7 Appendix

7.1 Details on models used

Throughout the paper four different models are considered: two variants of a Bayesian neural network (BNN-A and BNN-B), a hierarchical Poisson model, and Bayesian logistic regression.

Bayesian logistic regression: We consider two different datasets: "a1a" and "Mushrooms". In all cases the training set is given by $\mathcal{D} = \{x_i, y_i\}$, where y_i is binary. The model is specified by

$$\begin{aligned} w_i &\sim \mathcal{N}(0, 1), \\ p_i &= \left(1 + \exp(w_0 + w \cdot x_i)\right)^{-1}, \\ y_i &\sim \text{Bernoulli}(p_i). \end{aligned}$$

Hierarchical Poisson model: By Gelman et al. Gelman et al. (2007). The model measures the relative stop-and-frisk events in different precincts in New York city, for different ethnicities. The model is specified by

$$\begin{aligned} \mu &\sim \mathcal{N}(0, 10^2) \\ \log \sigma_\alpha &\sim \mathcal{N}(0, 10^2), \\ \log \sigma_\beta &\sim \mathcal{N}(0, 10^2), \\ \alpha_e &\sim \mathcal{N}(0, \sigma_\alpha^2), \\ \beta_p &\sim \mathcal{N}(0, \sigma_\beta^2), \\ \lambda_{ep} &= \exp(\mu + \alpha_e + \beta_p + \log N_{ep}), \\ Y_{ep} &\sim \text{Poisson}(\lambda_{ep}). \end{aligned}$$

In this case, e stands for ethnicity and p for precinct, Y_{ep} for the number of stops in precinct p within ethnicity group e (observed), and N_{ep} for the total number of arrests in precinct p within ethnicity group e (observed).

BNN-A: As done by Miller et al. Miller et al. (2017) we use a subset of 100 rows from the "Redwine" dataset (regression). We implement a neural network with one hidden layer with 50 units and Relu activations. Let $\mathcal{D} = \{x_i, y_i\}$ be the training set. The model is specified by

$$\begin{aligned} \log \alpha &\sim \mathcal{N}(0, 10^2), \\ \log \tau &\sim \mathcal{N}(0, 10^2), \\ w_i &\sim \mathcal{N}(0, \alpha^2), && \text{(weights and biases)} \\ \hat{y}_i &= \text{FeedForward}(x_i, W), \\ y_i &\sim \mathcal{N}(\hat{y}_i, \tau^2). \end{aligned}$$

BNN-B: We use a subset of 200 rows from the "Redwine" dataset (regression). We implement a neural network with one hidden layer with 50 units and Relu activations. Let $\mathcal{D} = \{x_i, y_i\}$ be the training set. The model is specified by

$$\begin{aligned} \log \tau &\sim \mathcal{N}(0, 5^2), \\ w_i &\sim \mathcal{N}(0, 5^2), && \text{(weights and biases)} \\ \hat{y}_i &= \text{FeedForward}(x_i, W), \\ y_i &\sim \mathcal{N}(\hat{Y}_i, \tau^2). \end{aligned}$$

The only difference between BNN-A and BNN-B is in the prior used for the weights and biases.

7.2 Estimators chosen by Algorithm 1 (Section 3.4)

Model	Time of choice		
	$T = 0$	$T = T_{opt}/10$	$T = T_{opt}/2$
Log Reg (a1a)	(Rep)	(Rep)	(STL)
Hier Poisson	(Miller)	(Miller)	(Miller)
BNN-A	(Rep)	(STL)	(STL)
BNN-B	(Rep)	(Rep)	(Rep)

Table 4: **Algorithm 1 selects different gradient estimators for different models.** Gradient estimators chosen by Algorithm 1. Each column shows the estimator selected at some selection time. (In different runs different choices may ocurr. We show the "most popular" choice across runs.)

7.3 Mixed Integer Quadratic Program

A mixed integer quadratic program is an optimization problem in which the objective function and constraints are quadratic (or linear), and some (or all) variables are restricted to be integers:

$$\begin{aligned} \underset{x}{\text{minimize}} \quad & \frac{1}{2} x^\top Q_0 x + r_0^\top x + u_0 \\ \text{s.t.} \quad & \frac{1}{2} x^\top Q_i x + r_i^\top x + u_i \geq 0 \quad i = 1, ..., m, \\ & Ax + b = 0, \end{aligned} \tag{12}$$

where $x \in \mathbb{R}^n$, $Q_0, ..., Q_m \in \mathbb{R}^{n \times n}$, and some components of x are restricted to be integers.

