Distributed Optimization and Learning

Solution Manual

STEFAN VLASKI and ALI H. SAYED

Imperial College London, United Kingdom
École Polytechnique Fédérale de Lausanne, Switzerland

1 Inference, Optimization and Learning

1.1 In this problem we implement the least-squares fusion technique in (1.60) in code to verify empirically verify the benefit of data fusion. You are free to choose any unspecified parameters.

(a) Generate local data sets $\{h_{k,n}, \gamma_{k,n}\}_{n=1}^{N}$ following the statistical model of Example 1.4.

(b) Compute $w_k^\star$ for each agent along with the globally optimal model $w^\star$.

(c) Generate a new independent test set $\left\{\widetilde{h}_n, \widetilde{\gamma}_n\right\}_{n=1}^{\widetilde{N}}$ and evaluate prediction performance of local models:

$$\frac{1}{\widetilde{N}}\sum_{n=1}^{\widetilde{N}}(\widetilde{\gamma}_n - \widetilde{h}_n^\mathsf{T} w_k^\star)^2$$

as well as of the performance of the global model:

$$\frac{1}{\widetilde{N}}\sum_{n=1}^{\widetilde{N}}(\widetilde{\gamma}_n - \widetilde{h}_n^\mathsf{T} w^\star)^2$$

Compare the performance for varying choices of N and K. What do you conclude?

Solution. The solution is provided as a Jupyter notebook in the separate file `Problem_1_1.ipynb`.

1.2 We consider again the setting of Section 1.3, but generalize to the setting where local data sets $\{h_{k,n}, \gamma_{k,n}\}_{n=1}^{N_k}$ have differing sizes N_k. Adjust the derivation leading up to (1.60) as necessary to find an optimal fusion protocol in this generalized setting. How do the local data sizes N_k affect the fusion?

Solution. We begin with (1.44), which, for a sample of N_k data points will take the form:

$$w_k^\star \triangleq \arg\max_{w\in\Omega} f_{\boldsymbol{w}|\{\boldsymbol{h}_{k,n},\boldsymbol{\gamma}_{k,n}\}_{n=1}^{N_k}}\left(w|\,\{h_{k,n},\gamma_{k,n}\}_{n=1}^{N_k}\right)$$

For the global model $w^\star$, we can define:

$$w^\star \triangleq \arg\max_{w \in \Omega} f_{\boldsymbol{w}|\left\{\{\boldsymbol{h}_{k,n}, \boldsymbol{\gamma}_{k,n}\}_{n=1}^{N_k}\right\}_{k=1}^{K}}\left(w| \left\{\{h_{k,n}, \gamma_{k,n}\}_{n=1}^{N_k}\right\}_{k=1}^{K}\right)$$

where we are aggregating all N_k samples from each agent k. We note that there is no structural change in the expression for the optimal estimate. To make the dependence on the sample size explicit, we specialize to the linear model leading to least-squares fusion. For the local models, it is straightforward to adjust the argument leading to (1.53) to allow for N_k samples, leading to:

$$\begin{aligned} w_k^\star &= \left(\sum_{n=1}^{N_k} h_{k,n}h_{k,n}^{\mathsf{T}} + \rho I\right)^{-1}\left(\sum_{n=1}^{N_k} \gamma_{k,n}h_{k,n}\right) \\ &= \left(H_k + \rho I\right)^{-1} d_k \end{aligned}$$

where we defined for compactness:

$$\begin{aligned} H_k &= \sum_{n=1}^{N_k} h_{k,n}h_{k,n}^{\mathsf{T}} \\ d_k &= \sum_{n=1}^{N_k} \gamma_{k,n}h_{k,n} \end{aligned}$$

For the global model, we find:

$$\begin{aligned} w^\star &= \left(\sum_{k=1}^{K}\sum_{n=1}^{N_k} h_{k,n}h_{k,n}^{\mathsf{T}} + \rho I\right)^{-1}\left(\sum_{k=1}^{K}\sum_{n=1}^{N_k} \gamma_{k,n}h_{k,n}\right) \\ &= \left(\sum_{k=1}^{K} H_k + \rho I\right)^{-1}\left(\sum_{k=1}^{K} d_k\right) \end{aligned}$$

and as the analogue of (1.60), we find:

$$w^\star = \left(\sum_{k=1}^{K} H_k + \rho I\right)^{-1}\left(\sum_{k=1}^{K} (H_k + \rho I)w_k^\star\right)$$

