Learning at the Turing

The Alan Turing Institute

Mathematics of Machine Learning - Summer School

Lecture 7
Stochastic Methods. Algorithmic Stability

July 1, 2021

Patrick Rebeschini
Department of Statistics, University of Oxford

Statistical/Computational Learning Theory (Lecture 1)

Problem formulation (out-of-sample prediction):

- Given n data $(X_1, Y_1), \dots, (X_n, Y_n) \in \mathbb{R}^d \times \mathbb{R}$ i.i.d. from $\mathbf{P}$ (**unknown**)
- Consider the *population risk* $r(a) = \mathbf{E}\, \phi(a(X), Y)$

Goal: Compute $A \in \sigma\{(X_i, Y_i)_{i=1}^n\}$ such that $\underbrace{r(A) - \inf_a r(a)}_{\text{excess risk}}$ is **small**

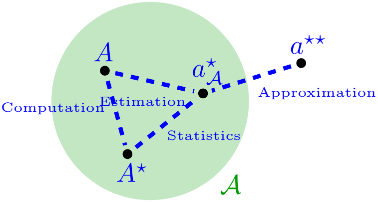
What does it mean to solve the problem **optimally**?

- **Statistics:** A is minimax-optimal w.r.t. the class of distrib. $\mathcal{P}$ if

$$\mathbf{E}\, r(A) - \inf_a r(a) \quad \sim \inf_{A \in \sigma\{Z_1, \dots, Z_n\}} \sup_{\mathbf{P} \in \mathcal{P}} \Big\{ \mathbf{E}\, r(A) - \inf_a r(a) \Big\}$$

- **Runtime:** Computing A takes same time to read the data, i.e. $O(nd)$ cost
- **Memory:** Storing $O(1)$ data point at a time, i.e. $O(d)$ storage cost
- **Distributed computations:** Runtime $O(1/m)$ if we have m machines
- (communication, privacy, robustness...)

Explicit regularization: uniform convergence (Lecture 1)



- Estimation/approximation: $r(A) - r(a^{\star\star}) = \underbrace{r(A) - r(a^{\star})}_{\text{Estimation}} + \underbrace{r(a^{\star}) - r(a^{\star\star})}_{\text{Approximation}}$

- Classical error decomposition for estimation error:

$$\underbrace{r(A) - r(a^{\star})}_{\text{Estimation}} = r(A) - R(A) + R(A) - R(A^{\star}) + \underbrace{R(A^{\star}) - R(a^{\star})}_{\leq 0} + R(a^{\star}) - r(a^{\star})$$

$$\boxed{r(A) - r(a^{\star\star}) \leq \underbrace{2 \sup_{a \in \mathcal{A}} |r(a) - R(a)|}_{\text{Statistics}} + \underbrace{R(A) - R(A^{\star})}_{\text{Computation}} + \underbrace{r(a^{\star}) - r(a^{\star\star})}_{\text{Approximation}}}$$

Recall: Subgradient Method with Euclidean Geometry

Risk minimization:

$$\underset{w}{\text{minimize}} \quad r(w) = \mathbf{E}\varphi(w^\top XY) \qquad \Longrightarrow \qquad \text{Let } w^\star \text{ be a minimizer}$$
$$\text{subject to} \quad \|w\|_2 \le c_2^{\mathcal{W}}$$

Empirical risk minimization:

$$\underset{w}{\text{minimize}} \quad R(w) = \frac{1}{n}\sum_{i=1}^{n} \varphi(w^\top X_i Y_i) \qquad \Longrightarrow \qquad \text{Let } W^\star \text{ be a minimizer}$$
$$\text{subject to} \quad \|w\|_2 \le c_2^{\mathcal{W}}$$

$$\boxed{r(\overline{W}_t) - r(w^\star) \le \underbrace{R(\overline{W}_t) - R(W^\star)}_{\texttt{Optimization}} + \underbrace{\sup_{w\in\mathcal{W}}\{r(w) - R(w)\} + \sup_{w\in\mathcal{W}}\{R(w) - r(w)\}}_{\texttt{Statistics}}}$$

$$\boxed{\mathbf{E}\,\texttt{Statistics} \le \frac{4c_2^{\mathcal{X}} c_2^{\mathcal{W}} \gamma_\varphi}{\sqrt{n}}} \qquad \boxed{\texttt{Optimization} \le \frac{2c_2^{\mathcal{X}} c_2^{\mathcal{W}} \gamma_\varphi}{\sqrt{t}}}$$

It seems a complete story but... what about the computational cost?