We now prove Theorem 4.1.

Theorem 4.1. *When different gradient estimators are indexed by a set of J control variate weights, the problem of finding $a^*(w)$ as in equation (10) can be*

reduced to solving a mixed integer quadratically constrained program with $2J + 2$ variables, one quadratic constraint, and one linear constraint.

Proof. Given

$$\bar{T}(g_a) = \bar{T}(g_{\text{base}}) + \sum_{i=1}^{J} \bar{T}(c_i)\ \mathbf{1}[a_i \neq 0] \tag{13}$$

and

$$\bar{G}^2(g_a, w) = \frac{1}{M} \sum_{m=1}^{M} \|g_{\text{base}}(w, \xi_m) + C(w, \xi_m)a\|^2, \tag{14}$$

we want to find

$$a^*(w) = \underset{a \in \mathbb{R}^J}{\arg\min}\ \bar{G}^2(g_a, w) \times \bar{T}(g_a). \tag{15}$$

To simplify notation, we use $\bar{G}^2 = \bar{G}^2(g_a, w)$, $g_{bm} = g_{\text{base}}(w, \xi_m)$ and $C_m = C(w, \xi_m)$. Expanding the squared norm in eq. 14 we get

$$\begin{aligned}
\bar{G}^2 &= \frac{1}{M} \sum_{m=1}^{M} \|g_{bm} + C_m a\|^2 \\
&= \frac{1}{M} \sum_{m=1}^{M} \left(g_{bm}^\top g_{bm} + 2 g_{bm}^\top C_m a + a^\top C_m^\top C_m a \right) \\
&= \underbrace{\frac{1}{M} \sum_{m=1}^{M} \|g_{bm}\|^2}_{u_1} + \underbrace{\left(\frac{2}{M} \sum_{m=1}^{M} g_{bm}^\top C_m \right)}_{r_1} a \qquad (16) \\
&\quad + \frac{1}{2} a^\top \underbrace{\left(\frac{2}{M} \sum_{m=1}^{M} C_m^\top C_m \right)}_{Q_1} a \\
&= u_1 + r_1^\top a + \frac{1}{2} a^\top Q_1 a. \qquad (17)
\end{aligned}$$

On the other hand, equation 13 can be expressed as

$$\bar{T}(g_a) = t_0 + t^\top b, \text{ s.t. } b_i = \mathbf{1}[a_i \neq 0], \tag{18}$$

where $t_0 = \bar{T}(g_{\text{base}})$, and $t_i = \bar{T}(c_i)$. Using equations 17 and 18, the minimization problem from equation 15 can be expressed as

$$\begin{aligned}
(a^*, b^*) = \underset{a \in \mathbb{R}^J, b \in \{0,1\}^J}{\arg\min} & \left(\frac{1}{2} a^\top Q_1 a + r_1^\top a + u_1 \right) \times \left(t_0 + t^\top b \right), \\
\text{s.t } & b_i = \mathbf{1}[a_i \neq 0].
\end{aligned} \tag{19}$$

Introducing two extra varaibles, V_G and V_T, we can express the minimization problem in eq. 19 as

$$\begin{aligned}
(a^*, b^*, V_G^*, V_T^*) = \underset{a \in \mathbb{R}^J, b \in \{0,1\}^J, V_G \in \mathbb{R}, V_T \in \mathbb{R}}{\arg\min} & V_G \times V_T, \\
\text{s.t } & V_G \geq \frac{1}{2} a^\top Q_1 a + r_1^\top a + u_1 \\
& V_T = t_0 + t^\top b \\
& b_i = \mathbf{1}[a_i \neq 0].
\end{aligned} \tag{20}$$

The final minimization problem shown in equation 20 has the form of a general MIQCP, shown in equation 12, with the exception of the last constraint $b_i = \mathbf{1}[a_i \neq 0]$. Despite not being in the original definition of a MIQCP, several solver accept constraints of this type (Gurobi Gurobi Optimization (2018), the solver used in our simulation, does).

□

7.4 SGD convergence rates

Recall the following definitions:

Convexity: F is convex iff it $F(\theta x + (1-\theta)y) \leq \theta F(x) + (1-\theta)F(y)\ \forall \theta \in [0, 1]$.

Strong convexity: F is λ-strongly convex iff $F(y) \geq F(x) + \nabla F(x)^\top (y - x) + \frac{\lambda}{2}||y - x||^2\ \forall x, y$.

