^2}$ ($a > 0$)

Gradient Descent Algorithms

For L_2 loss and $\mathscr{H} = \mathscr{H}_K$,

- Batch:

$$\hat{f}_{t+1} = \hat{f}_t - \eta_t \left[\frac{1}{T} \sum_{i=1}^{T} (\hat{f}_t(x_i) - y_i) K_{x_i} + \lambda_T \hat{f}_t \right]$$

- Online:

$$f_{t+1} = f_t - \eta_t [(f_t(x_{t+1}) - y_{t+1}) K_{x_{t+1}} + \lambda_t f_t]$$

Our Theoretical Goal

Convergence of $(\hat{f}_t) \in \mathscr{H}_K$ and $(f_t) \in \mathscr{H}_K$ to the regression function

$$f_\rho(x) := \mathbb{E}[y|x] \in \mathscr{L}^2_{\rho_X}$$

and its rates when f_ρ takes some sparse form.

Our Theoretical Goal

Convergence of $(\hat{f}_t) \in \mathscr{H}_K$ and $(f_t) \in \mathscr{H}_K$ to the regression function

$$f_\rho(x) := \mathbb{E}[y|x] \in \mathscr{L}^2_{\rho_X}$$

and its rates when f_ρ takes some sparse form.

But, $\mathscr{L}^2_{\rho_X}$ is too large a space to search, so we need *regularizations*.

Regularization

Two parameters: λ_t (or λ_T) and η_t:

- $\lambda_T = 0$ and $\eta_t = c$: Landwebter iterations
- $\lambda_T = 0$ and $\eta_t \downarrow 0$: Yao et al. (2005)
- $\lambda_t = \lambda > 0$ and $\eta_t \downarrow 0$: $f_t \to f_\lambda \neq f_\rho$, Smale and Yao (2005) etc.
- $\lambda_t \downarrow 0$ and $\eta_t \downarrow 0$: $f_t \to f_\rho$, Yao and Tarrès (2005)
- $\lambda_t = 0$ and $\eta_t \downarrow 0$: $f_t \to f_\rho$, Ying et al. (2006)

Sparsity of Regression Function

We are going to assume that the regression function is sparse/smooth w.r.t. the following *basis*

- roughly speaking, kernel principle components,
- or more precisely, the eigenfunctions of the *covariance operator* of $\rho_{\mathcal{X}}$ on $\mathscr{H}_K$.

Covariance operator

- Define an integral operator

$$\begin{aligned} L_K : \mathscr{L}^2_{\rho_X} &\rightarrow \mathscr{H}_K \\ f &\mapsto \int_X f(x')K(x', \cdot)d\rho_X \end{aligned}$$

Covariance operator

- Define an integral operator

$$\begin{aligned} L_K : \mathscr{L}^2_{\rho_X} &\to \mathscr{H}_K \\ f &\mapsto \int_X f(x')K(x',\cdot)d\rho_X \end{aligned}$$

- The *covariance operator*, is the restriction $L_K|_{\mathscr{H}_K} : \mathscr{H}_K \to \mathscr{H}_K$, i.e. $\mathbb{E}_x[\langle\ , K_x\rangle K_x]$

Covariance operator

- Define an integral operator

$$\begin{aligned} L_K : \mathscr{L}^2_{\rho_X} &\to \mathscr{H}_K \\ f &\mapsto \int_X f(x')K(x',\cdot)d\rho_X \end{aligned}$$

- The *covariance operator*, is the restriction $L_K|_{\mathscr{H}_K} : \mathscr{H}_K \to \mathscr{H}_K$, i.e. $\mathbb{E}_x[\langle\ ,K_x\rangle K_x]$
- $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$ compact $\Rightarrow$ orthonormal eigen-system $(\lambda_i, \phi_i)_{i\in\mathbb{N}}$, $\phi_i \in \mathscr{L}^2_{\rho_X} \cap \mathscr{H}_K$ bi-orthogonal and

$$\sum \lambda_i \le \sup_{x\in\mathcal{X}} K(x,x) =: \kappa < \infty$$

Sparsity Assumption

Assume that

$$f_\rho = L_K^r g, \quad g \in \mathscr{L}^2_{\rho_X}, r > 0$$

i.e. f_ρ has at least *power-law decay* coordinates w.r.t. the basis of eigenfunctions of $L_K : \mathscr{L}^2_{\rho_X} \to \mathscr{L}^2_{\rho_X}$:

$$f_\rho = \sum_i \lambda_i^r g_i \phi_i,$$

$$\sum \lambda_i \leq \kappa < \infty, \sum g_i^2 < \infty$$

Lower Rates in Learning

Let $\mathbb{P}(b, r)$ ($b > 1$ and $r \in (1/2, 1]$) be the set of probability measure ρ on $\mathcal{X} \times \mathcal{Y}$, such that:

