Constant Step Size Stochastic Gradient Descent for Probabilistic Modeling

Supplementary material

In this supplementary material we provide explicit expressions for the asymptotic expansions from the main paper. All assumptions from [2] are reused, namely, beyond the usual sampling assumptions, smoothness of the cost functions.

1 Explicit form of $B, \bar{B}, \bar{\bar{B}}^w$ and $\bar{\bar{B}}^m$

We have, even for mis-specified models:

$$\mathcal{F}(\eta) - \mathcal{F}(\eta_*) = \mathcal{G}(\mu) - \mathcal{G}(\mu_*) =$$

$$= \mathbb{E}\big[- \mathbb{E}_{p(y_n|x_n)} y_n \eta(x_n) + a(\eta(x_n)) + \mathbb{E}_{p(y_n|x_n)} y_n \eta_*(x_n) - a(\eta_*(x_n))\big] =$$

$$= \mathbb{E}\Big[a(\eta(x_n)) - a(\eta_*(x_n)) - a'(\eta_*(x_n))(\eta(x_n) - \eta_*(x_n))\Big] + \mathbb{E}\Big[\big(a'(\eta_*(x_n)) - \mathbb{E}_{p(y_n|x_n)} y_n\big) \cdot \big(\eta(x_n) - \eta_*(x_n)\big)\Big] =$$

$$\mathbb{E}\Big[D_a\big(\eta(x_n)\big|\eta_*(x_n)\big)\Big] + \mathbb{E}\Big[\big(\mu_*(x) - \mu_{**}(x)\big) \cdot \big(\eta(x_n) - \eta_*(x_n)\big)\Big] =$$

$$\mathbb{E}\Big[D_{a^*}\big(\mu(x_n)\big|\mu_*(x_n)\big)\Big] + \mathbb{E}\Big[\big(\mu_*(x) - \mu_{**}(x)\big) \cdot \eta(x_n)\Big] \tag{1}$$

for D_a the Bregman divergence associated to a, and D_{a^*} the one associated to a^*. We also use the optimality condition for the predictor $\eta_*(x)$: $\mathbb{E}\eta_*(x)[a'(\eta_*(x)) - \mathbb{E}_{p(x|y)} y] = 0$ in the last step.

When the model is well-specified, we have $a'(\eta_*(x_n)) = \mathbb{E}(y_n|x_n)$ and thus $F(\eta) - F(\eta_*) = \mathbb{E}\big[D_{a^*}(\mu_*(x_n)||\mu(x_n))\big]$. If η is linear in $\Phi(x)$, and even if the model is mis-specified, then we also have $F(\eta) - F(\eta_*) = \mathbb{E}\big[D_{a^*}(\mu_*(x_n)||\mu(x_n))\big]$.

Using asymptotic expansions of moments of the averaged SGD iterate with zero-mean statistically independent noise $f_n(\theta) = F(\theta) + \varepsilon_n(\theta)$ from [1], Theorem 2 one obtains:

$$\overline{\theta}_\gamma = \mathbb{E}_{\pi_\gamma}(\theta) = \theta_* + \gamma\Delta + O(\gamma^2), \tag{2}$$

$$\mathbb{E}_{\pi_\gamma}(\theta - \theta_*)(\theta - \theta_*)^\top = \gamma C + O(\gamma^2), \tag{3}$$

where

$$C = \big[F''(\theta_*) \otimes I + I \otimes F''(\theta_*)\big]^{-1}\Sigma.$$

and $\Sigma = \int\limits_{\mathbb{R}^d} \varepsilon(\theta)^{\otimes 2} \pi_\gamma(d\theta) \in \mathbb{R}^{d\times d}$.

The "drift" $\overline{\theta}_\gamma - \theta_*$ is linear in γ and can be interpreted as an additional error due to the function is not being quadratic and step sizes are not decaying to zero.

Connection between Δ and C can be easily obtained using $\theta_n = \theta_{n-1} - \gamma\big[F'(\theta_{n-1}) + \varepsilon_n\big]$. Taking expectation of both parts and using Tailor expansion up to the second order:

$$F''(\theta_*)(\overline{\theta}_\gamma - \theta_*) = -\frac{1}{2}F'''(\theta_*)\mathbb{E}_{\pi_\gamma}(\theta - \theta_*)^{\otimes 2} \Rightarrow$$

$$F''(\theta_*)\Delta = -\frac{1}{2}F'''(\theta_*)C. \tag{4}$$

1.1 Estimation without averaging

We start with the simplest estimator of the prediction function: $\mu_0(x) = a'(\Phi^\top \theta_n)$, where we do not use any averaging:

$$\mathcal{G}(\mu_n) - \mathcal{G}(\mu_*) = f(\theta_n) - f(\theta_*) = f'(\theta_*)(\theta_n - \theta_*) + \frac{1}{2} f''(\theta_*)(\theta_n - \theta_*)^{\otimes 2} + \frac{1}{6} f'''(\theta_*)(\theta_n - \theta_*)^{\otimes 3} + O(\gamma^{3/2})$$