Smoothness: F has Lipschitz continuous gradient with constant L (F is L-smooth) iff $||\nabla F(y) - \nabla F(x)|| \leq L||y - x||\ \forall x, y$.

We now state the convergence rates presented in Table 2 in more detail.

Theorem 7.1 (F strongly convex; decaying step-size; By Rakhlin et al. Rakhlin et al. (2012))**.** *Let F be a λ-strongly convex function over a convex set W. If we assume $\mathbb{E}_\xi[||g(w, \xi)||^2] \leq G^2\ \forall w \in W$ and set the step size $\eta_t = 1/(\lambda t)$, then, after K updates of SGD, we have*

$$\mathbb{E}\left[||w_K - w^*||^2\right] \leq \frac{4}{\lambda^2} \frac{G^2}{K}; \quad \textit{where } w^* = \textit{argmin}_w F(w)$$

Proof. See Rakhlin et al. Rakhlin et al. (2012). □

Corollary 7.1.1 (F strongly convex and smooth; decaying step-size; By Rakhlin et al. Rakhlin et al. (2012))**.** *Let F be an L-smooth and λ-strongly convex function over a convex set W. If we assume $\mathbb{E}_\xi[||g(w, \xi)||^2] \leq G^2\ \forall w \in W$ and set the step size $\eta_t = 1/(\lambda t)$, then, after K updates of SGD, we have*

$$\mathbb{E}\left[F(w_K) - F(w^*)\right] \leq \frac{2L}{\lambda^2} \frac{G^2}{K}.$$

Proof. See Rakhlin et al. Rakhlin et al. (2012). □

Theorem 7.2 (F convex; constant step-size; Nemirovski et al. Nemirovski et al. (2009))**.** *Let F be a convex function over a convex set W. Assume $\mathbb{E}_\xi[||g(w,\xi)||^2] \leq G^2 \, \forall w \in W$. Then, after K updates of SGD using the optimal learning rate $\eta^* = \frac{Dw}{G\sqrt{K}}$ we have*

$$\mathbb{E}\left[F(\bar{w}) - F(w^*)\right] \leq D_w \frac{G}{\sqrt{K}},$$

where $\bar{w} = \frac{1}{K}\sum_{i=1}^{K} w_i$ and $D_w = ||w_0 - w^*||$.

Proof. See Nemirovski et al. Nemirovski et al. (2009). □

Theorem 7.3 (F smooth; constant step-size)**.** *Let F be an L-smooth function. Assume $\mathbb{E}_\xi[||g(w,\xi)||^2] \leq G^2 \, \forall w$. Then, after K updates of SGD using the optimal learning rate $\eta^* = \sqrt{\frac{2D_f}{LKG^2}}$ we have*

$$\mathbb{E}\left[\frac{1}{K}\sum_{i=1}^{K} ||\nabla F(w_i)||^2\right] \leq \sqrt{LD_f}\frac{G}{\sqrt{K}},$$

where $D_f = \mathbb{E}[F(w_0) - F(w^*)]$.

Proof. This proof is a straightforward adaptation from the one by Bottou et al. Bottou et al. (2018).

$$\begin{aligned} F(w_{t+1}) &\leq F(w_t) + \nabla F(w_t)^\top (w_{t+1} - w_t) + \frac{L}{2}||w_{t+1} - w_t||^2 \\ &= F(w_t) - \eta_t \nabla F(w_t)^\top g_t + \frac{L}{2}||\eta_t g_t||^2 \\ &= F(w_t) - \eta_t \nabla F(w_t)^\top g_t + \frac{L\eta_t^2}{2}||g_t||^2. \end{aligned}$$

Taking the expectation on both sides we get

$$\begin{aligned} \mathbb{E}[F(w_{t+1}) - F(w_t)] &\leq -\eta_t \nabla \, \mathbb{E}[F(w_t)^\top g_t] + \frac{L\eta_t^2}{2}\,\mathbb{E}\,||g_t||^2 \\ &= -\eta_t \,\mathbb{E}\,||\nabla F(w_t)||^2 + \frac{L\eta_t^2}{2}G^2. \end{aligned}$$

If we take $\eta_t = \eta$, and sum up both sides of the inequality we get

$$\begin{aligned} \sum_{t=0}^{K-1} \mathbb{E}[F(w_{t+1}) - F(w_t)] &\leq -\eta \sum_{t=0}^{K-1} \mathbb{E}\,||\nabla F(w_t)||^2 + \frac{KL\eta^2}{2}G^2 \\ \mathbb{E}[F(w_K) - F(w_0)] &\leq -\eta \sum_{t=0}^{K-1} \mathbb{E}\,||\nabla F(w_t)||^2 + \frac{KL\eta^2}{2}G^2. \end{aligned}$$