- almost surely $|y| \leq M_\rho$
- $f_\rho = L_K^r g$ for some $g \in \mathscr{L}^2_{\rho_X}$
- the eigenvalues λ_i, arranged in a nonincreasing order, decay at $O(i^{-b})$

...Minimax Lower Rates

[Caponnetto-DeVito'05] The minimax lower rate:

$$\liminf_{t\to\infty} \inf_{\mathbf{z}_t \mapsto f_t} \sup_{\rho\in\mathbb{P}(b,r)} \mathbf{Prob}\left\{\mathbf{z}_t \in \mathbb{Z}^t : \frac{\|f_t - f_\rho\|_\rho}{t^{-\frac{rb}{2rb+1}}} > C\right\} = 1$$

where the $\inf$ is taken over all algorithms mapping $(z_i)_1^t \mapsto f_t$.

...Minimax Lower Rates

[Caponnetto-DeVito'05] The minimax lower rate:

$$\liminf_{t\to\infty} \inf_{\mathbf{z}_t \mapsto f_t} \sup_{\rho\in\mathbb{P}(b,r)}$$
$$\mathbf{Prob}\left\{\mathbf{z}_t \in \mathbb{Z}^t : \frac{\|f_t - f_\rho\|_\rho}{t^{-\frac{rb}{2rb+1}}} > C\right\} = 1$$

where the $\inf$ is taken over all algorithms mapping $(z_i)_1^t \mapsto f_t$.

- The ρ in $\sup_{\rho\in\mathbb{P}(b,r)}$ depends on sample size t!

...Minimax Lower Rates

[Caponnetto-DeVito'05] The minimax lower rate:

$$\liminf_{t\to\infty} \inf_{\mathbf{z}_t \mapsto f_t} \sup_{\rho\in\mathbb{P}(b,r)}$$
$$\mathbf{Prob}\left\{\mathbf{z}_t \in \mathbb{Z}^t : \frac{\|f_t - f_\rho\|_\rho}{t^{-\frac{rb}{2rb+1}}} > C\right\} = 1$$

where the $\inf$ is taken over all algorithms mapping $(z_i)_1^t \mapsto f_t$.

- The ρ in $\sup_{\rho\in\mathbb{P}(b,r)}$ depends on sample size t!
- Not suitable for batch learning, but ok for online learning.

...Individual Lower Rates

[Caponnetto-DeVito'05] The individual lower rate: for each $B > b$,

$$\inf_{((z_i)_1^t \mapsto f_t)_{t\in\mathbb{N}}} \sup_{\rho\in\mathbb{P}(b,r)} \limsup_{t\to\infty} \frac{\|f_t - f_\rho\|_\rho}{t^{-\frac{rB}{2rB+1}}} > 0.$$

...Individual Lower Rates

[Caponnetto-DeVito'05] The individual lower rate: for each $B > b$,

$$\inf_{((z_i)_1^t \mapsto f_t)_{t\in\mathbb{N}}} \sup_{\rho\in\mathbb{P}(b,r)} \limsup_{t\to\infty} \frac{\|f_t - f_\rho\|_\rho}{t^{-\frac{rB}{2rB+1}}} > 0.$$

Note: taking $b = 1$ and $B = 1$, it suggests *eigenvalue independent* minimax and individual lower rates:

$$t^{-\frac{r}{2r+1}}$$

Upper Bounds for Batch Learning

Theorem (Yao-Rosasco-Caponnetto'05). Assume that $f_\rho = L_K^r g$ ($r > 0$). There exist λ_T, η_t and an early stopping rule $t^* : \mathbb{N} \to \mathbb{N}$, such that

- if $r > 0$, $\|\hat{f}_{t^*(T)} - f_\rho\|_\rho \leq O(T^{-\frac{r}{2r+2}})$
- if $r > 1/2$, $\|\hat{f}_{t^*(T)} - f_\rho\|_K \leq O(T^{-\frac{r-1/2}{2r+2}})$

In fact, one may choose $\lambda_T = 0$, $\eta_t = \frac{1}{\kappa^2(t+1)^\theta}$ and $t^*(T) = \lceil T^{-\frac{1}{(2r+2)(1-\theta)}} \rceil$.