Taking expectation of both sides, when $n \to \infty$ and using Eq. (3) one obtains:

$$A(\gamma) = \mathbb{E}_{\pi_\gamma} f(\theta_n) - f(\theta_*) = \frac{1}{2} \mathrm{tr} f''(\theta_*) \gamma C + O(\gamma^{3/2}).$$

So, we have linear dependence of γ and $B = \frac{1}{2} \mathrm{tr} f''(\theta_*) C$.

1.2 Estimation with averaging parameters

Now, let us estimate $\overline{A}(\gamma)$:

$$\mathcal{G}(\bar{\mu}_n) - \mathcal{G}(\mu_*) = f(\bar{\theta}_n) - f(\theta_*) = f'(\theta_*)(\bar{\theta}_n - \theta_*) + \frac{1}{2} (\bar{\theta}_n - \theta_*) f''(\theta_*)(\bar{\theta}_n - \theta_*) + O(\gamma^3).$$

Taking expectation of both sides, when $n \to \infty$:

$$\mathcal{G}(\bar{\mu}_\gamma) - \mathcal{G}(\mu_*) = f(\bar{\theta}_\gamma) - f(\theta_*) = \frac{1}{2} \mathrm{tr} f''(\theta_*)(\bar{\theta}_\gamma - \theta_*)^{\otimes 2} + O\big(\gamma^3\big) = \frac{1}{2} \mathrm{tr} f''(\theta_*) \gamma^2 \Delta^{\otimes 2} + O(\gamma^3).$$

Finally we have a quadratic dependence of γ:

$$\bar{A}(\gamma) = \frac{1}{2} \mathrm{tr} f''(\theta_*) \gamma^2 \Delta^{\otimes 2} + O(\gamma^3).$$

And the coefficient $\bar{B} = \frac{1}{2} \mathrm{tr} f''(\theta_*) \Delta^{\otimes 2}$.

1.3 Estimation with averaging predictions

Recall, that by definition, $\bar{\bar{A}}(\gamma) = \mathcal{G}(\bar{\bar{\mu}}_\gamma) - \mathcal{G}(\mu_*)$, where $\bar{\bar{\mu}}_\gamma(x) = \mathbb{E}_{\pi_\gamma} a'\big(\theta^\top \Phi(x)\big)$. We again use Tailor expansion for $a'\big(\theta^\top \Phi(x)\big)$ at θ^*:

$$a'\big(\theta^\top \Phi(x)\big) = a'\big(\theta_*^\top \Phi(x)\big) + a''\big(\theta_*^\top \Phi(x)\big)(\theta - \theta_*)^\top \Phi(x) + \frac{1}{2} a'''\big(\theta_*^\top \Phi(x)\big) \cdot \Big((\theta - \theta_*)^\top \Phi(x)\Big)^2 + O(\gamma^{3/2}).$$

Taking expectation of both parts:

$$\bar{\bar{\mu}}_\gamma(x) = \mu_*(x) + a''\big(\theta_*^\top \Phi(x)\big)(\bar{\theta}_\gamma - \theta_*)^\top \Phi(x) + \frac{1}{2} a'''\big(\theta_*^\top \Phi(x)\big) \mathrm{tr}\big[\Phi(x)\Phi(x)^\top \mathbb{E}(\theta - \theta_*)^{\otimes 2}\big] + O(\gamma^{3/2}) =$$

$$= \mu_*(x) + a''\big(\theta_*^\top \Phi(x)\big) \gamma \Delta^\top \Phi(x) + \frac{1}{2} a'''\big(\theta_*^\top \Phi(x)\big) \mathrm{tr}\big[\Phi(x)\Phi(x)^\top \gamma C\big] + O(\gamma^{3/2}).$$

Finally, we showed, that:

$$\bar{\bar{\mu}}_\gamma(x) - \mu_*(x) = O(\gamma^{3/2}) + \gamma\Big[a''\big(\eta_*(x)\big)\Delta^\top \Phi(x) + \frac{1}{2} a'''\big(\eta_*(x)\big) \mathrm{tr}\big[\Phi(x)^{\otimes 2} C\big]\Big]$$

Now we use Bregram divergence notation Eq. (1):

$$\bar{\bar{A}}(\gamma) = \mathcal{G}(\bar{\bar{\mu}}_\gamma) - \mathcal{G}(\mu_*) = \mathcal{G}_1 + \mathcal{G}_2,$$

As mentioned above, the term $\mathcal{G}_2$ vanishes if model is well-specified or η is linear in $\Phi(x)$. Note, that for the case $\bar{A}(\gamma)$ indeed linear in $\Phi(x)$.