Re-arranging and using the fact that $F(w^*) \leq F(w_K)$ gives

$$\begin{aligned} \mathbb{E}\left[\frac{1}{K}\sum_{t=0}^{K-1} ||\nabla F(w_t)||^2\right] &\leq \frac{L\eta}{2}G^2 + \frac{F(w_0) - F(w^*)}{\eta K} \\ &= \frac{L\eta}{2}G^2 + \frac{D_f}{\eta K}. \end{aligned}$$

Where $D_f = F(w_0) - F(w^*)$. The value for η that minimizes the right hand side of the last inequality is $\eta^* = \sqrt{\frac{D_f}{LKG^2}}$. Using $\eta = \eta^*$ yields

$$\mathbb{E}\left[\frac{1}{K}\sum_{t=0}^{K-1} ||\nabla F(w_t)||^2\right] \leq \sqrt{\frac{L(F(w_0) - F(w^*))G^2}{K}}.$$

□

Finally, slightly different versions of the last two bounds in Table 2 were proposed by Yang et al. Yang et al. (2016). The authors carry out a unified analysis for stochastic momentum methods using a parameter s; if $s = 0$ the algorithm results in the heavy ball method, and if $s = 1$ in a stochastic variant of Nesterov's accelerated gradient. They state the following result:

Theorem 7.4 (F smooth; constant step-size; stochastic momentum methods; Yang et al. Yang et al. (2016))**.** *Let F be a (possibly non convex) L-smooth function, Assume $\mathbb{E}_\xi[||g(w,\xi) - \nabla F(w)||^2] \leq \delta^2$ and $||\nabla F(w)||^2 \leq V^2$. Then, after K updates of the proposed stochastic momentum method (β) with learning rate $\eta = \min\left\{\frac{1-\beta}{2L}, \frac{C}{\sqrt{T}}\right\}$, we have*

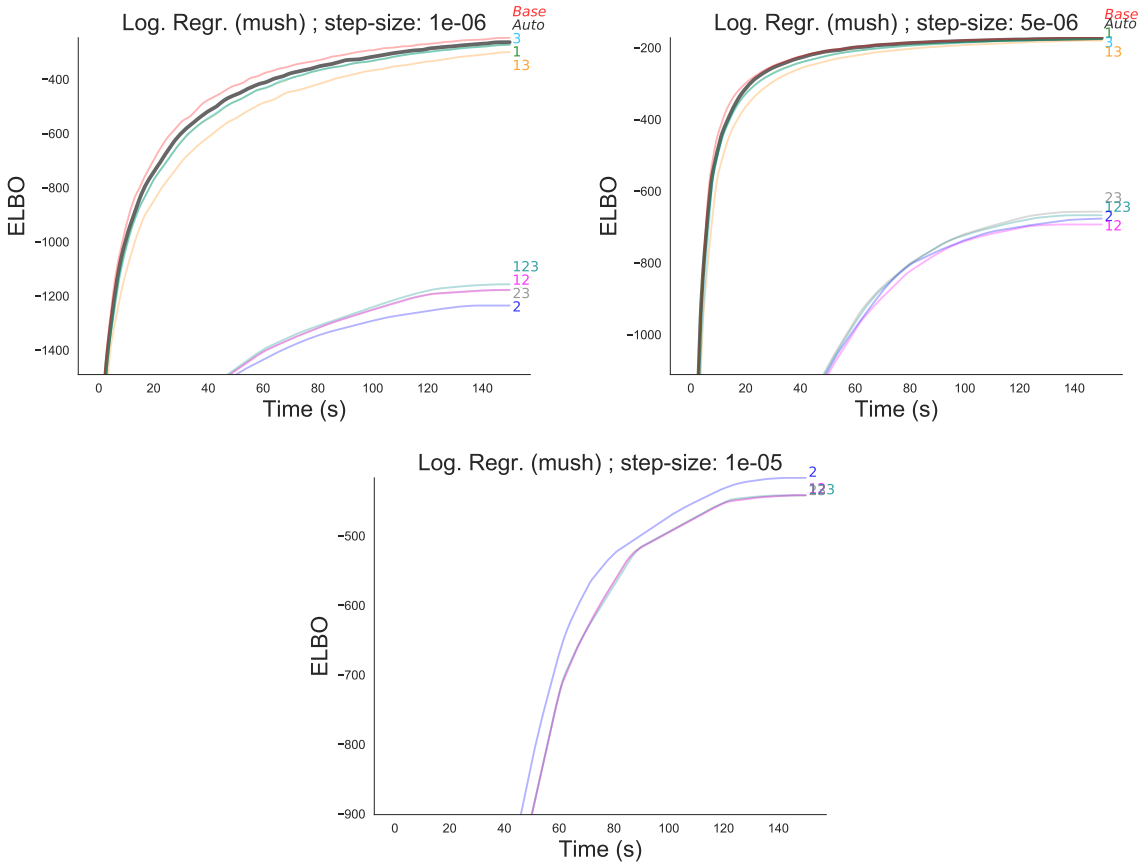
$$\min_{k=0,\ldots,K} \mathbb{E}\,||\nabla F(w_k)||^2 \leq \frac{2D_f(1-\beta)}{T}\max\left\{\frac{2L}{1-\beta}, \frac{\sqrt{T}}{C}\right\}$$