Computational Complexity and Stochastic Oracle Model

- Each subgradient computation costs $O(n)$ (prohibitive if n is large):

$$\partial R(w) = \frac{1}{n}\sum_{i=1}^{n}\partial_w \varphi(w^\top X_i Y_i)$$

- **Wish:** Can we use approximate/noisy subgradients and prove

$$\boxed{\mathbf{E}\,\texttt{Optimization} \le \frac{2c_2^{\mathcal{X}} c_2^{\mathcal{W}} \gamma_\varphi}{\sqrt{t}}} \quad ?$$

- **Answer:** Yes! And we just need $O(1)$ per subgradient computation
- Main idea: at each step use a single data point to approximate subgradient

$$\partial_w \varphi(w^\top X_i Y_i)$$

- This approach is motivated by the **stochastic oracle model**

Interplay between Optimization and Randomness

Stochastic Projected Subgradient Method

Goal: $\boxed{\min_{x \in \mathcal{C}} f(x)}$ with f convex, $\mathcal{C}$ convex

First Order Stochastic Oracle

Given X, the oracle yields back a random variable G that is an unbiased estimator of a subgradient of f at X conditionally on X, namely

$$\boxed{\mathbf{E}[G|X] \in \partial f(X)}$$

Projected Stochastic Subgradient Method

Projected Stochastic Subgradient Method

$$\boxed{\begin{aligned}\tilde{X}_{t+1} &= X_t - \eta_t G_t, \text{where } \mathbf{E}[G_t|X_t] \in \partial f(X_t) \\ X_{t+1} &= \Pi_{\mathcal{C}}(\tilde{X}_{t+1})\end{aligned}}$$

Projected Stochastic Subgradient Method (Theorem 11.1)

- Assume $\mathbf{E}[\|G_s\|_2^2] \le \gamma^2$ for any $s \in [t]$
- Assume $\mathbf{E}[\|X_1 - x^\star\|_2^2] \le b^2$

Then, projected subgradient method with $\eta_s \equiv \eta = \frac{b}{\gamma\sqrt{t}}$ satisfies

$$\boxed{\mathbf{E} f\left(\frac{1}{t}\sum_{s=1}^{t} X_s\right) - f(x^\star) \le \frac{\gamma b}{\sqrt{t}}}$$

Proof of Theorem 11.1

- By convexity and the properties of conditional expectations:

$$f(X_s)-f(x^\star) \le \partial f(X_s)^\top (X_s-x^\star) = \mathbf{E}[G_s|X_s]^\top (X_s-x^\star) = \mathbf{E}[G_s^\top (X_s-x^\star)|X_s]$$

- Proceeding as in the proof of Theorem 9.3:

$$G_s^\top (X_s - x^\star) \le \frac{1}{2\eta}(\|X_s - x^\star\|_2^2 - \|X_{s+1} - x^\star\|_2^2) + \frac{\eta}{2}\|G_s\|_2^2$$

- Taking the expectation, by the tower property of conditional expectations:

$$\begin{aligned}\mathbf{E}f(X_s) - f(x^\star) &\le \mathbf{E}\mathbf{E}[G_s^\top (X_s-x^\star)|X_s] = \mathbf{E}G_s^\top (X_s - x^\star)\\ &\le \frac{1}{2\eta}(\mathbf{E}\|X_s - x^\star\|_2^2 - \mathbf{E}\|X_{s+1} - x^\star\|_2^2) + \frac{\eta}{2}\mathbf{E}\|G_s\|_2^2\end{aligned}$$

  and using the assumption $\mathbf{E}\|G_s\|_2^2 \le \gamma^2$ we obtain

$$\frac{1}{t}\sum_{s=1}^{t}(\mathbf{E}f(X_s)-f(x^\star)) \le \frac{1}{2\eta t}\left(\mathbf{E}\|X_1-x^\star\|_2^2 - \mathbf{E}\|X_{t+1}-x^\star\|_2^2\right)+\frac{\eta}{2}\gamma^2 \le \frac{b^2}{2\eta t}+\frac{\eta\gamma^2}{2}$$

- Proof follows minimizing right-hand side ($\eta = \frac{b}{\gamma\sqrt{t}}$)