Improvements

[Bauer-Pereverzev-Rosasco'06] For $\theta = 0$ and $r > 1/2$,

$$\|\hat{f}_{t^*(T)} - f_\rho\|_\rho \leq O\big(T^{-\frac{r}{2r+1}}\big)$$

which meets the lower rates.

Upper Bounds for Online Learning

Theorem (Tarrès-Yao'06). Assume that $f_\rho = L_K^r g$ $(r > 0)$. There exist λ_t and η_t such that

- if $r > 0$, $\|f_t - f_\rho\|_\rho \leq O(t^{-\max\{\frac{r}{2r+1}, 1/3\}})$
- if $r > 1/2$, $\|f_t - f_\rho\|_K \leq O(t^{-\max\{\frac{r-1/2}{2r+1}, 1/4\}})$

In fact, $\lambda_t \sim O(t^{-1/(2r+1)})$ and $\eta_t \sim O(t^{-2r/(2r+1)})$.

Upper Bounds for Online Learning

Theorem (Tarrès-Yao'06). Assume that $f_\rho = L_K^r g$ $(r > 0)$. There exist λ_t and η_t such that

- if $r > 0$, $\|f_t - f_\rho\|_\rho \leq O(t^{-\max\{\frac{r}{2r+1}, 1/3\}})$
- if $r > 1/2$, $\|f_t - f_\rho\|_K \leq O(t^{-\max\{\frac{r-1/2}{2r+1}, 1/4\}})$

In fact, $\lambda_t \sim O(t^{-1/(2r+1)})$ and $\eta_t \sim O(t^{-2r/(2r+1)})$.

Note: the upper rates *saturate* when $r \geq 1$ and $r \geq 3/2$!

Breaking Saturation

It is expected that with $\lambda_t = 0$ and suitable choices $\eta_t \to 0$ and $\sum_t \eta_t = \infty$, one has

$$\|f_t - f_\rho\|_\rho \leq O\left(t^{-\frac{r}{2r+1}}\right)$$

for *all* $r > 0$.

Breaking Saturation

It is expected that with $\lambda_t = 0$ and suitable choices $\eta_t \to 0$ and $\sum_t \eta_t = \infty$, one has

$$\|f_t - f_\rho\|_\rho \leq O\big(t^{-\frac{r}{2r+1}}\big)$$

for *all* $r > 0$.

A Positive Answer: Ying et al. (2006) give results suggesting its truth.

Random Projection Perspective

Given $\mathbf{x}_T \in \mathcal{X}^T$, define a sampling operator on $\mathscr{H}_K$

$$\begin{aligned} S_{\mathbf{x}_T} : \mathscr{H}_K &\rightarrow l_2(\mathbf{x}_T) \\ f &\mapsto (f(x_i))_1^T = (\langle f, K_{x_i} \rangle_K)_1^T \end{aligned}$$

Random Projection Perspective

Given $\mathbf{x}_T \in \mathcal{X}^T$, define a sampling operator on $\mathscr{H}_K$

$$\begin{aligned} S_{\mathbf{x}_T} : \mathscr{H}_K &\to l_2(\mathbf{x}_T) \\ f &\mapsto (f(x_i))_1^T = (\langle f, K_{x_i} \rangle_K)_1^T \end{aligned}$$

- $S_{\mathbf{x}_T} f$ takes T random measurements/projections of f.

Random Projection Perspective

Given $\mathbf{x}_T \in \mathcal{X}^T$, define a sampling operator on $\mathscr{H}_K$

$$\begin{aligned} S_{\mathbf{x}_T} : \mathscr{H}_K &\rightarrow l_2(\mathbf{x}_T) \\ f &\mapsto (f(x_i))_1^T = (\langle f, K_{x_i} \rangle_K)_1^T \end{aligned}$$

- $S_{\mathbf{x}_T} f$ takes T random measurements/projections of f.
- Adjoint operator $S^*_{\mathbf{x}_T}\mathbf{y} = \frac{1}{T}\sum_{i=1}^T y_i K_{x_i}$.

Random Projection Perspective

Given $\mathbf{x}_T \in \mathcal{X}^T$, define a sampling operator on $\mathscr{H}_K$

$$\begin{aligned} S_{\mathbf{x}_T} : \mathscr{H}_K &\rightarrow l_2(\mathbf{x}_T) \\ f &\mapsto (f(x_i))_1^T = (\langle f, K_{x_i} \rangle_K)_1^T \end{aligned}$$

- $S_{\mathbf{x}_T} f$ takes T random measurements/projections of f.
- Adjoint operator $S^*_{\mathbf{x}_T} \mathbf{y} = \frac{1}{T} \sum_{i=1}^T y_i K_{x_i}$.
- $S_{\mathbf{x}_T} S^*_{\mathbf{x}_T}$ is the Gram matrix $(K(x_i, x_j))^{T \times T}$.