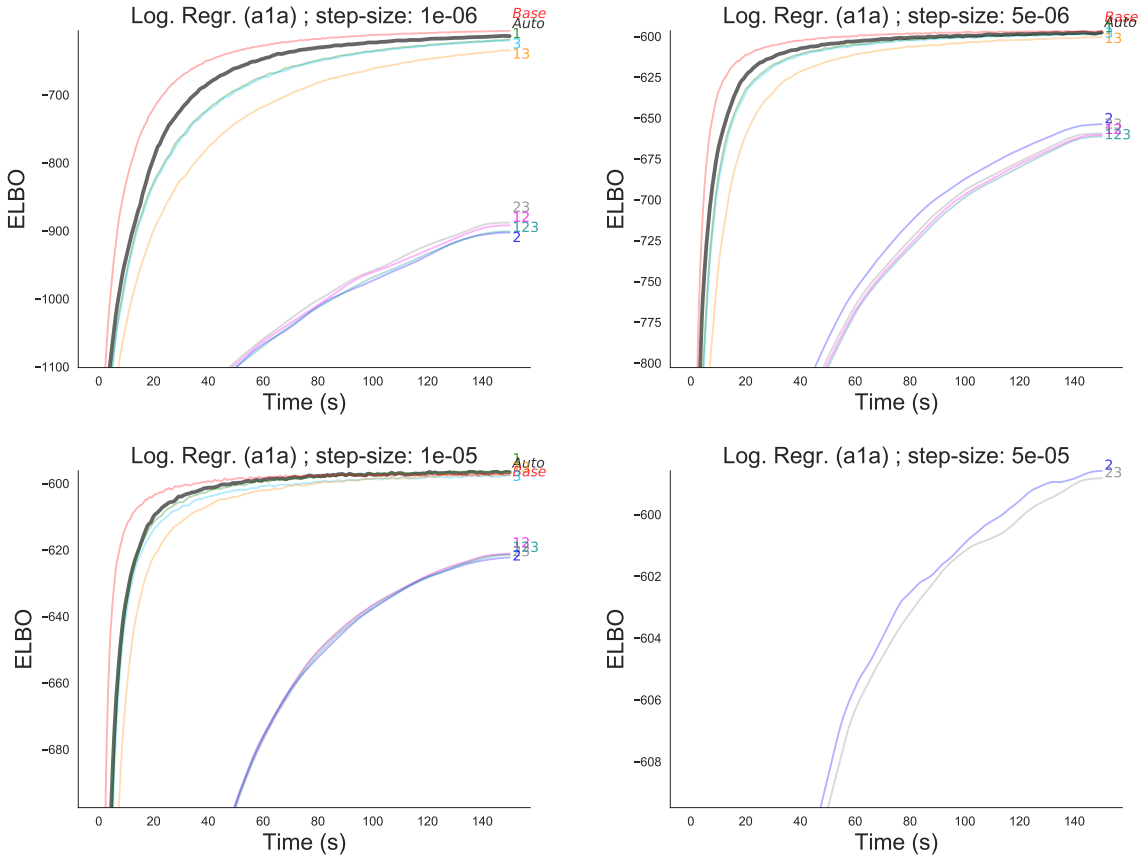
$$+\frac{C}{\sqrt{T}}\frac{L\beta^2\left((1-\beta)s-1\right)^2(V^2+\delta^2) + L\delta^2(1-\beta)^2}{(1-\beta)^3},$$