It is important to note that while N_k does not explicitly appear, it does implicitly affect the fusion, as H_k and d_k both grow with N_k. To make this dependence more explicit, we can normalize as follows. Define the total number of samples

$N = \sum_{k=1}^{K} N_k$. Then:

$$\begin{aligned}
w^\star &= \left(\sum_{k=1}^{K} H_k + \rho I\right)^{-1} \left(\sum_{k=1}^{K} (H_k + \rho I) w_k^\star\right) \\
&= \left(\frac{1}{N}\sum_{k=1}^{K} H_k + \frac{\rho}{N} I\right)^{-1} \left(\frac{1}{N}\sum_{k=1}^{K} (H_k + \rho I) w_k^\star\right) \\
&= \left(\sum_{k=1}^{K} \frac{N_k}{N}\frac{1}{N_k} H_k + \frac{\rho}{N} I\right)^{-1} \left(\sum_{k=1}^{K} \frac{N_k}{N}\left(\frac{1}{N_k} H_k + \frac{\rho}{N_k} I\right) w_k^\star\right) \\
&= \left(\sum_{k=1}^{K} \frac{N_k}{N}\overline{H}_k + \frac{\rho}{N} I\right)^{-1} \left(\sum_{k=1}^{K} \frac{N_k}{N}\left(\overline{H}_k + \frac{\rho}{N_k} I\right) w_k^\star\right)
\end{aligned}$$

where we introduced the sample covariances:

$$\overline{H}_k = \frac{1}{N_k} H_k = \frac{1}{N_k}\sum_{n=1}^{N_k} h_{k,n} h_{k,n}^{\mathsf{T}}$$

In contrast to H_k, the sample covariance $\overline{H}_k$ is normalized by the sample size N_k, and stable as N_k grows. In turn, the effect of $\overline{H}_k$ and $w_k^\star$ in (3.1) is mordulate by $\frac{N_k}{N}$, which corresponds to the fraction of total data samples available at agent k.

2 Deterministic Optimization

2.1 Find the smoothness and strong-convexity parameters δ and ν of the logistic regression cost (2.27a). Use these and Theorem 2.1 to determine the rate of convergence of gradient descent when applied to problem (2.27a). Which choice of the step-size parameter μ gives the fastest rate of convergence?

Solution. The most straightforward way for finding δ, ν is through the conditions on the Hessian matrix. To this end, we find:

$$\nabla_w^2 \, J(w) = \qquad \rho I_M \; + \; \frac{1}{N}\sum_{n=1}^{N} h_n h_n^{\mathsf{T}} \, \frac{e^{-\gamma_n h_n^{\mathsf{T}} w}}{\left(1+e^{-\gamma_n h_n^{\mathsf{T}} w}\right)^2}$$

from which we readily conclude that

$$0 \; < \; \underbrace{\rho}_{\triangleq\, \nu} \, I_M \; \leq \; \nabla_w^2 \, J(w) \; \leq \; \underbrace{\rho I_M + \lambda_{\max}\left(\frac{1}{N}\sum_{n=1}^{N} h_n h_n^{\mathsf{T}}\right) I_M}_{\triangleq\, \delta}$$

For the rate of convergence, we then find from (2.32) that:

$$\lambda = 1 - 2\mu\nu + \mu^2\delta^2 = 1 - 2\mu\rho + \mu^2\left(\rho I_M + \lambda_{\max}\left(\frac{1}{N}\sum_{m=1}^{N} h_n h_n^{\mathsf{T}}\right)\right)^2$$

From Figure 2.4, we know that the fastest rate of convergence is obtained for:

$$\lambda^o = \frac{\nu}{\delta^2} = \frac{\rho}{\left(\rho I_M + \lambda_{\max}\left(\frac{1}{N}\sum_{m=1}^{N} h_n h_n^{\mathsf{T}}\right)\right)^2}$$