1.3.1 Estimation of $\mathcal{G}_1$.

By definition $D_{a^*}(\mu_*(x)||\mu(x)) = \frac{1}{2}\big(\mu_*(x)-\mu(x)\big)(a^*)''\big(\mu_*(x)\big)\big(\mu_*(x)-\mu(x)\big)$ and

$$\mathcal{G}_1 = \frac{1}{2}\mathbb{E}\frac{\big(\mu_*(x)-\bar{\bar{\mu}}_\gamma(x)\big)^2}{a''\big(\theta_*^\top\Phi(x)\big)} = \frac{\gamma^2}{2}\mathbb{E}\Big[\frac{(a''\big(\eta_*(x)\big)\Delta^\top\Phi(x)+\frac{1}{2}a'''\big(\eta_*(x)\big)\mathrm{tr}\big[\Phi(x)^{\otimes 2}C\big])^2}{a''\big(\theta_*^\top\Phi(x)\big)}\Big].$$

Since

$$\mathbb{E}_x\Big[a''(\big(\theta_*^\top\Phi(x)\big)(\Delta^\top\Phi(x))^2\Big] = \Delta^\top f''(\theta_*)\Delta$$

and

$$\mathbb{E}_x\Big[\Delta^\top a'''\big(\theta_*^\top\Phi(x)\big)\Phi(x)^{\otimes 3}C\Big] = \Delta^\top f'''(\theta_*)C = -2\Delta^\top f''(\theta_*)\Delta,$$

$$\mathcal{G}_1 = \gamma^2\Bigg[-\frac{1}{2}\Delta^\top f''(\theta_*)\Delta + \frac{1}{8}\mathbb{E}\Big[\frac{a'''(\eta_*(x))^2}{a''(\eta_*(x))}\cdot\big(\mathrm{tr}\big[\Phi(x)^{\otimes 2}C\big]\big)^2\Big]\Bigg] + O(\gamma^3)$$

And the coefficient $\bar{\bar{B}}^w = -\frac{1}{2}\Delta^\top f''(\theta_*)\Delta + \frac{1}{8}\mathbb{E}\Big[\frac{a'''(\eta_*(x))^2}{a''(\eta_*(x))}\cdot\big(\mathrm{tr}\big[\Phi(x)^{\otimes 2}C\big]\big)^2\Big]$.

1.3.2 Estimation of $\mathcal{G}_2$.

$$\mathcal{G}_2 = \mathbb{E}\Big[\big(\mu_*(x)-\mu_{**}(x)\big)\cdot\big(\bar{\bar{\eta}}(x_n)-\eta_*(x_n)\big)\Big],$$

using properties of conjugated functions,

$$\mathcal{G}_2 = \mathbb{E}\Big[\big((a^*)'(\bar{\bar{\mu}}(x)) - (a^*)'(\mu_*(x))\big)\cdot\big(\mu_*(x)-\mu_{**}(x)\big)\Big] =$$

$$\mathbb{E}\Big[(a^*)''\big(\bar{\bar{\mu}}_*(x)\big)(\bar{\bar{\mu}}(x)-\mu_*(x)\big)\cdot\big(\mu_*(x)-\mu_{**}(x)\big) + O(\gamma^2) =$$

$$= \mathbb{E}\ \frac{\bar{\bar{\mu}}(x)-\mu_*(x)}{a''(\eta_*(x))}\cdot\big(\mu_*(x)-\mu_{**}(x)\big) + O(\gamma^2) =$$

$$= \gamma\cdot\mathbb{E}\Bigg[\Big(\Delta^\top\Phi(x) + \frac{a'''(\eta_*(x))}{2a''(\eta_*(x))}\mathrm{tr}\big[\Phi(x)^{\otimes 2}C\big]\Big)\cdot\Big(\mu_*(x)-\mu_{**}(x)\Big)\Bigg] + O(\gamma^2).$$

And the coefficient $\bar{\bar{B}}^m = \mathbb{E}\Big(\Delta^\top\Phi(x) + \frac{a'''(\eta_*(x))}{2a''(\eta_*(x))}\mathrm{tr}\big[\Phi(x)^{\otimes 2}C\big]\Big)\cdot\Big(\mu_*(x)-\mu_{**}(x)\Big)$.

References

[1] A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *Ann. Statist.*, 44(4):1363–1399, 08 2016.

[2] A. Dieuleveut, A. Durmus, and F. Bach. Bridging the gap between constant step size stochastic gradient descent and markov chains. Technical Report 1707.06386, arXiv, 2017.