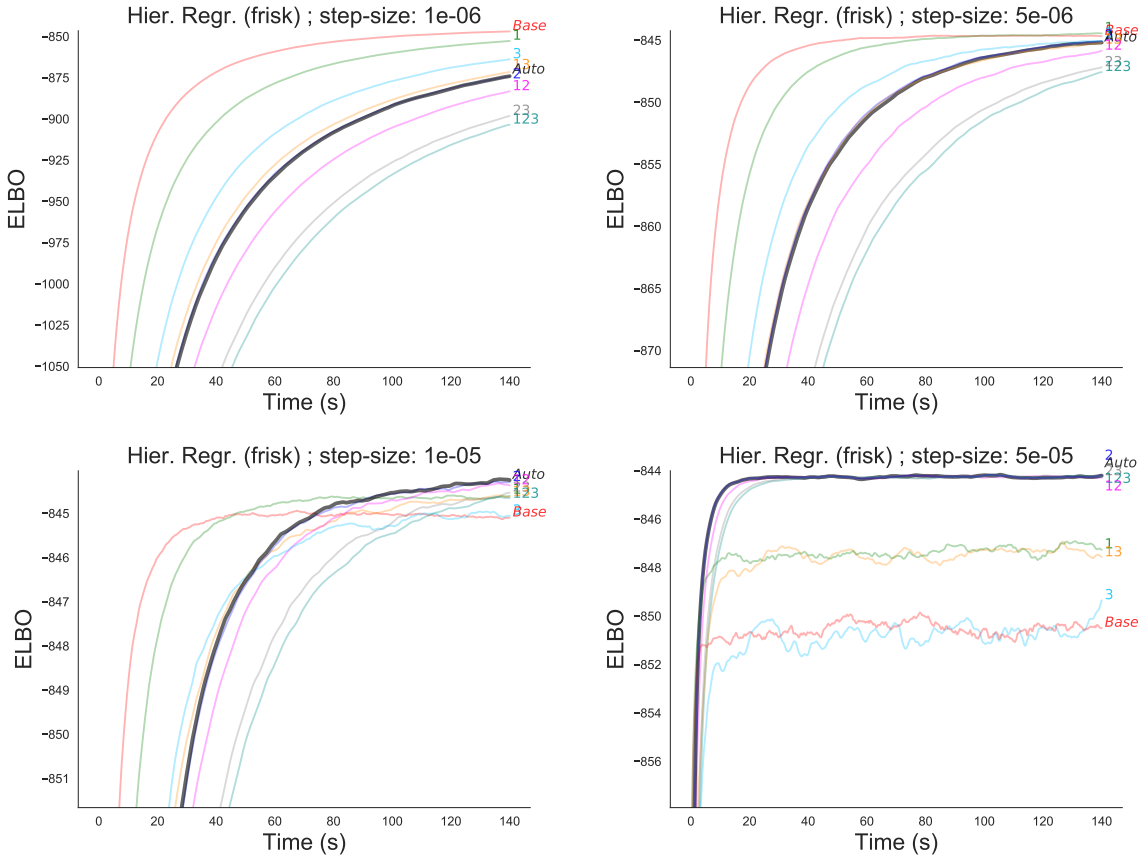
where $D_f = F(w_0) - F(w^*)$.