3 Stochastic Optimization

3.1 We are interested in solving the logistic regression problem (3.38) by employing the ordinary gradient approximation $\widehat{\nabla J}(w) \triangleq \nabla Q(w; \boldsymbol{h}_i, \boldsymbol{\gamma}_i)$, giving rise to the recursion:

$$\boldsymbol{w}_i = (1 - \mu\rho)\boldsymbol{w}_{i-1} + \mu \frac{\boldsymbol{\gamma}_i \boldsymbol{h}_i}{1 + e^{\boldsymbol{\gamma}_i \boldsymbol{h}_i^{\mathsf{T}} \boldsymbol{w}_{i-1}}}$$

Determine the strong convexity parameter ν of the logistic cost (3.38). Use it along with the gradient noise parameters derived in Example 3.4 and Theorem 3.1 to find a bound for the steady-state performance $\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2$. How do you need to choose the step-size in terms of the other parameters to ensure that

$$\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \le \epsilon?$$

How will a very small ϵ affect the rate of convergence λ?

Solution. Follow an analogous argument to that of Problem 2.1, we find that for the expected logistic regression problem (3.38) $\nu = \rho$. From (3.40) we know that $\sigma^2 = \mathrm{Tr}(R_h)$. It then follows from (3.54) that:

$$\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \le \frac{\mu\sigma^2}{\nu} = \frac{\mu \mathrm{Tr}(R_h)}{\rho}$$

To ensure $\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \le \epsilon$, we can rearrange and solve from μ to find:

$$\mu \le \frac{\epsilon\rho}{\mathrm{Tr}(R_h)}$$

In other words, to achieve a small limiting error, we need to choose a small step-size μ. However, this choice implies for the rate of convergence from (3.62) that:

$$\lambda \ge 1 - 2\frac{\epsilon\rho^2}{\mathrm{Tr}(R_h)}$$

Small values of ϵ result in small values of μ, which in turn results in λ close to one, and hence slow convergence.

3.2 We consider again the logistic regression problem (3.38), but now employ the mini-batch stochastic gradient approximation (3.41), giving rise to:

$$\boldsymbol{w}_i = (1 - \mu\rho)\boldsymbol{w}_{i-1} + \frac{\mu}{B}\sum_{b=1}^{B} \frac{\boldsymbol{\gamma}_{b,i}\boldsymbol{h}_{b,i}}{1 + e^{\boldsymbol{\gamma}_{b,i}\boldsymbol{h}_{b,i}^{\mathsf{T}}\boldsymbol{w}_{i-1}}}$$

Determine an expression for the gradient noise parameters of this improved approximation, and use them with Theorem 3.1 to find a bound for the steady-state performance $\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2$. Keeping the step-size constant, how do you need to choose the mini-batch size B as a function of the remaining parameters to ensure:

$$\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \le \epsilon?$$

How will a very small ϵ affect the rate of convergence λ?

Solution. From (3.46), we can conclude that the mini-batch approximation of the gradient of the logistic cost will exhibit an absolute variance component $\frac{\mathrm{Tr}(R_h)}{B}$ with B-fold reduction in the variance. Following the same argument as in Problem 3.1, we then find:

$$\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \le \frac{\mu\sigma^2}{\nu} = \frac{\mu \mathrm{Tr}(R_h)}{B\rho} \tag{3.1}$$

Solving $\lim_{i\to\infty} \mathbb{E}\|\widetilde{\boldsymbol{w}}_i\|^2 \le \epsilon$ for B we then find:

$$B \ge \frac{\mu \mathrm{Tr}(R_h)}{\epsilon\rho} \tag{3.2}$$

We conclude that high accuracy requires small values of ϵ, which in turn necessitates large mini-batch sizes B. In contrast to Problem 3.1, however, we find for the convergence rate:

$$\lambda = 1 - 2\mu\nu + O(\mu^2) \approx 1 - 2\mu\nu \tag{3.3}$$

which to first order in the step-size is unaffected by ϵ since we are keeping the step-size constant and controlling accuracy through the mini-batch size B. This is of course desirable, but comes at the increased computational burden of having to compute B gradients at every iteration.

3.3 Consider a gradient approximation similar to the mini-batch construction (3.41), where the number of available samples at time i is now a random variable $\boldsymbol{B}_i$, distributed uniformly between 1 and B. We can then construct:

$$\widehat{\nabla J}(w) \triangleq \frac{1}{\boldsymbol{B}_i}\sum_{b=1}^{\boldsymbol{B}_i} \nabla Q(w; \boldsymbol{x}_{b,i})$$

Verify that this construction satisfies the unbiasedness condition (3.22) and determine an expression for the gradient noise constants of (3.23) in terms of those of the ordinary gradient approximation.

