Generalization Properties and Implicit Regularization for Multiple Passes SGM
by
Lin J., Camoriano R., Rosasco L.

ADEIKALAM Pierre
CHEN Guangyue
XU Kevin

The paper

Aim of the paper

Analyze how the step-size (a.k.a learning rate) and the number of passes in the stochastic gradient method (SGM) induce an implicit regularization of the model. This is done by :

- Finding explicit bounds on the generalization risk that depend on the step-size and the number of passes.
- Exploiting different strategies for setting the step-sizes and number of passes to optimize this bound, thus showing the regularisation effect of these parameters.

Our aim in this short presentation

- Clearly and easily explain the regularization effect of these parameters to people that have not read the paper.

Contenu

Introduction

- Regression function : $f_w(x) = \langle w, \phi(x) \rangle$ where $\phi(x) = K(x, \cdot)$ is a positive definite kernel.
- Loss function : $V(y, \cdot)$ left differentiable.
- SGM Algorithm :

$$w_{t+1} = w_t - \eta_t V'(y_{j_t}, \langle w_t, \phi(x_{j_t}) \rangle \phi(x_{j_t})$$

where

- j_t is a sample from $(x_1, ..., x_m)$
- $(\eta_t)_{t \geq 1}$ is a non-increasing sequence of step-sizes.

Introduction

- Expected excess risk of the last iterate :

$$\mathbb{E}[\varepsilon(w_T) - \inf_{w} \varepsilon(w)]$$

where

$$\varepsilon(w) = \frac{1}{m} \sum_{j=1}^{m} V(y_j, f(x_j))$$

Assumption 1

- $\forall y \in Y$, $x \in X \mapsto V(y, x)$ is convex.
- The loss V is bounded (True if Y and X are bounded).
- The derivative of the loss V' is bounded (True if V is Lipschitz and the above holds).
- $K = \sup_{x \in X} \|\phi(x)\|_2 < \infty$ (True if ϕ is continuous and all of the above hold).

The constant K will be very important later as it will set the initial step-size for every strategy.

Assumption 2

We define the approximation error of (V, ϕ) as :

$$D(\lambda) = \inf_w \left\{ \varepsilon(w) + \frac{\lambda}{2} \|w\|^2 \right\} - \inf_w \varepsilon(w)$$

and we assume $\exists \beta \in]0, 1]$ such that $D(\lambda) \leq c_\beta \lambda^\beta$ for some $c_\beta > 0$.

Assumption 3

We assume that $\forall y \in Y$, $V(y, \cdot)$ is differentiable and $x \mapsto V'(y, x)$ is L-Lipschitz for some $L > 0$.

Main Theorem

If Assumptions 1, 2 and 3 hold and for all t, we set $\eta_t \leq \frac{2}{K^2 L}$, then

$$\mathbb{E}[\varepsilon(w_t) - \inf_w \varepsilon(w)] \lesssim \frac{\sum_{k=1}^t \eta_k}{m} \sum_{k=1}^{t-1} \frac{\eta_k}{\eta_t (t-k)} \tag{1}$$

$$+ \sum_{k=1}^{t-1} \frac{\eta_k^2}{\eta_t (t-k)} + \eta_t \tag{2}$$

$$+ \frac{\left(\sum_{k=1}^t \eta_k\right)^{1-\beta}}{\eta_t t} \tag{3}$$

This implies that the 3 error terms can be balanced to find optimal choices for the number of steps t and step-sizes $(\eta_k)_{k=1}^t$, thus proving the existence of their regularisation effect.

Strategies

We will now look at 4 different strategies to minimize this upper bound on the generalization error.

- The first 2 strategies consist in defining step-sizes a priori (with no knowledge of β) and fine-tune the number of iterations t.
- The other 2 strategies consist in reaching the optimum in one pass by fine-tuning the step-sizes instead.

Strategy 1 : Constant Step-Sizes

If Assumptions 1, 2 and 3 hold, we set $\eta_t = \frac{\eta_1}{\sqrt{m}}$ for some initial step-size $0 < \eta_1 \leq \frac{2}{K^2 L}$.