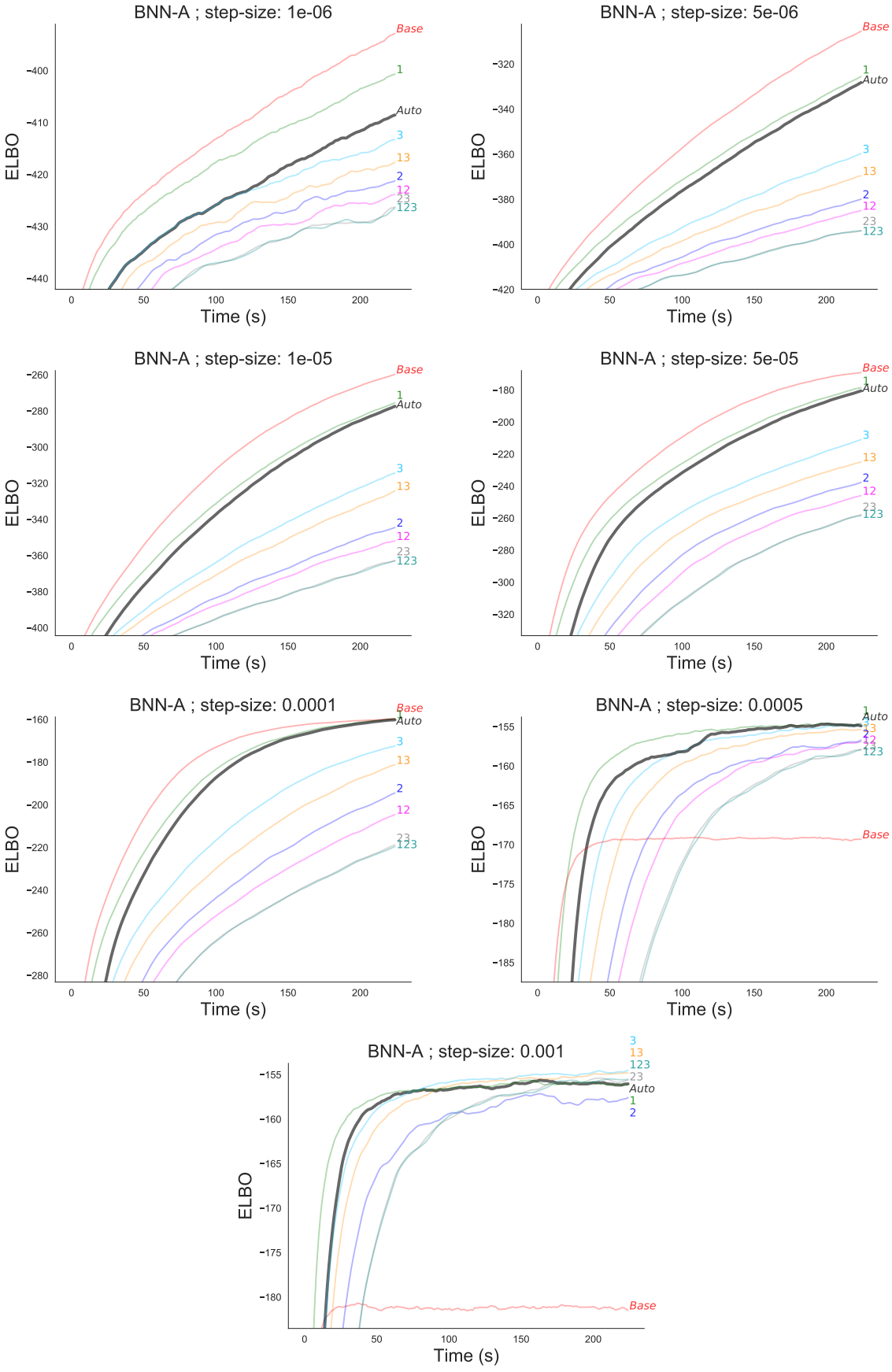
The results shown in Table 2 are obtained by: 1) setting $s = 0$ (momentum) or $s = 1$ (Nesterov); 2) finding the optimal C; 3) bounding $\mathbb{E}_\xi\,||g(w,\xi)||^2 \leq \delta^2 + V^2 = G^2$; and 4) assuming that optimization is performed for a large enough number of steps $\left(K \geq \frac{4L^2C^2}{1-\beta}\right)$.