Back to Learning: Single and Multiple Passes O(1) Cost

- **Multiple Passes through the Data:**
  - **Goal:** Minimize **regularized** empirical risk R over $\mathcal{W}_2$
  - $G_s = \partial_w \varphi(W_s^\top X_{I_{s+1}} Y_{I_{s+1}})$ $\qquad$ ($I_2, I_3, I_4, \ldots$ are i.i.d. uniform in $[n]$)
  - $\mathbf{E}[\partial_w \varphi(W_s^\top X_{I_{s+1}} Y_{I_{s+1}}) | S, W_s] = \frac{1}{n} \sum_{i=1}^n \mathbf{E}[\partial \varphi(W_s^\top X_i Y_i) | S, W_s] = \partial R(W_s)$

$$\boxed{\mathbf{E}\,\texttt{Optimization} = \mathbf{E}[R(\overline{W}_t) - R(W^\star)] \leq \frac{2 c_2^{\mathcal{X}} c_2^{\mathcal{W}} \gamma_\varphi}{\sqrt{t}}}$$

- **Single Pass through the Data:**
  - **Goal:** Minimize **regularized** expected risk r over $\mathcal{W}_2$
  - $G_s = \partial_w \varphi(W_s^\top X_s Y_s)$
  - $\mathbf{E}[\partial_w \varphi(W_s^\top X_s Y_s) | W_s] = \partial r(W_s)$

$$\boxed{\mathbf{E}\, r(\overline{W}_t) - r(w^\star) \leq \frac{2 c_2^{\mathcal{X}} c_2^{\mathcal{W}} \gamma_\varphi}{\sqrt{t}}}$$

Direct bound on estimation error.
No need to go through empirical risk, Rademacher complexity, etc...

Projected Stochastic Mirror Descent

Projected Stochastic Mirror Descent

$$\boxed{\begin{aligned} \nabla\Phi(\widetilde{X}_{t+1}) &= \nabla\Phi(X_t) - \eta_t G_t, \text{where } \mathbf{E}[G_t|X_t] \in \partial f(X_t) \\ X_{t+1} &= \Pi_{\mathcal{C}}^{\Phi}(\widetilde{X}_{t+1}) \end{aligned}}$$

Projected Stochastic Mirror Descent (Theorem 11.2)

- Assume that $\mathbf{E}[\|G_s\|_*^2] \leq \gamma^2$ for any $s \in [t]$
- Mirror map Φ is α-strongly convex on $\mathcal{C} \cap \mathcal{D}$ w.r.t. the norm $\|\cdot\|$
- Initial condition is $X_1 \equiv x_1 \in \operatorname{argmin}_{x \in \mathcal{C} \cap \mathcal{D}} \Phi(x)$
- Assume $c^2 = \sup_{x \in \mathcal{C} \cap \mathcal{D}} \Phi(x) - \Phi(x_1)$

Then, projected mirror descent with $\eta_s \equiv \eta = \frac{c}{\gamma}\sqrt{\frac{2\alpha}{t}}$ satisfies

$$\boxed{\mathbf{E} f\left(\frac{1}{t}\sum_{s=1}^{t} X_s\right) - f(x^\star) \leq c\gamma\sqrt{\frac{2}{\alpha t}}}$$

Recap: Statistical and Computational optimality

Linear models $f(x,a) = \langle a, x \rangle$ with Lipschitz loss function ℓ

- **Ridge regression:**

$\mathcal{A}_\rho = \{w^\top x : \|w\|_2 \le \rho\}$

Statistics $\lesssim \rho\sqrt{\frac{d}{n}}$

Computation $\lesssim \rho\sqrt{\frac{d}{t}}$ (proj. gradient descent)

- **Lasso:**

$\mathcal{A}_\rho = \{w^\top x : \|w\|_1 \le \rho\}$

Statistics $\lesssim \rho\sqrt{\frac{\log d}{n}}$

Computation $\lesssim \rho\sqrt{\frac{\log d}{t}}$ (proj. mirror descent) (entropy mirror map)

We need $t \sim n$ iterations, i.e. computational complexity $O(n^2 d)$
(Stochastic gradient descent yields optimal computational complexity $O(nd)$)

Limitations leading to implicit regularization...

Explicit regularization and uniform convergence:

$$r(A) - r(a^\star) \leq \underbrace{2 \sup_{a \in \mathcal{A}} |r(a) - R(a)|}_{\text{Statistics}} + \underbrace{R(A) - R(A^\star_{\mathcal{A}})}_{\text{Computation}} + \underbrace{r(a^\star_{\mathcal{A}}) - r(a^\star)}_{\text{Approximation}}$$

Statistics:

- If the empirical risk R has multiple global minima, it can be $r(A^\star) \ll r(A^{\star\prime})$ but the bound above does not differentiate

Computation:

- If the empirical risk R is non-convex, it is typically not feasible to make $R(A) - R(A^\star)$ arbitrarily small

Approximation:

- In practice, optimal choices of the class $\mathcal{A}$ involve unknown quantities, e.g. level of the noise, so one has to resort to model selection (expensive)

Limitations prompt to study **implicit** regularization of solvers applied in practice

Can Avoid Supremum and Directly Bound Excess Risk?