Then, there exists an optimal number of iterations $t^* = \lceil m^{\frac{\beta+3}{2(\beta+1)}} \rceil$ such that

$$\mathbb{E}[\varepsilon(w_{t^*}) - \inf_w \varepsilon(w)] \leq m^{-\frac{\beta}{\beta+1}} \log(m)$$

which is the optimal bound.

Strategy 2 : Decaying Step-Sizes

If Assumptions 1, 2 and 3 hold, we set $\eta_t = \frac{\eta_1}{\sqrt{t}}$ for some initial step-size $0 < \eta_1 \leq \frac{2}{K^2 L}$.

Then, there exists an optimal number of iterations $t^* = \lceil m^{\frac{2}{\beta+1}} \rceil$ such that

$$\mathbb{E}[\varepsilon(w_{t^*}) - \inf_w \varepsilon(w)] \leq m^{-\frac{\beta}{\beta+1}} \log(m)$$

which is the optimal bound.

Strategy 3 : One pass Constant Step-Sizes

If Assumptions 1, 2 and 3 hold, we set $t^* = m$. Then, the optimal bound is reached by using the step-sizes given by :

$$\eta_t = \eta_1 m^{-\frac{\beta}{\beta+1}}$$

for some initial step-size $0 < \eta_1 \leq \frac{2}{K^2 L}$.

Strategy 4 : One pass Decaying Step-Sizes

If Assumptions 1, 2 and 3 hold, we set $t^* = m$. Then, the optimal bound is reached by using the step-sizes given by :

$$\eta_t = \eta_1 t^{-\frac{\beta}{\beta+1}}$$

for some initial step-size $0 < \eta_1 \leq \frac{2}{K^2 L}$.

Simulations

The authors did some numerical simulations for showing different regularization effects of the step-size(fixed or decaying) and the number of passes in SGM and SIGM.

- Test error with respect to the number of passes.
- Test error cross-validation.

The number of passes

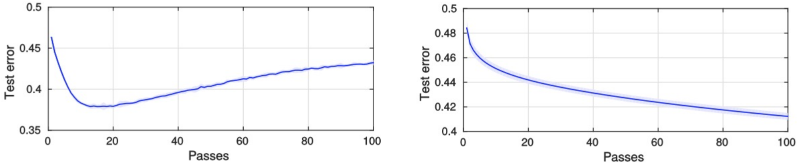

Figure 1 – Test error for SIGM with fixed (a) and decaying (b) step-size with respect to the number of passes on Adult (n = 1000)

- For fixed step-size, it has overfitting regime. Which clearly illustrates the regularization effect of the number of passes.
- For decaying step-size, overfitting is not observed in the first 100 passes, the convergence to the optimal solution slower than fixed case.

Cross-validation

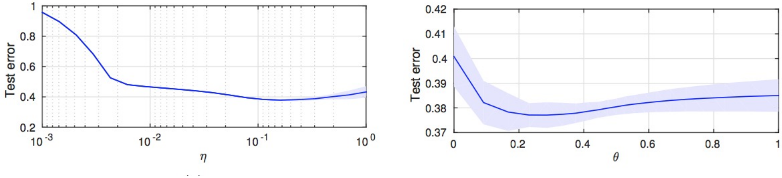

Figure 2 – Test error for SGM with fixed (a) and decaying (b) step-size cross-validation on Adult (n = 1000).

- For decaying step-size, the author fixed $\eta_1 = \frac{1}{4}$, and it shows that the decay rate has a regularization effect.
- For fixed step-size, a large step-size ($\eta = 1$) leads to overfitting, while a smaller one (10^{-3}) is associated to oversmoothing.

Conclusion

- Both the step-sizes and the number of passes have a regularisation effect.
- Each effect needs to be balanced to achieve optimal generalization.
- Different strategies are available. By setting the step-sizes a priori we can use Early Stopping to find the optimal number of passes.

Conclusion

Thank you !