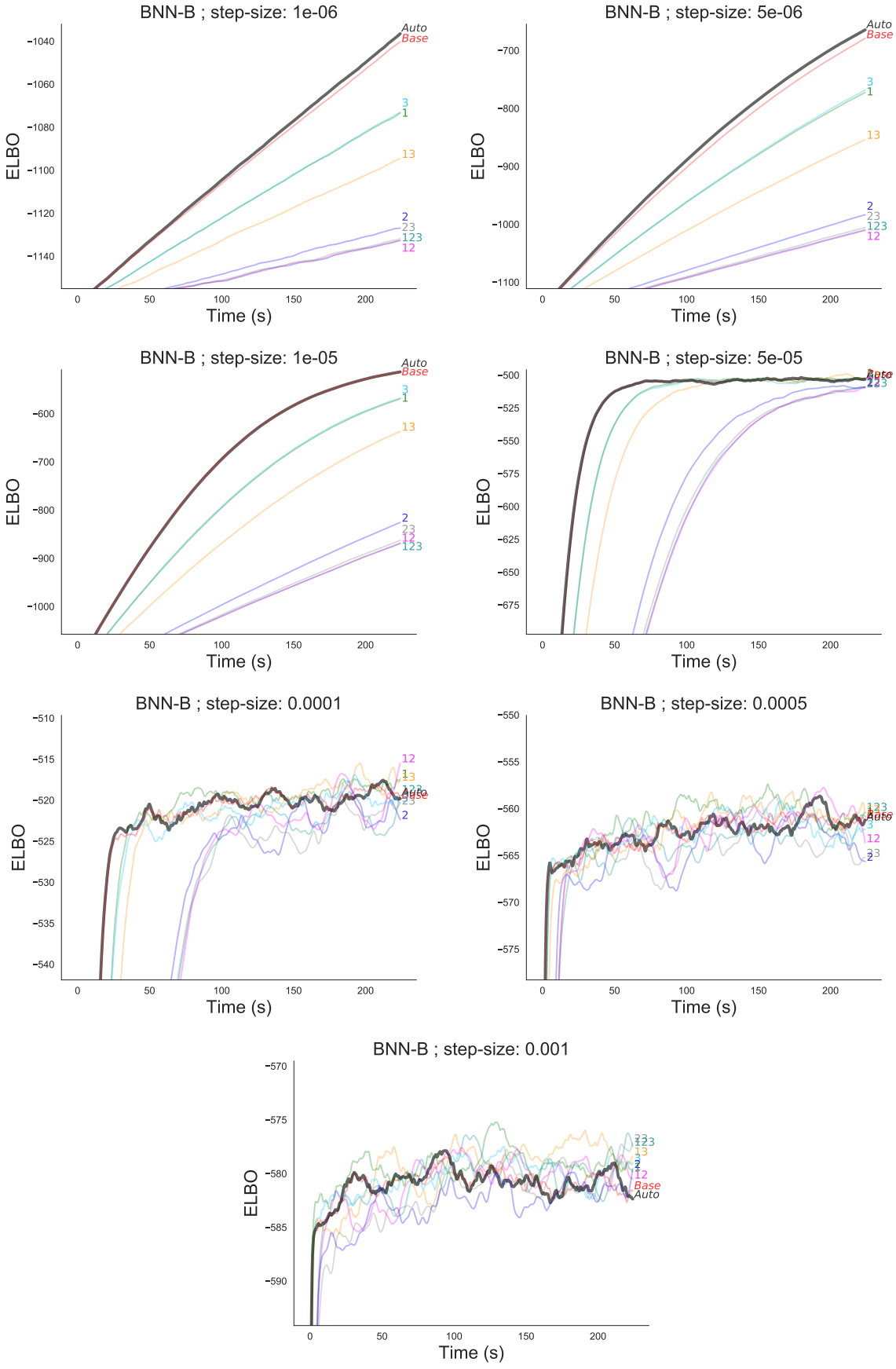
7.5 Raw results for individual step-sizes


Log. Regr. (mush) ; step-size: 1e-06
Base
Auto
3
1
13
123
12
23
2
ELBO
Time (s)
Log. Regr. (mush) ; step-size: 5e-06
Base
Auto
3
1
13
23
123
2
12
ELBO
Time (s)
Log. Regr. (mush) ; step-size: 1e-05
2
123
ELBO
Time (s)


Log. Regr. (a1a) ; step-size: 1e-06
Log. Regr. (a1a) ; step-size: 5e-06
Log. Regr. (a1a) ; step-size: 1e-05
Log. Regr. (a1a) ; step-size: 5e-05
ELBO
Time (s)
Base
Auto
1
3
13
23
12
123
2


Hier. Regr. (frisk) ; step-size: 1e-06
Hier. Regr. (frisk) ; step-size: 5e-06
Hier. Regr. (frisk) ; step-size: 1e-05
Hier. Regr. (frisk) ; step-size: 5e-05
ELBO
Time (s)
Base
Auto
1
2
3
12
13
23
123


BNN-A ; step-size: 1e-06
BNN-A ; step-size: 5e-06
BNN-A ; step-size: 1e-05
BNN-A ; step-size: 5e-05
BNN-A ; step-size: 0.0001
BNN-A ; step-size: 0.0005
BNN-A ; step-size: 0.001
ELBO
Time (s)
Base
Auto
1
3
13
2
12
23
123


BNN-B ; step-size: 1e-06
BNN-B ; step-size: 5e-06
BNN-B ; step-size: 1e-05
BNN-B ; step-size: 5e-05
BNN-B ; step-size: 0.0001
BNN-B ; step-size: 0.0005
BNN-B ; step-size: 0.001
ELBO
Time (s)
Auto
Base
1
2
3
12
13
23
123