Solution. Let us first verify the zero-mean condition (3.22). We have:

$$
\begin{aligned}
\mathbb{E}\widehat{\nabla J}(w) &\overset{(a)}{=} \sum_{\overline{B}=1}^{B} \Pr\left\{\boldsymbol{B}_i = \overline{B}\right\} \cdot \mathbb{E}\left\{\widehat{\nabla J}(w) | \boldsymbol{B}_i = \overline{B}\right\} \\
&= \frac{1}{B} \sum_{\overline{B}=1}^{B} \mathbb{E}\left\{ \frac{1}{\overline{B}} \sum_{b=1}^{\overline{B}} \nabla Q(w; \boldsymbol{x}_{b,i}) | \boldsymbol{B}_i = \overline{B} \right\} \\
&\overset{(b)}{=} \frac{1}{B} \sum_{\overline{B}=1}^{B} \mathbb{E}\left\{ \frac{1}{\overline{B}} \sum_{b=1}^{\overline{B}} \nabla Q(w; \boldsymbol{x}_{b,i}) \right\} \\
&= \frac{1}{B} \sum_{\overline{B}=1}^{B} \frac{1}{\overline{B}} \sum_{b=1}^{\overline{B}} \mathbb{E}\left\{\nabla Q(w; \boldsymbol{x}_{b,i})\right\} \\
&= \frac{1}{B} \sum_{\overline{B}=1}^{B} \nabla J(w) \\
&= \nabla J(w)
\end{aligned} \tag{3.4}
$$

Here, (a) follows from the law of total probability and (b) follows by indpendence of $\boldsymbol{B}_i$ and the data $\boldsymbol{x}_{b,i}$. For the variance, we have:

$$
\begin{aligned}
&\mathbb{E}\|\widehat{\nabla J}(w) - \nabla J(w)\|^2 \\
&\overset{(a)}{=} \sum_{\overline{B}=1}^{B} \Pr\left\{\boldsymbol{B}_i = \overline{B}\right\} \cdot \mathbb{E}\left\{ \left\|\widehat{\nabla J}(w) - \nabla J(w)\right\|^2 | \boldsymbol{B}_i = \overline{B} \right\} \\
&= \frac{1}{B} \sum_{\overline{B}=1}^{B} \mathbb{E}\left\{ \left\| \frac{1}{\overline{B}} \sum_{b=1}^{\overline{B}} \nabla Q(w; \boldsymbol{x}_{b,i}) - \nabla J(w) \right\|^2 | \boldsymbol{B}_i = \overline{B} \right\} \\
&= \frac{1}{B} \sum_{\overline{B}=1}^{B} \mathbb{E}\left\{ \left\| \frac{1}{\overline{B}} \sum_{b=1}^{\overline{B}} \nabla Q(w; \boldsymbol{x}_{b,i}) - \nabla J(w) \right\|^2 \right\} \\
&\overset{(3.45)}{=} \frac{1}{B} \sum_{\overline{B}=1}^{B} \left(\frac{\alpha_{\text{ord}}^2}{\overline{B}} \|\nabla J(\boldsymbol{w}_{i-1})\|^2 + \frac{\beta_{\text{ord}}^2}{\overline{B}} \|w^o - \boldsymbol{w}_{i-1}\|^2 + \frac{\gamma_{\text{ord}}^2}{\overline{B}} \left(J(\boldsymbol{w}_{i-1}) - J^o\right) + \frac{\sigma_{\text{ord}}^2}{\overline{B}} \right) \\
&= \frac{1}{B} \left(\sum_{\overline{B}=1}^{B} \frac{1}{\overline{B}} \right) \left(\alpha_{\text{ord}}^2 \|\nabla J(\boldsymbol{w}_{i-1})\|^2 + \beta_{\text{ord}}^2 \|w^o - \boldsymbol{w}_{i-1}\|^2 + \gamma_{\text{ord}}^2 \left(J(\boldsymbol{w}_{i-1}) - J^o\right) + \sigma_{\text{ord}}^2 \right)
\end{aligned} \tag{3.5}
$$

We conclude that all gradient noise constants get multiplied by a factor of $\frac{1}{B}\left(\sum_{\overline{B}=1}^{B}\right) < 1$.