Compressed Sensing

- f is sparse w.r.t. certain basis/frames (unknown)
- $S_{\mathbf{x}_T}$ takes some random measurements of f such that the Uniform Uncertainty Principle holds, or equivalently, for small enough T_0 and all $T \leq T_0$, $S_{\mathbf{x}_T} S^*_{\mathbf{x}_T}$ has a *uniform lower bound* (depending on the sparsity of f) on the smallest eigenvalue.

Compressed Sensing

- f is sparse w.r.t. certain basis/frames (unknown)
- $S_{\mathbf{x}_T}$ takes some random measurements of f such that the Uniform Uncertainty Principle holds, or equivalently, for small enough T_0 and all $T \leq T_0$, $S_{\mathbf{x}_T} S^*_{\mathbf{x}_T}$ has a *uniform lower bound* (depending on the sparsity of f) on the smallest eigenvalue.

However in Learning, since

$$\mathbb{E}[S^*_{\mathbf{x}_T} S_{\mathbf{x}_T}] = L_K|_{\mathscr{H}_K}$$

where L_K is a compact operator with eigenvalues convergent to 0, NO lower bound!

Learning vs. Compressed Sensing

To control the *condition number* (or smallest eigenvalue) of the Gram matrix $S_{\mathbf{x}_T} S_{\mathbf{x}_T}^*$:

- Learning uses *regularization*
- Compressed sensing uses *Random Matrix Theory*

Learning vs. Compressed Sensing

To control the *condition number* (or smallest eigenvalue) of the Gram matrix $S_{\mathbf{x}_T} S_{\mathbf{x}_T}^*$:

- Learning uses *regularization*
- Compressed sensing uses *Random Matrix Theory*

Moreover, there is another kind of "condition number" in machine learning:

margin

Margin

Definitions.

- $f \in \mathscr{H}_K$ has *margin* $\gamma > 0$, if

$$\rho_X\{x \in X : \angle(f, K_x) \geq \arccos \gamma\} = 1$$

- $f \in \mathscr{H}_K$ has *margin* $\gamma > 0$ with error $\epsilon \in [0,1]$, if

$$\rho_X\{x \in X : \angle(f_t, K_x) \geq \arccos \gamma\} \geq 1 - \epsilon$$

Margin

Definitions.

- $f \in \mathscr{H}_K$ has *margin* $\gamma > 0$, if

$$\rho_X\{x \in X : \angle(f, K_x) \geq \arccos\gamma\} = 1$$

- $f \in \mathscr{H}_K$ has *margin* $\gamma > 0$ with error $\epsilon \in [0, 1]$, if

$$\rho_X\{x \in X : \angle(f_t, K_x) \geq \arccos\gamma\} \geq 1 - \epsilon$$

Note: $f \in \mathscr{H}_K$ has margin $\gamma > 0$ simply says that f can't *jump* arbitrarily small at zero value, i.e.

$$|f(x)| \geq \gamma\|f\|\|K_x\|$$

Margin and Random Projections

[Balcan-Blum-Vempala'05] If $f \in \mathscr{H}_K$ has margin γ, then with i.i.d. examples of number

$$t \geq \frac{8}{\epsilon} \max \left\{ \frac{1}{\gamma^2}, \ln \frac{1}{\delta} \right\}$$

there is a f_t such that with confidence $1 - \delta$, f_t has margin $\gamma/2$ with error ϵ.

Margin and Random Projections

[Balcan-Blum-Vempala'05] If $f \in \mathscr{H}_K$ has margin γ, then with i.i.d. examples of number

$$t \geq \frac{8}{\epsilon} \max\left\{\frac{1}{\gamma^2}, \ln\frac{1}{\delta}\right\}$$

there is a f_t such that with confidence $1-\delta$, f_t has margin $\gamma/2$ with error ϵ.

- In fact, f_t can be realized by the *Gram-Schmidt Orthonormalization*.

Future Directions

- Step-Size Adaptation
  - Cross-Validation
  - Averaging process acceleration
  - Stochastic Meta-Descent (SMD)
- Dependent Sampling
  - Markov Chain sampling
  - Mixing process
- Various aspects of Random Projections
- Applications in time series, etc.