Recall from Lecture 1:

$$\underbrace{r(A) - r(a^{\star\star})}_{\text{excess risk}} = \underbrace{r(A) - r(a^{\star})}_{\text{estimation error}} + \underbrace{r(a^{\star}) - r(a^{\star\star})}_{\text{approximation error}}$$

So far we used the following decomposition (apart from proof of Theorem 7.10...):

$$\begin{aligned}\underbrace{r(A)-r(a^{\star})}_{\text{estimation error}} &= r(A)-R(A) + \underbrace{R(A)-R(A^{\star})}_{\text{optimization error}} + \underbrace{R(A^{\star})-R(a^{\star})}_{\leq 0} + R(a^{\star})-r(a^{\star})\\ &\leq \underbrace{R(A)-R(A^{\star})}_{\text{optimization error}} + \underbrace{\sup_{a\in\mathcal{A}}(r(a)-R(a)) + \sup_{a\in\mathcal{A}}(R(a)-r(a))}_{\text{statistics error}}\end{aligned}$$

Question. Can we analyze directly excess risk without explicit regularization (i.e., without admissible set $\mathcal{A} \subseteq \mathcal{B}$)?

Question. Can we analyze directly behavior of A without taking the supremum (i.e., without notions of complexity for set $\mathcal{A} \subseteq \mathcal{B}$)?

Answer. Yes to both! Use algorithmic stability and implicit regularization

Algorithmic Stability: New Error Decomposition

New error decomposition (Proposition 11.3)

For any $A \in \mathcal{B}$ we have

$$\boxed{\mathbf{E}\,\underbrace{r(A) - r(a^{\star\star})}_{\text{excess risk}} \leq \mathbf{E}\,\underbrace{[r(A) - R(A)]}_{\text{generalization error}} + \mathbf{E}\,\underbrace{[R(A) - R(A^{\star\star})]}_{\text{optimization error}}}$$

Proof. We have

$$r(A) - r(a^{\star\star}) = r(A) - R(A) + R(A) - R(A^{\star\star}) + R(A^{\star\star}) - r(a^{\star\star}).$$

Note that $\mathbf{E}R(A^{\star\star}) \leq r(a^{\star\star})$, as for any $a \in \mathcal{B}$ we have $R(A^{\star\star}) \leq R(a)$ (as, by definition, $A^{\star\star}$ is a minimizer of the empirical risk R over $\mathcal{B}$) so that

$$\mathbf{E}R(A^{\star\star}) \leq \mathbf{E}R(a) = r(a),$$

which holds also for $a = a^{\star\star}$.

Algorithmic Stability

Let $\widetilde{A}(i)$ be algorithm trained on perturbed dataset $\{Z_1, ..., Z_{i-1}, \widetilde{Z}_i, Z_{i+1}, ..., Z_n\}$

Generalization error bound via algorithmic stability (Proposition 11.5)

If for any $z \in \mathcal{Z}$ the function $a \to \ell(a, z)$ is γ-Lipschitz, then

$$\boxed{\mathbf{E}[\,\underbrace{r(A) - R(A)}_{\text{generalization error}}\,] \leq \gamma\,\frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\|A - \widetilde{A}(i)\|}$$

Stability: $\|A - \widetilde{A}(i)\|$ small.

Proof. We have $\mathbf{E}\,r(A) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\,\ell(A, \widetilde{Z}_i)$.
As (A, Z_i) has the same distribution as $(\widetilde{A}(i), \widetilde{Z}_i)$:

$$\mathbf{E}R(A) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\,\ell(A, Z_i) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\,\ell(\widetilde{A}(i), \widetilde{Z}_i)$$

Stability for Stochastic Gradient Descent. Early Stopping

Take $A = W_t$, stochastic gradient descent **(no projection as no constraints!)**

Generalisation error for convex Lipschitz and smooth losses (Lemma 11.6)

- Function $w \in \mathbb{R}^d \to \ell(w, z)$ is convex, γ-Lipschitz and β-smooth
- $\eta_s \equiv \eta$ satisfying $\eta\beta \leq 2$
- Let $W_1 = 0$

$$\boxed{\mathbf{E}[\underbrace{r(W_t) - R(W_t)}_{\text{generalization error}}] \leq \frac{2\eta\gamma^2}{n}(t-1)}$$

Early stopping: find time that minimizes upper bounds using Proposition 11.3:

- Generalization error: increasing with time
- Optimization error: decreasing with time

Example of implicit/algorithmic regularization, as opposed to explicit/structural