Algorithmic Foundations of Learning

Lecture 11
Stochastic Oracle Model
Algorithmic Stability and Implicit Regularization

Patrick Rebeschini

Department of Statistics
University of Oxford
November 13th 2018
Recall: Subgradient Descent with Euclidean Geometry

Risk minimization:
\[
\begin{align*}
\text{minimize} & \quad r(w) = \mathbf{E} \varphi(w^\top X Y) \\
\text{subject to} & \quad \|w\|_2 \leq c_2^W
\end{align*}
\]
\[\implies \text{Let } w^* \text{ be a minimizer}\]

Empirical risk minimization:
\[
\begin{align*}
\text{minimize} & \quad R(w) = \frac{1}{n} \sum_{i=1}^{n} \varphi(w^\top X_i Y_i) \\
\text{subject to} & \quad \|w\|_2 \leq c_2^W
\end{align*}
\]
\[\implies \text{Let } W^* \text{ be a minimizer}\]

\[
r(\overline{W}_t) - r(w^*) \leq R(\overline{W}_t) - R(W^*) + \sup_{w \in \mathcal{W}} \{r(w) - R(w)\} + \sup_{w \in \mathcal{W}} \{R(w) - r(w)\}
\]

\[
\mathbf{E} \text{Statistics} \leq \frac{4 c_2^X c_2^W \gamma \varphi}{\sqrt{n}} \quad \text{Optimization} \leq \frac{2 c_2^X c_2^W \gamma \varphi}{\sqrt{t}}
\]

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

$$\mathbb{E} \text{Optimization} \leq \frac{2c_2^X c_2^W \gamma \varphi}{\sqrt{t}}$$

▶ 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 Descent

**Goal:** \( \min_{x \in C} f(x) \) with \( f \) convex, \( C \) 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

\[
E[G | X] \in \partial f(X)
\]
Projected Stochastic Subgradient Descent

\[ \tilde{X}_{t+1} = X_t - \eta_t G_t, \text{ where } \mathbb{E}[G_t | X_t] \in \partial f(X_t) \]
\[ X_{t+1} = \Pi_C(\tilde{X}_{t+1}) \]

Projected Stochastic Subgradient Descent (Theorem 11.1)

- Assume \( \mathbb{E}[\|G_s\|^2_2] \leq \gamma^2 \) for any \( s \in [t] \)
- Assume \( \mathbb{E}[\|X_1 - x^*\|^2_2] \leq b^2 \)

Then, projected subgradient descent with \( \eta_s \equiv \eta = \frac{b}{\gamma\sqrt{t}} \) satisfies

\[ \mathbb{E} f \left( \frac{1}{t} \sum_{s=1}^{t} X_s \right) - f(x^*) \leq \frac{\gamma b}{\sqrt{t}} \]
Back to Learning: Single and Multiple Passes O(1) Cost

▶ Multiple Passes through the Data:

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

\[
\mathbb{E} \text{ Optimization} = \mathbb{E}[R(\overline{W}_t) - R(W^*)] \leq \frac{2c_s^x c_2^\mathcal{W} \gamma \varphi}{\sqrt{t}}
\]

▶ Single Pass through the Data:

- **Goal:** Minimize regularized expected risk $r$ over $\mathcal{W}_2$
- $G_s = \partial_w \varphi(W_s^T X_s Y_s)$
- $\mathbb{E}[\partial_w \varphi(W_s^T X_s Y_s)|W_s] = \partial r(W_s)$

\[
\mathbb{E} r(\overline{W}_t) - r(w^*) \leq \frac{2c_s^x c_2^\mathcal{W} \gamma \varphi}{\sqrt{t}}
\]

Direct bound on estimation error. No need to go through empirical risk!
Projected Stochastic Mirror Descent

\[ \nabla \Phi(\tilde{X}_{t+1}) = \nabla \Phi(X_t) - \eta_t G_t, \text{where } E[G_t | X_t] \in \partial f(X_t) \]
\[ X_{t+1} = \Pi_\Phi^C(\tilde{X}_{t+1}) \]

Projected Stochastic Mirror Descent (Theorem 11.2)

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

Then, projected mirror descent with \( \eta_s \equiv \eta = \frac{c}{\gamma} \sqrt{\frac{2\alpha}{t}} \) satisfies
\[ Ef \left( \frac{1}{t} \sum_{s=1}^{t} X_s \right) - f(x^*) \leq c\gamma \sqrt{\frac{2}{\alpha t}} \]
Can Avoid Supremum and Directly Bound Excess Risk?

Recall from Lecture 1:

\[
\begin{align*}
\underbrace{r(A) - r(a^{**})}_{\text{excess risk}} &= \underbrace{r(A) - r(a^*)}_{\text{estimation error}} + \underbrace{r(a^*) - r(a^{**})}_{\text{approximation error}}
\end{align*}
\]

So far we used the following decomposition (apart from proof of Theorem 7.9…):

\[
\begin{align*}
\underbrace{r(A) - r(a^*)}_{\text{estimation error}} &= \underbrace{r(A) - R(A)}_{\text{optimization error}} + \underbrace{R(A) - R(A^*)}_{\text{optimization error}} + \underbrace{R(A^*) - R(a^*)}_{\text{optimization error}} + \underbrace{R(a^*) - r(a^*)}_{\text{statistics error}} \\
&\leq \underbrace{R(A) - R(A^*)}_{\text{optimization error}} + \sup_{a \in A} \left( r(a) - R(a) \right) + \sup_{a \in A} \left( R(a) - r(a) \right)
\end{align*}
\]

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

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

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

New error decomposition (Proposition 11.3)

For any $A \in B$ we have

$$
\mathbb{E}[r(A) - r(a^{**})] \leq \mathbb{E}[r(A) - R(A)] + \mathbb{E}[R(A) - R(A^{**})]
$$

excess risk  
generalization error  
optimization error

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

Generalization error bound via algorithmic stability (Proposition 11.4)

If for any $z \in Z$ the function $a \rightarrow \ell(a, z)$ is $\gamma$-Lipschitz, then

$$
\mathbb{E}[r(A) - R(A)] \leq \gamma \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\|A - \tilde{A}(i)\|
$$

generalization error

Stability: $\|A - \tilde{A}(i)\|$ is small
Stability for Stochastic Gradient Descent. Early Stopping

Take $A = W_t$, stochastic gradient descent \textit{(no projection as no constraints!)}

Generalisation error for convex Lipschitz and smooth losses (Lemma 11.5)

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

\[
\mathbb{E}[r(W_t) - R(W_t)] \leq \frac{2\eta \gamma^2}{n} (t - 1)
\]

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/algorithic regularization, as opposed to explicit/structural